When depth is redundant: Efficient transformer-based speech anti-spoofing

Hoan My Tran[1], Damien Lolive[2], Aghilas Sini[3]
Arnaud Delhay[1], Pierre-François Marteau[2], David Guennec[1]
[1]Univ Rennes, CNRS, IRISA, Lannion, France
[2]Univ Bretagne Sud, CNRS, IRISA, Vannes, France
[3]Univ Le Mans, LIUM, Le Mans, France
hoan.tran@irisa.fr, damien.lolive@irisa.fr, aghilas.sini@univ-lemans.fr
arnaud.delhay@irisa.fr, pierre-francois.marteau@irisa.fr, david.guennec@irisa.fr

Abstract

Detecting speech deepfakes is critical for protecting society against fraud, identity theft, and the misuse of modern speech synthesis technologies. Despite recent progress, existing countermeasures often exhibit limited generalization to unseen spoofing attacks, particularly in out-of-domain evaluation settings, even when achieving strong in-domain performance. Transformer architectures have become ubiquitous in anti-spoofing, serving both as feature extractors (e.g., wav2vec 2.0) and as classifiers. However, deep transformer stacks exhibit substantial representational redundancy across adjacent layers, with similarity increasing toward deeper layers. As a result, task-specific specialization is largely concentrated in the final layers, while shallow layers remain underutilized during fine-tuning. In this work, we analyze the layer-wise behavior of transformer-based classifiers for speech deepfake detection and propose a training strategy that explicitly aligns shallow and intermediate representations with those of the final transformer layer. By encouraging all layers to mimic the task-specialized representation learned at depth, the model more effectively exploits early-layer features while preserving discriminative capacity in deeper layers. This design improves robustness to unseen spoofing attacks and enhances out-of-domain generalization. Extensive experiments across multiple benchmark datasets demonstrate consistent performance gains over strong baselines.

1 Introduction

Recent advances in speech generation technologies, particularly text-to-speech (TTS) and voice conversion (VC), have enabled the synthesis of highly natural and realistic speech (Tan et al., 2021). These technologies support beneficial applications such as virtual assistants and assistive communication for individuals who have lost their voices (Medeiros, 2015). However, as the quality of synthetic speech improves, so does the risk of misuse. Speech deepfakes can be exploited to impersonate real individuals, deceiving both humans and automatic speaker verification (ASV) systems, thereby facilitating fraud and identity theft (Folorunsho and Boamah, 2025). This growing threat has motivated the development of anti-spoofing countermeasures (CMs) to distinguish bona fide from synthetic speech (Todisco et al., 2019; Yamagishi et al., 2021; Wang et al., 2024).

State-of-the-art CMs typically rely on self-supervised learning (SSL) speech foundation models (SFMs) as feature extractors, followed by a classifier that maps high-dimensional representations to a binary real/fake decision. Many widely used SFMs are based on the *wav2vec 2.0* architecture (Baevski et al., 2020), including English-only models such as *WavLM* (Chen et al., 2022) and multilingual variants such as *XLS-R* (Babu et al., 2022). These models have demonstrated strong performance in speech deepfake detection. Nevertheless, transformer-based SFMs exhibit substantial representational redundancy across layers, with adjacent hidden states often encoding highly similar information (Pasad et al., 2021, 2023; Ashihara et al., 2024; Dorszewski et al., 2025).

The design of the classifier plays a critical role in effectively leveraging SSL representations. Transformer-based classifiers have achieved SOTA performance on several in-domain spoofing benchmarks (Rosello et al., 2023; Truong et al., 2024; Li et al., 2024; Hao et al., 2025; Dat and Dat, 2025; Kim et al., 2025; Phuong et al., 2025; Tran et al., 2025a). However, their generalization to unseen, out-of-domain spoofing attacks remains limited (Reimao and Tzerpos, 2019; Müller et al., 2022, 2024; Jung et al., 2025; Wang et al., 2026). While increasing model depth can improve in-domain discrimination, it often exacerbates overfitting and complicates interpretation due to limited understanding of layer-wise feature evolution (Gro-

Findings of the Association for Computational Linguistics: ACL 2026, pages 6380–6397
July 2-7, 2026 ©2026 Association for Computational Linguistics

mov et al., 2025; Jiang et al., 2025). Moreover, the quadratic computational complexity of multi-head attention (MHA) (Vaswani et al., 2017) with respect to sequence length poses significant efficiency challenges. Recent work has explored alternatives such as state-space models (SSMs) (Gu and Dao, 2024) to reduce inference cost, but these approaches have not yet resolved the generalization gap in deepfake detection (Xiao and Das, 2025; Tran et al., 2025b; Xuan et al., 2025).

From a theoretical perspective, recent studies have linked deep neural networks to the phenomenon of neural collapse (NC) (Papyan et al., 2020), in which class-conditional representations converge toward a symmetric equiangular tight frame. For deep transformers, increasing depth enforces a progressively tighter approximation to this optimal geometry through layer-wise saturation events (Súkeník et al., 2025; Jiang et al., 2025). While such saturation can enhance class separation and improve in-domain performance, it may also amplify representational redundancy and reduce robustness to distribution shifts (Gromov et al., 2025). In the context of deepfake detection, this suggests that excessive depth may underutilize shallow layers while over-specializing deeper representations.

In this work, we revisit transformer-based classifiers for speech deepfake detection from a layer-wise perspective. We focus on the intrinsic redundancy induced by identical transformer blocks, residual connections, and fixed-dimensional representations. To improve computational efficiency, we adopt multi-head temporal latent attention (MTLA) as a low-cost alternative to standard MTA. More importantly, we introduce a novel regularization strategy that aligns shallow and intermediate transformer layer representations with those of the final layer. This alignment mitigates early-layer underutilization while preserving task-specific specialization at depth, leading to improved robustness against unseen spoofing attacks.

Our contributions are summarized as follows:

1. We systematically analyze the impact of transformer depth in spoofing classifiers, showing that deeper models do not reliably improve out-of-domain generalization and may increase overfitting.

2. We propose an angular-distance regularization that aligns shallow and intermediate layer representations with those of the final layer, improving feature utilization across depth.

3. Extensive experiments across multiple benchmark datasets demonstrate competitive in-domain performance and consistent improvements in out-of-domain robustness over prior approaches.

2 Related works

2.1 Speech deepfake synthesis

Recent years have witnessed rapid advances in speech generation technologies, particularly TTS and VC. Modern TTS systems typically follow a multi-stage pipeline in which input text is first converted into linguistic representations, mapped to acoustic features such as mel-spectrograms, and finally synthesized into speech waveforms using neural vocoders. VC systems, in contrast, aim to preserve linguistic content while modifying speaker-specific characteristics using reference speech from a target speaker (Tan et al., 2021). With large-scale training data and powerful neural architectures, both TTS and VC models can now generate highly natural speech that is often perceptually indistinguishable from genuine human speech (Eskimez et al., 2024; Ju et al., 2024; Chen et al., 2025), enabling increasingly convincing speech deepfakes.

2.2 Anti-spoofing countermeasures

To counter the risks posed by speech deepfakes, anti-spoofing CMs have evolved from systems based on handcrafted acoustic features, such as MFCCs, LFCCs (Lei and Lopez, 2009), and CQCCs (Todisco et al., 2017), toward deep learning approaches leveraging SSL. Pretrained SFMs such as *WavLM* (Chen et al., 2022) and *XLS-R* (Babu et al., 2022) are widely used as front-end feature extractors, followed by trainable classifiers. A variety of classifier architectures have been explored, including LSTMs (Guan et al., 2025), graph-based models (Tak et al., 2022b; Yang et al., 2025b,a), and transformer-based designs, which currently dominate the field due to their strong modeling capacity (Rosello et al., 2023; Truong et al., 2024; Li et al., 2024; Hao et al., 2025; Dat and Dat, 2025; Kim et al., 2025; Phuong et al., 2025; Tran et al., 2025a; Truong et al., 2025b,a). To reduce the high computational cost of self-attention, recent works have explored alternatives such as SSMs (Tran et al., 2025b; Xiao and Das, 2025; Xuan et al., 2025) and gated MLPs (Tran et al., 2025c), reporting improvements in performance.

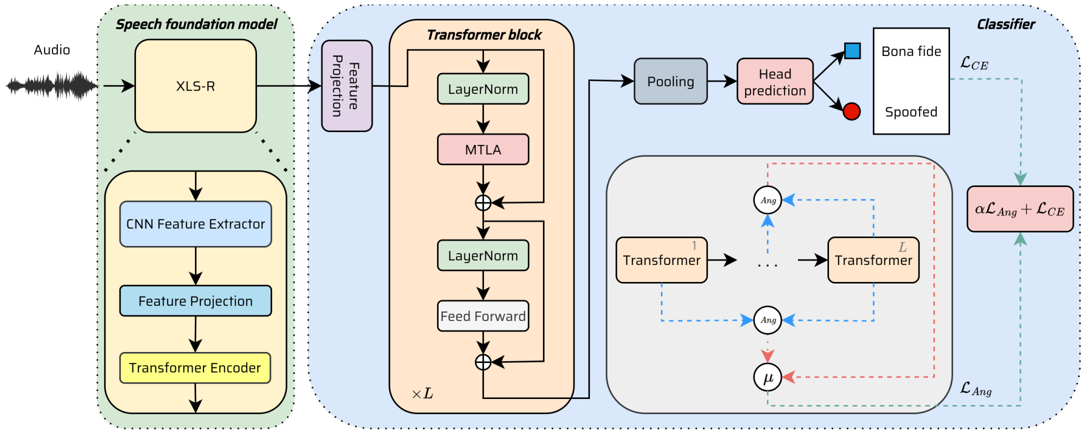

Figure 1: Overview of the proposed model architecture.

2.3 Similarity and representation learning

Pretrained SFMs exhibit strong representational similarity across adjacent transformer layers, resulting in redundant learned features (Dorszewski et al., 2025). From a theoretical standpoint, this phenomenon can be linked to NC (Papyan et al., 2020), in which class-conditional representations converge toward a symmetric equiangular tight frame. For deep transformers, increasing depth enforces a progressively tighter approximation to this geometry through layer-wise saturation events (Súkeník et al., 2025; Jiang et al., 2025). While such saturation improves class separability and in-domain performance, it can amplify redundancy and reduce robustness to distribution shifts (Gromov et al., 2025). Although prior work has attempted to mitigate representational similarity to promote diversity and generalization (Tran et al., 2025c), effectively exploiting or regularizing layer-wise similarity in transformer-based classifiers for speech deepfake detection remains underexplored.

3 Preliminaries

This section introduces the core components of our transformer-based anti-spoofing framework: a pretrained SFM, an efficient attention mechanism, and a metric for analyzing representational similarity across transformer layers.

3.1 Pretrained speech foundation model

We adopt *XLS-R* (Babu et al., 2022), a large-scale multilingual self-supervised SFM pretrained on approximately 436,000 hours of unlabeled audio spanning 128 languages. *XLS-R* extends the *wav2vec 2.0* architecture (Baevski et al., 2020) and consists of a convolutional feature extractor followed by a deep transformer encoder.

Given a raw speech waveform $\mathcal{X}$ sampled at 16 kHz, the convolutional feature extractor $f : \mathcal{X} \rightarrow \mathcal{Z}$ maps $\mathcal{X}$ to a sequence of latent speech representations $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T\}, \quad \mathbf{z}_t \in \mathbb{R}^{d_z}$, where T denotes the number of latent time steps and $d_z = 512$. Due to the convolutional strides, each latent vector corresponds to an effective temporal resolution of approximately 20 ms.

During pretraining, a subset of latent vectors is randomly masked, and the model is optimized using a contrastive masked prediction objective inspired by BERT (Devlin et al., 2019). The masked latent sequence is then processed by a stack of 24 transformer layers, $g : \mathcal{Z} \rightarrow \mathcal{C}$, producing contextualized representations $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T\}, \quad \mathbf{c}_t \in \mathbb{R}^{d_c}$, with hidden dimension $d_c = 1024$. These representations encode long-range temporal dependencies and serve as high-level speech features for downstream tasks, including speech deepfake detection.

3.2 Multi-head temporal latent attention

Given an input sequence $\mathcal{X} \in \mathbb{R}^{T \times d}$, standard MHA computes queries, keys, and values as

$$\begin{aligned} \mathbf{Q} &= \mathcal{X}\mathbf{W}_Q \in \mathbb{R}^{T \times (n_h d_h)}, \\ \mathbf{K} &= \mathcal{X}\mathbf{W}_K \in \mathbb{R}^{T \times (n_h d_h)}, \\ \mathbf{V} &= \mathcal{X}\mathbf{W}_V \in \mathbb{R}^{T \times (n_h d_h)}, \end{aligned} \tag{1}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times (n_h d_h)}$ are learnable parameters and $d_h = d/n_h$ with n_h heads. The attention output is given by

$$\mathrm{Attn}(\mathcal{X}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}, \tag{2}$$

resulting in a space complexity of $\mathcal{O}(T^2)$ and a key-value (KV) cache that grows linearly with T.

Latent-space compression. MTLA (Deng and Woodland, 2025) adopts the low-rank projection logic of MLA (Liu et al., 2024) to compress the feature dimension. The input $\mathcal{X}$ is projected into a latent representation $\mathbf{c'}_t$ for each token:

$$\mathbf{c'}_t = \mathcal{X}\mathbf{W}_r, \quad \mathbf{W}_r \in \mathbb{R}^{r \times d}, \quad r \ll d \tag{3}$$

where r is the latent rank.

Temporal latent compression. MTLA further compresses $\mathbf{C'} = \{\mathbf{c'}_1, \mathbf{c'}_2, \dots, \mathbf{c}_T\}$ along the temporal dimension using learnable weighted aggregation with stride s. Adjacent latent vectors within each temporal group are merged using dynamically generated gating weights computed via a hyper-network, producing a compressed sequence $\hat{\mathbf{C}} \in \mathbb{R}^{\lceil T/s \rceil \times r}$. This reduces the KV cache size by a factor of s, decreasing memory usage and inference cost while maintaining performance.

3.3 Angular distance between transformer layers

To quantify representational similarity across transformer layers, we adopt the angular distance metric (Gromov et al., 2025). Let $\mathcal{T}_{(\ell)}$ denote the ℓ-th transformer layer, and let

$$\mathbf{H}^{(\ell)} = \{\mathbf{h}_1^{(\ell)}, \dots, \mathbf{h}_T^{(\ell)}\} \tag{4}$$

be its sequence of hidden representations. We obtain a sequence-level representation by temporal pooling:

$$\bar{\mathbf{h}}^{(\ell)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}_t^{(\ell)}. \tag{5}$$

Given two layers ℓ and $\ell{+}n$, the cosine similarity between their pooled representations is defined as

$$\cos\theta = \frac{\bar{\mathbf{h}}^{(\ell)} \cdot \bar{\mathbf{h}}^{(\ell+n)}}{\|\bar{\mathbf{h}}^{(\ell)}\|_2 \, \|\bar{\mathbf{h}}^{(\ell+n)}\|_2}. \tag{6}$$

The normalized angular distance is then given by

$$d_{Ang}\left(\bar{\mathbf{h}}^{(\ell)}, \bar{\mathbf{h}}^{(\ell+n)}\right) = \frac{1}{\pi}\arccos(\cos\theta), \tag{7}$$

which lies in the interval $[0, 1]$. Smaller values indicate stronger alignment between representations.

Consistently small angular distances across adjacent layers indicate limited representational evolution, revealing redundancy across depth. In this work, we leverage this observation to motivate explicit alignment between shallow and deep transformer representations.

4 Method

4.1 Overall architecture

As described in Figure 1, given a raw speech waveform $\mathcal{X} \in \mathbb{R}$, sampled at 16 kHz, we first extract contextualized frame-level representations using a pretrained *XLS-R* encoder:

$$\mathbf{H} = \textit{XLS-R}(\mathcal{X}) \in \mathbb{R}^{T \times D}, \tag{8}$$

where $D = 1024$ is the hidden dimension.

To adapt these representations to the classifier, we apply a linear projection p followed by a SiLU nonlinearity:

$$\mathbf{H}^0 \leftarrow \mathrm{SiLU}(\mathbf{H}\mathbf{W}_p + b_p), \tag{9}$$

where $\mathbf{W}_p \in \mathbb{R}^{D \times d}$, $b_p \in \mathbb{R}^d$, and $d = 128$.

The projected features are processed by a stack of L transformer blocks equipped with MTLA. Let $\mathbf{H}^{(\ell)} \in \mathbb{R}^{T \times d}$ denote the output of the ℓ-th block. The stack is defined recursively as

$$\mathbf{H}^{(\ell)} = \mathcal{T}_\ell\big(\mathbf{H}^{(\ell-1)}\big), \quad \ell = 1, \dots, L, \tag{10}$$

with $\mathbf{H}^{(0)}$ as the projected *XLS-R* output.

Each transformer block $\mathcal{T}_\ell(\cdot)$ follows a pre-layer normalization (LN) with residual connections:

$$\tilde{\mathbf{H}}^{(\ell)} = \mathbf{H}^{(\ell-1)} + \mathrm{MTLA}\big(\mathrm{LN}(\mathbf{H}^{(\ell-1)})\big), \tag{11}$$

$$\mathbf{H}^{(\ell)} = \tilde{\mathbf{H}}^{(\ell)} + \mathrm{FFN}\big(\mathrm{LN}(\tilde{\mathbf{H}}^{(\ell)})\big), \tag{12}$$

where FFN denotes a two-layer feed-forward network with SiLU activation. We set $L \in \{1, \dots, 4\}$ to not increase the classifier module parameters.

To obtain a fixed-dimensional utterance-level embedding, we apply global average pooling over time:

$$\mathbf{z}^{(\ell)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{H}_t^{(\ell)} \in \mathbb{R}^d. \tag{13}$$

The final-layer representation $\mathbf{z}^{(L)}$ is passed to a linear classifier $\hat{c}$ to produce logits:

$$\hat{\mathbf{y}} = \mathbf{W}_{\hat{c}}\mathbf{z}^{(L)} + b_{\hat{c}}, \tag{14}$$

where $\mathbf{W}_{\hat{c}} \in \mathbb{R}^{d \times 2}$ and $b_{\hat{c}} \in \mathbb{R}^2$ and $\hat{\mathbf{y}}$ correspond to the bona fide and spoof classes.

4.2 Layer-wise representation alignment

To mitigate the underutilization of shallow layers and reduced generalization, we explicitly encourage intermediate layers to align with the task-specialized final representation. For each intermediate layer $\ell \in \{1, \ldots, L\}$, we compute the angular distance between its representation $\mathbf{z}^{(\ell)}$ and the final-layer representation $\mathbf{z}^{(L)}$:

$$a_\ell = d_{Ang}(\mathbf{z}^{(\ell)}, \mathbf{z}^{(L)}). \tag{15}$$

The layer-wise alignment loss is defined as

$$\mathcal{L}_{Ang} = \frac{1}{L} \sum_{k=1}^{L} a_\ell. \tag{16}$$

Minimizing this term encourages shallow and intermediate layers to produce representations that are geometrically aligned with the final task-optimized embedding, thereby improving feature consistency across depth.

The model is trained end-to-end using a composite loss combining standard cross-entropy $\mathcal{L}_{CE}$ and the proposed alignment regularization. For a binary classification task with $C = 2$ classes, let the ground-truth label be $y \in \{0, 1\}$, the classification loss is:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_{n,c} \log(\hat{y}_{n,c}) \tag{17}$$

with N samples. The final training objective is:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \, \mathcal{L}_{Ang}, \tag{18}$$

where $\alpha > 0$ controls the strength of the alignment regularization.

5 Experiments

5.1 Datasets and evaluation metric

Training and development. All models are trained on the `ASVspoof 2019` logical access (`19LA`) training set. Model selection is performed on the `19LA` development set, which includes a disjoint set of speakers. The `19LA` dataset comprises spoofed speech generated using TTS and VC techniques. Figure 2 sums up the in-domain statistics of training, development and evaluation.

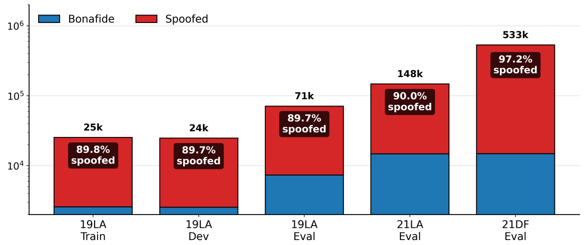

Figure 2: In-domain dataset statistics for training, development and evaluation.

In-domain evaluation. In-domain performance is evaluated on the `19LA` evaluation set, as well as the `ASVspoof 2021` logical access (`21LA`) and deepfake (`21DF`) partitions. The `21LA` dataset extends `19LA` by incorporating codec and transmission effects, while `21DF` further introduces diverse lossy compression artifacts. These datasets provide increasingly realistic test conditions while remaining within the `ASVspoof` benchmark framework.

Out-of-domain evaluation. To assess robustness and generalization beyond the training distribution, we evaluate our models on a diverse suite of out-of-domain datasets spanning in-the-wild recordings, diffusion- and vocoder-based synthesis, cross-lingual scenarios, and large-scale multilingual benchmarks. Detailed dataset descriptions and statistics are provided in Appendix A.

Performance metric. Following prior work, we evaluate system performance using the equal error rate (EER). The EER provides a threshold-independent summary of detection performance, where lower values indicate better discrimination between bona fide and spoofed speech. Detailed definitions are provided in Appendix B.

5.2 Implementation details

We employ the pretrained *XLS-R*[1] and fine-tune it during training. Audio inputs within each batch are dynamically padded to match the length of the longest utterance. To address the class imbalance in the `19LA` (Figure 2), we adopt a weighted $\mathcal{L}_{\text{CE}}$, assigning a higher weight to the bona fide class and a lower weight to the spoofed class.

To further improve robustness, we apply data augmentation (Tak et al., 2022a)[2], including linear and nonlinear convolutive noise, impulsive signal-dependent additive noise, stationary additive noise, and randomly colored noise. All hyperparameter settings are summarized in Table 3.

[1] https://huggingface.co/facebook/wav2vec2-xls-r-300m
[2] https://github.com/TakHemlata/RawBoost-antispoofing

Table 1: Overall performance (EER %). **Bold** indicates best results, the second-best are underlined. Angular-aligned layers are shaded gray and denoted by $\langle\cdot\rangle$.

#Blocks	Layer	In-domain			Out-of-domain		
		19LA	21LA	21DF	FOR	ITW	M-EN
1	$\mathcal{T}_1$	0.10	2.49	**1.42**	0.97	3.91	10.65
2	$\mathcal{T}_1$	0.12	2.15	1.81	7.73	4.00	8.82
	$\mathcal{T}_2$	0.12	1.44	1.88	6.62	3.86	7.34
	$\langle\mathcal{T}_1\rangle$	**0.08**	2.23	1.85	0.50	**3.36**	9.97
	$\langle\mathcal{T}_2\rangle$	0.09	1.62	1.84	**0.13**	3.85	8.54
3	$\mathcal{T}_1$	0.15	1.02	2.52	1.81	4.75	9.00
	$\mathcal{T}_2$	0.22	0.64	2.44	1.02	4.47	5.32
	$\mathcal{T}_3$	0.22	**0.63**	2.77	1.94	4.46	**4.98**
	$\langle\mathcal{T}_1\rangle$	0.16	2.72	1.80	0.63	4.20	9.71
	$\langle\mathcal{T}_2\rangle$	0.15	1.84	1.86	0.58	4.18	8.00
	$\langle\mathcal{T}_3\rangle$	0.16	1.20	2.13	0.93	4.35	7.38
4	$\mathcal{T}_1$	9.75	6.59	3.48	16.83	4.12	31.82
	$\mathcal{T}_2$	0.15	1.34	2.74	7.64	4.42	11.71
	$\mathcal{T}_3$	0.16	1.40	2.79	7.64	4.10	10.89
	$\mathcal{T}_4$	0.19	1.31	2.80	5.44	4.15	10.32
	$\langle\mathcal{T}_1\rangle$	2.80	5.23	2.86	1.68	3.83	15.67
	$\langle\mathcal{T}_2\rangle$	0.14	3.58	2.76	1.64	3.72	13.69
	$\langle\mathcal{T}_3\rangle$	0.13	2.27	2.72	1.55	3.71	11.79
	$\langle\mathcal{T}_4\rangle$	0.14	3.24	2.75	1.64	3.72	13.18

5.3 Experimental results

Representational redundancy across layers and domains. To investigate representational redundancy, the trained $\hat{c}$ on the output of $\mathbf{z}^{(L)}$ is applied without retraining to all intermediate layers across both in-domain (19LA, 21LA, 21DF) and out-of-domain (FOR, ITW, M-EN) datasets (Table 1). In the 2- and 3-block models, all $\mathcal{T}$ layers consistently achieve closely aligned EERs, with adjacent layers differing by less than 0.6 to 1 pp across both in- and out-of-domain datasets, indicating that task-relevant information becomes linearly separable early and remains largely preserved across depth. For example, in the three-block model, layers $\mathcal{T}_2$ and $\mathcal{T}_3$ exhibit nearly identical performance on 21LA (0.64% vs 0.63%) and on M-EN (5.32% vs 4.98%), demonstrating that redundancy emerges rapidly and is robust to domain shift. In contrast, the four-block model reveals a notable outlier in $\mathcal{T}_1$, which performs substantially worse than deeper layers (e.g., M-EN: 31.82% vs 11.71–10.32%), suggesting that very early layers are insufficiently aligned with the final decision boundary under both in- and out-of-domain conditions. Overall, these results indicate that redundancy emerges early in smaller models and persists across deeper layers, while extreme depth can introduce initial layers that are less domain-general, emphasizing the value of analyzing shallower configurations to reveal the intrinsic representational alignment of the transformer stack.

Impact of angular alignment on shallow layers. Table 1 also reports EERs for models trained using the $\mathcal{L}_{Ang}$ objective, which encourages internal representations to become more geometrically aligned with the final task-optimized embedding $\mathbf{z}^{(L)}$. Compared to the unaligned layers, shallow layers, particularly $\langle\mathcal{T}_1\rangle$ in the 2-, 3-, and 4-block models show substantially improved performance across both in-domain and out-of-domain datasets. For instance, in the 2-block model, $\langle\mathcal{T}_1\rangle$ decreases from 0.12% to 0.08% on 19LA, while in the 3-block model, $\langle\mathcal{T}_1\rangle$ improves from 2.52% to 1.80% on 21DF, showing increased consistency with deeper layers. Similarly, out-of-domain improvements are notable: in the 4-block model, $\langle\mathcal{T}_1\rangle$ drops on most datasets (FOR, ITW) and achieves better results across all layers, illustrating that d_{Ang} adjustment brings shallow layer representations much closer to $\mathbf{z}^{(L)}$. Overall, aligning shallow layers with the final block via d_{Ang} reduces discrepancies between layers, strengthens representational redundancy, and enables early layers to capture discriminative features that were previously only in deeper layers.

Ablation on angular alignment strength. Following the results in Table 1, we perform an ablation study on the 2-block model as it performs well and requires less parameters to evaluate the effect of angular alignment strength $\alpha \in \{0.1, 0.3, 0.5\}$ on layer-wise representations (Figure 4). For in-domain datasets, all α values maintain consistently low EERs across both layers ($\langle\mathcal{T}_1\rangle$, $\langle\mathcal{T}_2\rangle$). For out-of-domain performance, at $\alpha = 0.1$, EERs drop on FOR ($\langle\mathcal{T}_1\rangle$: 0.50%, $\langle\mathcal{T}_2\rangle$: 0.13%). Increasing α to 0.3 provides smaller gains (e.g., M-EN: 7.08–7.02%), while $\alpha = 0.5$ slightly degrades out-of-domain performance (12.36–12.29%), indicating that excessive alignment can over-constrain representations. Overall, these results highlight that moderate angular alignment suffices to unify shal-

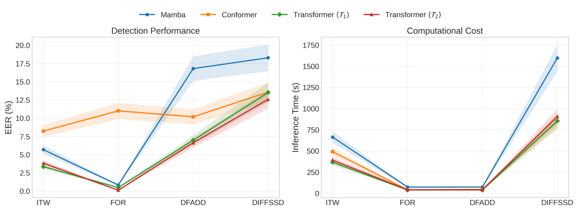

Figure 3: Performance (EER %) and inference computational cost of Mamba, Conformer, and Transformer $\langle\mathcal{T}_1\rangle$ and $\langle\mathcal{T}_2\rangle$ on out-of-domain datasets in terms of wall-clock time.

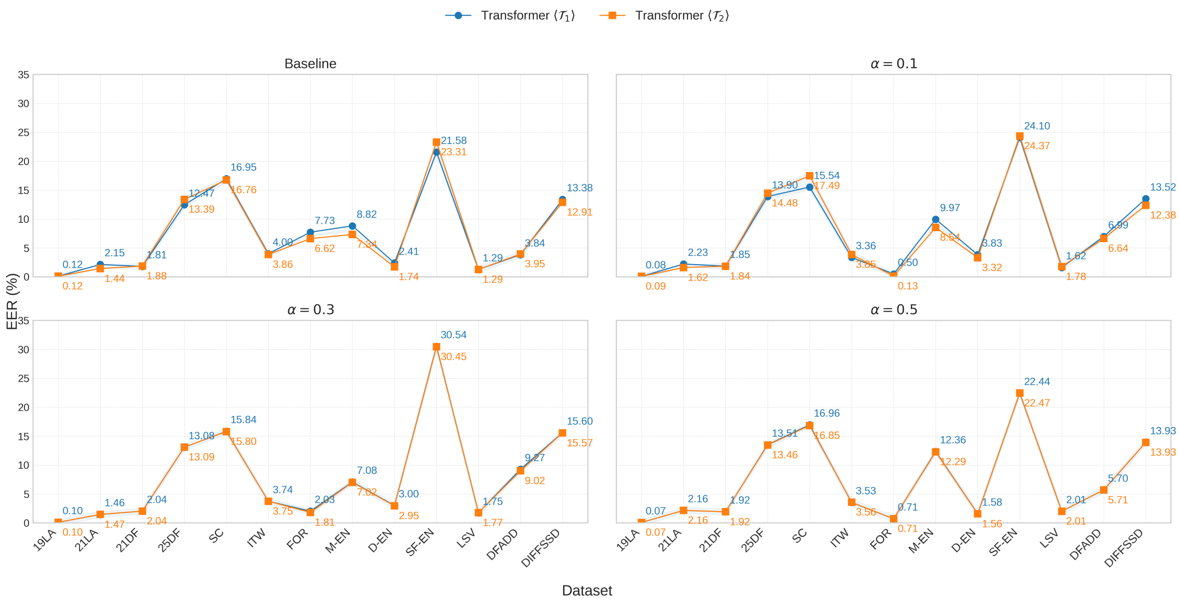

Figure 4: Effect of alignment strength $\alpha \in \{0.1, 0.3, 0.5\}$ on overall performance (EER %) of the 2-block Transformer across multiple datasets.

low and deep representations, enhancing redundancy and out-of-domain generalization.

Detection performance and efficiency gains. We compare the 2-block model with angular alignment $\alpha = 0.1$ against existing SOTA approaches with Conformer (Truong et al., 2024) and Mamba (Xiao and Das, 2025) on out-of-domain datasets (ITW, FOR, DFADD, DIFFSSD). As shown in Figure 3, our aligned shallow transformer layers ($\langle \mathcal{T}_1 \rangle$ and $\langle \mathcal{T}_2 \rangle$) achieve competitive or superior EERs compared to strong baselines. For instance, on ITW, $\langle \mathcal{T}_1 \rangle$ and $\langle \mathcal{T}_2 \rangle$ obtain 3.36% and 3.85% EER, outperforming Mamba (5.70%) and Conformer (8.24%). On FOR, $\langle \mathcal{T}_2 \rangle$ achieves 0.13% EER, a substantial improvement over all baselines. On diffusion-based datasets (DFADD and DIFFSSD), our models achieved 13.52% and 12.38%, respectively for $\langle \mathcal{T}_1 \rangle$ and $\langle \mathcal{T}_2 \rangle$, demonstrating the effective approach while detecting better diffusion- and flow-based artifacts.

In terms of efficiency, our 2-block model is significantly faster than deeper or more complex approaches. Comparing to Mamba and Conformer, $\langle \mathcal{T}_1 \rangle$ and $\langle \mathcal{T}_2 \rangle$ are faster across all evaluated datasets, while still maintaining better performance. This highlights that aligned shallow transformers provide a highly favorable trade-off between performance and inference efficiency.

Multilingual generalization. We further evaluate the robustness of our approach on the multilingual MLAAD benchmark, covering eight languages (M-EN, M-FR, M-IT, M-ES, M-PL, M-RU, M-UK, M-DE). As shown in Figure 5, the proposed shallow transformer ($\langle \mathcal{T}_1 \rangle$) demonstrates competitive performance against strong SOTA approaches across most languages, despite its substantially lower architectural depth with only 479.11K parameters. In particular, $\langle \mathcal{T}_1 \rangle$ outperforms both Conformer and Mamba on M-IT (5.69%), M-PL (8.38%), and M-RU (7.18%), indicating strong robustness to linguistic variability. On M-ES and M-FR, $\langle \mathcal{T}_1 \rangle$ remains competitive, achieving EERs close to the best-performing approach, while on M-EN it significantly improves over Conformer (9.97% vs. 14.35%). In contrast, $\langle \mathcal{T}_1 \rangle$ underperforms on M-DE. Overall, these results highlight that shallow, well-aligned transformer representations can generalize effectively across diverse languages, supporting their suitability for multilingual scenarios.

Overall comparison with state-of-the-art. Table 2 presents a comprehensive classifier module comparison between the proposed approach and representative SOTA anti-spoofing systems across both in-domain and out-of-domain benchmarks. Despite using substantially fewer parameters (479K), our method achieves performance that is competitive with or superior to significantly larger models. On in-domain datasets, our approach matches the best reported result on 19LA and remains competitive on 21LA and 21DF, demon-

Table 2: Overall performance (EER %). **Bold** indicates best results, the second-best are underlined. Angular-aligned layers are denoted by $\langle\cdot\rangle$. † denotes results reported from (Dowerah et al., 2025). ⋄ denotes results computed by using official released checkpoints. Reported parameter counts exclude the *XLS-R* (315M) backbone.

Model	#Params.	In-domain			Out-of-domain					Average
		`19LA`	`21LA`	`21DF`	`25DF`	`ITW`	`DFADD`	`LSV`	`SC`	
SLS (Zhang et al., 2024)†⋄	23.40M	0.23	2.87	1.91	18.76	7.46	7.54	1.97	24.51	8.16
Conformer (Truong et al., 2024)†⋄	3.82M	0.19	1.03	2.06	18.85	7.79	8.89	2.35	38.15	9.91
MultiConv (Tran et al., 2025c)⋄	2.64M	**0.08**	2.77	1.43	15.19	4.44	**6.60**	1.70	18.71	6.37
KAN (Tran et al., 2025a)⋄	2.11M	0.11	2.29	1.49	14.10	5.31	15.63	1.52	22.06	7.81
Mamba (Tran et al., 2025b)⋄	2.08M	0.11	1.78	1.51	**13.58**	5.12	8.62	1.82	17.87	6.30
Mamba (Xiao and Das, 2025)†⋄	1.94M	0.42	0.93	1.88	14.40	6.71	10.70	2.23	27.58	8.10
Nes2NetX (Liu et al., 2025)†⋄	512.04K	0.12	1.88	1.49	22.06	5.52	11.15	2.88	54.28	12.42
AASIST (Tak et al., 2022b)†⋄	447.24K	0.22	**0.82**	2.85	16.25	11.20	11.93	11.21	26.72	10.15
RASA (Yang et al., 2025b)	–	0.09	0.89	**1.24**	–	4.74	–	–	–	–
Poin-HierNet (Yang et al., 2025a)	–	0.11	0.94	1.40	–	4.91	–	–	–	–
Transformer $\mathcal{T}_1$ (Ours)	479.11K	0.10	2.49	1.42	13.97	3.91	7.14	**1.10**	19.09	6.15
Transformer $\langle\mathcal{T}_1\rangle$ (Ours)	479.11K	**0.08**	2.23	1.85	13.90	**3.36**	6.99	1.62	**15.54**	**5.70**

strating that shallow, aligned representations are sufficient to capture discriminative spoofing cues.

More notably, our method consistently excels in out-of-domain scenarios, where robustness to distribution shift is critical. On `ITW`, our approach achieves the lowest EER (3.36%), outperforming all compared methods, including larger architectures such as Mamba (Xiao and Das, 2025) and Conformer (Truong et al., 2024). Similarly, strong generalization is observed on `SC` (15.54%), `DFADD` (6.99%), and `LSV` (1.62%), where our model ranks among the top-performing approaches despite its compact size. Compared to recent parameter-efficient methods (e.g., Nes2NetX (Liu et al., 2025), AASIST (Tak et al., 2022b)), our approach consistently yields lower EERs across most datasets, highlighting the effectiveness of angular alignment in improving cross-domain robustness.

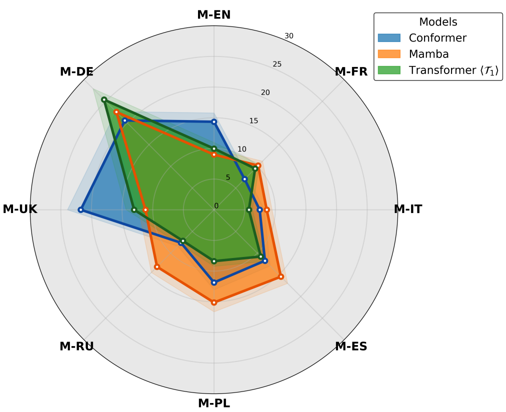

Figure 5: Performance (EER %) Mamba, Conformer, and Transformer $\langle\mathcal{T}_1\rangle$ on `MLAAD` dataset.

Overall, these results indicate that carefully aligned shallow transformer representations can rival or surpass deeper and more complex models, offering a favorable trade-off between robustness and model efficiency. This supports the central claim that representational redundancy across transformer layers can be exploited to design lightweight yet highly effective anti-spoofing CMs.

6 Conclusion

We presented a systematic analysis of representational redundancy in transformer-based anti-spoofing models and demonstrated that discriminative information for deepfake detection emerges early and is largely preserved across depth. By training a linear classifier on the final transformer block and reusing it across intermediate layers in classifier module, we showed that adjacent layers, particularly in shallow configurations exhibit highly aligned representations under both in-domain and out-of-domain conditions. Building on this observation, we introduced an angular alignment strategy that explicitly reduces representational misalignment between shallow layers and the final transformer representation. Extensive experiments across diverse benchmarks, including multilingual and real-world datasets, demonstrate that our approach achieves competitive or superior performance compared to state-of-the-art systems while using substantially fewer parameters. These results highlight that depth alone is not the primary driver of robustness in anti-spoofing, and that exploiting representational redundancy enables the design of efficient and effective detection models.

Limitations

Despite promising results, our study has several limitations. First, the analysis is restricted to transformer-based architectures in classifier module and not the pretrained speech foundation model. Second, model families such as state-space model can be explored. Third, although we evaluate across a wide range of datasets, real-world spoofing attacks continue to evolve, and performance on unseen future attacks cannot be guaranteed. Finally, angular alignment is controlled by a fixed hyperparameter, which may require tuning when transferring to new domains or datasets.

Ethical considerations

Deepfake detection technologies have important societal implications, including protecting individuals from fraud, misinformation, and identity misuse. Our work aims to improve the robustness and efficiency of such systems, facilitating broader deployment in real-world scenarios. However, like all detection methods, our approach may be imperfect and could produce false positives or false negatives, potentially leading to unintended consequences if used in high-stakes decision-making without human oversight. Moreover, publishing detailed analyses of model behavior may indirectly inform adversaries; we therefore emphasize that our contributions are intended to strengthen defensive capabilities rather than enable misuse. We encourage responsible deployment, transparency in system limitations, and continued evaluation to ensure fair and ethical use.

References

Takanori Ashihara, Marc Delcroix, Takafumi Moriya, Kohei Matsuura, Taichi Asami, and Yusuke Ijima. 2024. What do self-supervised speech and speaker models learn? new findings from a cross model layer-wise analysis. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10166–10170.

Zhongjie Ba, Qing Wen, Peng Cheng, Yuwei Wang, Feng Lin, Li Lu, and Zhenguang Liu. 2023. Transferring audio deepfake detection capability across languages. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2033–2044, New York, NY, USA. Association for Computing Machinery.

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. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Kratika Bhagtani, Amit Kumar Singh Yadav, Paolo Bestagini, and Edward J. Delp. 2025. Diffssd: A diffusion-based dataset for speech forensics. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271, Vienna, Austria. Association for Computational Linguistics.

Phuong Tuan Dat and Tran Huy Dat. 2025. Xlsr-kanformer: A kan-intergrated model for synthetic speech detection. In *2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS)*, pages 1–6.

Keqi Deng and Phil Woodland. 2025. Multi-head temporal latent attention. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Teresa Dorszewski, Albert Kjøller Jacobsen, Lenka Tětková, and Lars Kai Hansen. 2025. How redundant is the transformer stack in speech representation models? In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Sandipana Dowerah, Atharva Kulkarni, Ajinkya Kulkarni, Hoan My Tran, Joonas Kalda, Artem Fedorchenko, Benoit Fauve, Damien Lolive, Tanel Alumäe, and Matthew Magimai Doss. 2025. Speech df arena: A leaderboard for speech deepfake detection models. *arXiv preprint arXiv:2509.02859*.

Jiawei Du, I-Ming Lin, I-Hsiang Chiu, Xuanjun Chen, Haibin Wu, Wenze Ren, Yu Tsao, Hung-Yi Lee, and Jyh-Shing Roger Jang. 2024. Dfadd: The diffusion and flow-matching based audio deepfake dataset. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 921–928.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 682–689.

Femi Folorunsho and Benedicta Frema Boamah. 2025. Deepfake technology and its impact: ethical considerations, societal disruptions, and security threats in ai-generated media. *International journal of information technology and management information systems*, 16(1):1060–1080.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Dan Roberts. 2025. The unreasonable ineffectiveness of the deeper layers. In *The Thirteenth International Conference on Learning Representations*.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*.

Yu Guan, Wu Guo, Jie Zhang, and Zhijun Zhang. 2025. Fusing multi-layer features of the pre-trained model with grouped cross attention for spoofing speech detection. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 601–606.

Yunqi Hao, Minqiang Xu, Yihao Chen, Yanyan Liu, Liang He, Lei Fang, and Lin Liu. 2025. Integrating spectro-temporal cross aggregation and multiscale dynamic learning for audio deepfake detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, and Yanmin Qian. 2025. SpeechFake: A large-scale multilingual speech deepfake dataset incorporating cutting-edge generation methods. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9985–9998, Vienna, Austria. Association for Computational Linguistics.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Jiachen Jiang, Jinxin Zhou, and Zhihui Zhu. 2025. Tracing representation progression: Analyzing and enhancing layer-wise similarity. In *The Thirteenth International Conference on Learning Representations*.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and sheng zhao. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Forty-first International Conference on Machine Learning*.

Jee-weon Jung, Yihan Wu, Xin Wang, Ji-Hoon Kim, Soumi Maiti, Yuta Matsunaga, Hye-jin Shim, Jinchuan Tian, Nicholas Evans, Joon Son Chung, Wangyou Zhang, Seyun Um, Shinnosuke Takamichi, and Shinji Watanabe. 2025. Spoofceleb: Speech deepfake detection and sasv in the wild. *IEEE Open Journal of Signal Processing*, 6:68–77.

Taewoo Kim, Guisik Kim, Choongsang Cho, and Young Han Lee. 2025. Naturalness-Aware Curriculum Learning with Dynamic Temperature for Speech Deepfake Detection. In *Interspeech 2025*, pages 5318–5322.

Howard Lei and Eduardo Lopez. 2009. Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. In *Interspeech 2009*, pages 2323–2326.

Xinfeng Li, Kai Li, Yifan Zheng, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. 2024. Safeear: Content privacy-preserving audio deepfake detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, page 3585–3599, New York, NY, USA. Association for Computing Machinery.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, and 1 others. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Tianchi Liu, Duc-Tuan Truong, Rohan Kumar Das, Kong Aik Lee, and Haizhou Li. 2025. Nes2net: A lightweight nested architecture for foundation model driven speech anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 20:12005–12018.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and 1 others. 2023. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522.

Joao Medeiros. 2015. How intel gave stephen hawking a voice. *Wired Retrieved from https://www. wired. com/2015/01/intel-gave-stephen-hawking-voice*.

Nicolas Müller, Pavel Czempin, Franziska Diekmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does audio deepfake detection generalize? In *Interspeech 2022*, pages 2783–2787.

Nicolas M. Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. 2024. Mlaad: The multi-language audio anti-spoofing dataset. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Vardan Papyan, XY Han, and David L Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Tuan Dat Phuong, Long-Vu Hoang, and Huy Dat Tran. 2025. Pushing the Performance of Synthetic Speech Detection with Kolmogorov-Arnold Networks and Self-Supervised Learning Models. In *Interspeech 2025*, pages 5633–5637.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pages 2757–2761.

Ricardo Reimao and Vassilios Tzerpos. 2019. For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–10.

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.

Peter Súkeník, Christoph H. Lampert, and Marco Mondelli. 2025. Neural collapse is globally optimal in deep regularized resnets and transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. 2023. Ai-synthesized voice detection using neural vocoder artifacts. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 904–912.

Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2022a. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6382–6386. IEEE.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022b. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop (Odyssey 2022)*, pages 112–119.

Pablo Andrés Tamayo Flórez, Rubén Manrique, and Bernardo Pereira Nunes. 2023. Habla: A dataset of latin american spanish accents for voice anti-spoofing. In *Interspeech 2023*, pages 1963–1967.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. 2017. Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 45:516–535.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi H. Kinnunen, and Kong Aik Lee. 2019. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In *Interspeech 2019*, pages 1008–1012.

Hoan My Tran, Damien Lolive, David Guennec, Aghilas Sini, Arnaud Delhay, and Pierre-François Marteau. 2025a. Kan you hear the truth? audio deepfake detection with kolmogorov–arnold networks. In *2025 33rd European Signal Processing Conference (EUSIPCO)*, pages 561–565.

Hoan My Tran, Damien Lolive, David Guennec, Aghilas Sini, Arnaud Delhay, and Pierre-François Marteau. 2025b. Leveraging SSL Speech Features and Mamba for Enhanced DeepFake Detection. In *Interspeech 2025*, pages 5323–5327.

Hoan My Tran, Damien Lolive, Aghilas Sini, Arnaud Delhay, Pierre-François Marteau, and David Guennec. 2025c. Multi-level ssl feature gating for audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11766–11775, New York, NY, USA. Association for Computing Machinery.

Duc-Tuan Truong, Tianchi Liu, Junjie Li, Ruijie Tao, Kong Aik Lee, and Eng Siong Chng. 2025a. Addressing gradient misalignment in data-augmented training for robust speech deepfake detection. *arXiv preprint arXiv:2509.20682*.

Duc-Tuan Truong, Tianchi Liu, Ruijie Tao, Junjie Li, Kong Aik Lee, and Eng Siong Chng. 2025b.

Qamo: Quality-aware multi-centroid one-class learning for speech deepfake detection. *arXiv preprint arXiv:2509.20679*.

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*, pages 537–541.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, Junichi Yamagishi, Myeonghun Jeong, Ge Zhu, Yongyi Zang, You Zhang, Soumi Maiti, Florian Lux, and 10 others. 2026. Asvspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Computer Speech & Language*, 95:101825.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi H. Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi. 2024. Asvspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 1–8.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, and 1 others. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

Yang Xiao and Rohan Kumar Das. 2025. Xlsr-mamba: A dual-column bidirectional state space model for spoofing attack detection. *IEEE Signal Processing Letters*, 32:1276–1280.

Xi Xuan, Zimo Zhu, Wenxin Zhang, Yi-Cheng Lin, and Tomi Kinnunen. 2025. Fake-mamba: Real-time speech deepfake detection using bidirectional mamba as self-attention's alternative. *arXiv preprint arXiv:2508.09294*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive:(http://web. ku. edu/~ idea/readings/rainbow. htm)*.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 47–54.

Mingru Yang, Yanmei Gu, Qianhua He, Yanxiong Li, Peirong Zhang, Yongqiang Chen, Zhiming Wang, Huijia Zhu, Jian Liu, and Weiqiang Wang. 2025a. Generalizable Audio Deepfake Detection via Hierarchical Structure Learning and Feature Whitening in Poincaré sphere. In *Interspeech 2025*, pages 2255–2259.

Mingru Yang, Yanmei Gu, Qianhua He, Peirong Zhang, Haolin He, Zhiming Wang, Huijia Zhu, Jian Liu, and Weiqiang Wang. 2025b. Generalizable audio deepfake detection via risk-aware style alignment and structural empirical risk minimization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11600–11609, New York, NY, USA. Association for Computing Machinery.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, and 1 others. 2023. Add 2023: the second audio deepfake detection challenge. *arXiv preprint arXiv:2305.13774*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech 2019*, pages 1526–1530.

Qishan Zhang, Shuangbing Wen, and Tao Hu. 2024. Audio deepfake detection with self-supervised XLS-r and SLS classifier. In *ACM Multimedia 2024*.

Zhenyu Zhang, Yewei Gu, Xiaowei Yi, and Xianfeng Zhao. 2021. Fmfcc-a: a challenging mandarin dataset for synthetic speech detection. In *International Workshop on Digital Watermarking*, pages 117–131. Springer.

A Dataset details

In this section, we provide the information about different datasets used in this study.

A.1 ASVspoof 2019

The ASVspoof 2019 logical access (`19LA`[3]) scenario (Wang et al., 2020) is a benchmark dataset designed to evaluate CMs against spoofing attacks on ASV systems. The `19LA` focuses specifically on LA attacks, where spoofed speech is generated using advanced TTS and VC techniques. It includes bona fide speech from the `VCTK` corpus (Yamagishi et al., 2019) alongside spoofed utterances created with 17 different attack methods (6 known in training/development sets and 13 unknown in

[3] https://doi.org/10.7488/ds/2555

evaluation), promoting research into robust detection of synthesized and converted speech that can deceive ASV systems. The dataset has become a standard for assessing generalization to unseen spoofing conditions in controlled environments.

A.2 ASVspoof 2021

ASVspoof 2021 (Liu et al., 2023) logical access (`21LA`[4]) scenario builds upon the `19LA` task to evaluate CMs against spoofing attacks in more realistic scenarios. It includes evaluation data with various codec encodings, transmission effects, and channel variability to simulate communication over telephony and VoIP networks. ASVspoof 2021 deepfake (`21DF`[5]) scenario shifts focus toward general audio deepfake detection beyond ASV-specific vulnerabilities. The dataset comprises bona fide speech and spoofed utterances processed through a variety of lossy codecs with different configurations to introduce compression artifacts.

A.3 ASVspoof 5

ASVspoof 5 (Wang et al., 2026) (`25DF`[6]) is the fifth edition of the ASVspoof challenge series, introducing a significantly more challenging and realistic database for evaluating spoofing CMs and deepfake detection in the context of ASV systems. Unlike previous editions that relied on studio-quality recordings from limited speakers, ASVspoof 5 utilizes crowdsourced bona fide speech from approximately 2,000 speakers in diverse acoustic conditions, primarily derived from the Multilingual Librispeech (`MLS`) English dataset (Pratap et al., 2020). Spoofed utterances incorporate advanced TTS/VC methods and novel adversarial attacks, with evaluation data including unseen codecs and compression artifacts to promote generalization to in-the-wild conditions.

A.4 SpoofCeleb

The SpoofCeleb (`SC`[7]) dataset (Jung et al., 2025) is a large-scale benchmark for speech deepfake detection, featuring real-world, in-the-wild bona fide speech from 1,251 unique celebrities sourced from the processed `VoxCeleb1` corpus through an automated pipeline involving transcription, segmentation, noise reduction, and quality filtering. It includes over 2.5 million utterances, with spoofed samples generated using 23 TTS systems, creating diverse synthetic attacks. The evaluation set comprises 40 speakers, 9 attacks emphasizing noisy, reverberant conditions and high speaker diversity to better simulate real-world deepfakes, addressing generalization challenges and supporting advanced CMs against sophisticated TTS-generated spoofs.

A.5 In-the-wild

The In-the-wild (`ITW`[8]) dataset (Müller et al., 2022) comprises approximately 37.9 hours of short audio clips from 58 English-speaking celebrities and politicians. It includes 20.7 hours of bona fide speech sourced from public podcasts and speeches, paired with 17.2 hours of spoofed speech derived from 219 publicly available audio files. Clips are carefully matched for style, emotion, background noise, and duration to reflect realistic conditions. This `ITW` dataset captures diverse, unseen TTS algorithms, social media artifacts, compression, and environmental variability, serving as a challenging out-of-domain benchmark for evaluating the generalization of spoofing CMs and audio deepfake detectors, where ASVspoof-trained models often exhibit severe performance degradation.

A.6 Fake-or-Real

The Fake-or-Real (`FoR`[9]) dataset (Reimao and Tzerpos, 2019) comprises over 195,000 English utterances with approximately 87,000 synthetic and 111,000 bona fide, generated using advanced commercial and research TTS systems. Bona fide speech is sourced from diverse open datasets and real-world recordings, incorporating variability in speakers, genders, accents, ages, microphones, and acoustic conditions. The dataset offers multiple versions, including `FOR`-original (raw full utterances), `FOR`-norm (normalized and balanced), `FOR`-2seconds (truncated to 2 seconds), and `FOR`-rerecorded (re-recorded to simulate real-world playback attacks with noise and channel effects). In this study, we evaluate the generalizability of our models on the original version.

A.7 Multi-language audio anti-spoofing dataset

The multi-language audio anti-spoofing dataset (Müller et al., 2024) (`MLAAD`[10]) is a large-scale multilingual benchmark for audio

[4] https://doi.org/10.5281/zenodo.4817650
[5] https://doi.org/10.5281/zenodo.4835107
[6] https://doi.org/10.5281/zenodo.14498691
[7] https://huggingface.co/datasets/jungjee/spoofceleb
[8] https://deepfake-total.com/in_the_wild
[9] https://bil.eecs.yorku.ca/datasets
[10] https://deepfake-total.com/mlaad

deepfake detection and spoofing CMs, designed to mitigate language bias prevalent in prior datasets and enhance model generalization across diverse linguistic and synthesis conditions. It comprises over 570 hours of synthesized spoofed speech with approximately 243,000 utterances, generated using 119 diverse TTS models, applied to bona fide speech sourced from the M-AILABS dataset in 8 original languages (English (M-EN), French (M-FR), German (M-DU), Italian (M-IT), Polish M-PL, Russian (M-RU), Spanish (M-ES), Ukrainian (M-UK)) and extended to 40 languages via translation. Audio is provided at 22,050 Hz, with evaluation augmentations incorporating noise, music, and various codecs to simulate real-world variability. In contrast to predominantly English-focused datasets, MLAAD's extensive multilingual scope and modern TTS diversity enable superior cross-dataset performance, promoting robust detectors resilient to unseen languages, synthesis methods, and in-the-wild deepfakes.

A.8 Deepfake cross-lingual evaluation

The deepfake cross-lingual evaluation (DC[11]) dataset (Ba et al., 2023) is a bilingual benchmark specifically designed to assess the cross-lingual generalization of audio deepfake detection systems by evaluating the impact of language differences on detector performance. It consists of two subsets: DC-E in English and DC-C in Chinese. Bona fide speech in the Chinese subset is sourced from the FMFCC-A dataset (Zhang et al., 2021), while corresponding English utterances are synthesized via online TTS APIs. Spoofed samples are generated using a variety of modern TTS and VC systems, including commercial ones. DC emphasizes cross-lingual transfer, typically training on English and evaluating on Chinese, to highlight language-specific vulnerabilities and promote robust detectors capable of handling unseen linguistic variations in real-world deepfake scenarios.

A.9 SpeechFake

The SpeechFake (SF[12]) dataset (Huang et al., 2025) comprises over 3 million deepfake utterances totaling more than 3,000 hours of audio, divided into a bilingual dataset focusing on English (SF-EN) and Chinese (SF-CH) and a multilingual dataset spanning 46 languages. Spoofed samples are generated using 40 diverse tools including 30 open-source and 10 commercial APIs, covering TTS, VC, and neural vocoders. SpeechFake incorporates cutting-edge generation methods, extensive multilingual coverage, and speaker variability to better simulate real-world deepfakes and address cross-lingual generalization challenges.

A.10 Librisevoc

The LibriSeVoc (LSV[13]) dataset (Sun et al., 2023), introduced in 2023, is a specialized benchmark for audio deepfake and synthesized speech detection, with a primary focus on identifying artifacts introduced by modern neural vocoders rather than full TTS or VC pipelines. Derived from the clean, multi-speaker LibriTTS corpus (Zen et al., 2019), it comprises 92,407 utterances totaling approximately 244 hours at 24 kHz sampling rate, including 13,201 bona fide samples and 79,206 spoofed samples generated via self-vocoding-resynthesizing the original waveforms using six diverse neural vocoders representing major architectures: autoregressive, diffusion-based, and GAN-based.

A.11 Diffusion and flow-matching based audio deepfake dataset

The diffusion and flow-matching based audio deepfake dataset (Du et al., 2024) (DFADD[14]) is a specialized benchmark focusing on highly natural synthetic speech generated by diffusion and flow-matching TTS models that pose significant challenges to existing CMs. Built upon the VCTK, it comprises approximately 180 hours of audio, including 44,455 bona fide utterances and 163,500 spoofed audios.

A.12 Diffusion-based synthetic speech dataset

The diffusion-based synthetic speech dataset (DIFFSSD[15]) (Bhagtani et al., 2025) is a benchmark addressing the generalization failures of CMs trained on controlled datasets when faced with highly natural speech from modern diffusion-based TTS systems. It comprises approximately 200 hours of audio, including bona fide utterances sourced from the LJ Speech (Ito and Johnson, 2017) and LibriSpeech (Panayotov et al., 2015) corpora across 74 speakers, paired with 70,000 spoofed utterances generated using 10 advanced

[11]https://doi.org/10.5281/zenodo.7601506
[12]https://github.com/YMLLG/SpeechFake
[13]https://github.com/csun22/Synthetic-Voice-Detection-Vocoder-Artifacts
[14]https://github.com/isjwdu/DFADD
[15]https://huggingface.co/datasets/purdueviperlab/diffssd

TTS systems with 8 open-source diffusion-based and 2 commercial.

A.13 Audio Deepfake Detection 2023

The audio deepfake detection (ADD[16]) (Yi et al., 2023) track 1.2 dataset, serves as the evaluation benchmark for the detection sub-task in an adversarial "fake game" setting, where systems must identify spoofed utterances, including those crafted to evade detectors. The evaluation features two rounds: Round 1 (ADD-R1) with 111,976 test utterances (80,000 bona fide and 31,976 spoofed via TTS/VC and partial Track 1.1 submissions); Round 2 (ADD-R2) with 118,477 test utterances (87,500 bona fide, 30,977 spoofed, incorporating more adversarial generations). In contrast to ASVspoof datasets (focused on controlled English spoofing with known/unknown attacks), ADD Track 1.2 emphasizes large-scale Mandarin speech, real-world diversity in test sets, and adversarial robustness through integration of participant-generated fakes, better simulating in-the-wild deepfake threats.

A.14 Latin American Spanish accents datasets

The HABLA[17] dataset (Tamayo Flórez et al., 2023), is the first dedicated voice anti-spoofing benchmark in Spanish, emphasizing Latin American accents (Argentinian, Chilean, Colombian, Peruvian, and Venezuelan) from 162 speakers (male and female). It comprises 22,816 bona fide utterances and approximately 58,000 spoofed utterances generated using six modern methods including voice conversion systems with cross-accent. In contrast to predominantly English-centric datasets like ASVspoof, HABLA highlights accent diversity and contemporary VC/TTS methods to better evaluate cross-lingual generalization and robustness of spoofing CMs against regional variations in real-world deepfakes.

B Evaluation metric

The EER corresponds to the operating point at which the false acceptance (FA) rate equals the false rejection (FR) rate. In the context of spoofing CMs, both error rates are defined as functions of a decision threshold τ_{CM} applied to the detection score.

[16] http://addchallenge.cn/databases2023

[17] https://doi.org/10.5281/zenodo.7370804

The false acceptance rate is defined as

$$P_{\mathrm{fa}}^{\mathrm{CM}}(\tau_{\mathrm{CM}}) = \frac{\#\{\text{spoofed with scores} > \tau_{\mathrm{CM}}\}}{\#\{\text{spoofed trials}\}}, \tag{19}$$

while the false rejection (miss) rate is given by

$$P_{\mathrm{miss}}^{\mathrm{CM}}(\tau_{\mathrm{CM}}) = \frac{\#\text{bona fide with scores} \leq \tau_{\mathrm{CM}}}{\#\{\text{bona fide trials}\}}. \tag{20}$$

A false acceptance occurs when a spoofed utterance is incorrectly classified as bona fide, whereas a false rejection occurs when a bona fide utterance is incorrectly rejected. The EER is obtained by sweeping the threshold τ_{CM} and identifying the operating point at which

$$P_{\mathrm{fa}}^{\mathrm{CM}}(\tau_{\mathrm{CM}}) = P_{\mathrm{miss}}^{\mathrm{CM}}(\tau_{\mathrm{CM}}). \tag{21}$$

Lower EER values indicate stronger spoofing detection performance.

In our detection framework, the model outputs confidence scores for the bona fide and spoofed hypotheses. We compute the final detection score as a log-likelihood ratio (LLR):

$$\mathrm{LLR}_t = \log p(X_t \mid \mathcal{H}_0) - \log p(X_t \mid \mathcal{H}_1), \tag{22}$$

where X_t denotes the input utterance of trial t. The null hypothesis $\mathcal{H}_0$ corresponds to bona fide speech, while the alternative hypothesis $\mathcal{H}_1$ corresponds to spoofed speech. The LLR score is used to sweep the decision threshold τ_{CM} for EER computation.

C Additional ablation study results

C.1 English datasets

Diffusion-based speech synthesis. Diffusion- and flow-matching–dominated datasets (DFADD, DIFFSSD) consistently reveal the limitations of deeper architectures. On DFADD, shallow configurations perform best, with the 2-block model achieving the lowest EER (3.84–3.95%). In contrast, increasing depth degrades performance: 4-block models exhibit substantially higher EERs (10–13% on DFADD and 17–18% on DIFFSSD). Notably, intermediate layers ($\mathcal{T}_2$, $\mathcal{T}_3$) outperform the final layer ($\mathcal{T}_4$), indicating that the deepest transformer block tends to overfit and generalizes poorly.

These results suggest that additional depth does not enhance sensitivity to diffusion-induced artifacts, which are typically subtle, globally coherent, and weakly correlated with localized temporal cues. Instead, deeper stacks appear to dilute discriminative information through redundant transformations.

While angular alignment slightly degrades performance across all configurations, the proposed models remain competitive with SOTA approaches.

Vocoder artifact detection. On vocoder-based datasets such as LSV, shallow architectures exhibit strong robustness, achieving EERs of 1.10% with a 1-block model and 1.29% with a 2-block model. In contrast, the 3-block configuration shows clear overfitting on training dataset and correspondingly degraded performance. Angular-aligned variants yield comparable but slightly less stable results, with EERs ranging between 1.0% and 2.0%. Overall, these findings indicate that vocoder artifacts are predominantly captured by early-to-mid-level representations, while additional depth provides limited marginal benefit for detecting vocoder-specific inconsistencies.

Cross-domain performance. We evaluate robustness on cross-domain datasets (25DF, M-EN, D-EN, SF-EN). Across configurations, D-EN exhibits low EERs (<5%), with performance generally improving as depth increases, except for the 3-block model. Angular-aligned variants further reduce EERs under stronger alignment. On 25DF, all configurations achieve EERs below 15%, despite the presence of adversarially generated samples unseen during training. In contrast, SF-EN, which combines domain mismatch with advanced, unseen TTS, VC, and neural vocoder samples, remains challenging, with EERs exceeding 20% across all models. For M-EN, the 3-block model attains the best in-domain performance (4.98–5.32%) but fails to generalize to other datasets, whereas angular-aligned variants exhibit slightly worse on M-EN but improved other cross-dataset generalization.

C.2 Analysis on cross-lingual benchmarks.

Table 5 summarizes performance across a diverse set of non-English and cross-lingual benchmarks, including Mandarin adversarial data (ADD-R1/R2, SF-CH, D-CH), multilingual European languages (MLAAD), and accent- and language-specific datasets (HABLA). Several consistent trends emerge. First, shallow architectures exhibit markedly stronger robustness: models with two or three transformer blocks substantially outperform deeper configurations on average, with the 2-block model achieving the lowest pooled EER (23.35%) among non-angular variants. In contrast, increasing depth to four blocks leads to a pronounced degradation, particularly in MLAAD languages (e.g., M-DE, M-FR) and accent-rich HABLA, indicating that excessive depth amplifies language- and accent-specific biases rather than improving generalization. Second, intermediate layers ($\mathcal{T}_2$ and $\mathcal{T}_3$) consistently outperform the first block and closely match or exceed the final block, reinforcing the observation that task-relevant, language-agnostic cues are encoded at intermediate depths rather than progressively refined at later stages. Third, angular supervision ($\alpha = 0.1$) systematically improves cross-lingual robustness for shallow and mid-depth models, reducing average EERs by up to 1.0–1.5 absolute points for 2- and 3-block configurations, while simultaneously narrowing performance gaps across languages. Notably, angular constraints mitigate but do not fully eliminate the degradation observed in deeper models, suggesting that representational over-specialization rather than classifier misalignment is the primary failure mode at larger depths. Overall, these results indicate that compact transformer stacks with moderate depth and angular regularization provide a more favorable bias–variance trade-off for non-English and cross-lingual audio deepfake detection, yielding representations that are both discriminative and resilient to linguistic, phonetic, and accentual variation.

Configurations	**Setting**
Batch size	5
Epochs	10
GPUs	1 NVIDIA GeForce RTX 4090
Optimizer	Adam
Learning rate	2.5×10^{-6}
Weight decay	$1e^{-4}$
Weighted cross-entropy loss	0.9 for real, 0.1 for fake
Early-stop patience	3
Data augmentation	RawBoost
Model architecture	**Parameters**
XLS-R feature extractor	315.44M
Feature projection	131.20K
Transformer	
1 block	479.11K
2 blocks	826.75K
3 blocks	1.17M
4 blocks	1.52M
Classification head	258

Table 3: Hyperparameters and architecture details of the models.

Table 4: Performance (EER %) comparison on English datasets. Reported parameter counts exclude the *XLS-R* (315M) backbone.

Layer	#Params	Dataset													Average	Pooled
		In-domain			Out-of-domain											
		19LA	21LA	21DF	25DF	SC	ITW	FOR	M-EN	D-EN	SF-EN	LSV	DFADD	DIFFSSD		
1 block																
$\mathcal{T}_1$	479.11K	0.10	2.49	1.42	13.97	19.09	3.91	0.97	10.65	2.37	22.25	1.10	7.14	10.83	7.41	13.92
2 blocks																
$\mathcal{T}_1$	479.11K	0.12	2.15	1.81	12.47	16.95	4.00	7.73	8.82	2.41	21.58	1.29	3.84	13.38	7.43	13.95
$\mathcal{T}_2$	826.75K	0.12	1.44	1.88	13.39	16.76	3.86	6.62	7.34	1.74	23.31	1.29	3.95	12.91	7.28	13.84
2 Blocks + Angular $\alpha = 0.1$																
$\langle\mathcal{T}_1\rangle$	479.11K	0.08	2.23	1.85	13.90	15.54	3.36	0.50	9.97	3.83	24.10	1.62	6.99	13.52	7.50	13.94
$\langle\mathcal{T}_2\rangle$	826.75K	0.09	1.62	1.84	14.48	17.49	3.85	0.13	8.54	3.32	24.37	1.78	6.64	12.38	7.42	14.43
2 Blocks + Angular $\alpha = 0.3$																
$\langle\mathcal{T}_1\rangle$	479.11K	0.10	1.46	2.04	13.08	15.84	3.74	2.03	7.08	3.00	30.54	1.75	9.27	15.60	8.12	14.24
$\langle\mathcal{T}_2\rangle$	826.75K	0.10	1.47	2.04	13.09	15.80	3.75	1.81	7.02	2.95	30.45	1.77	9.02	15.57	8.07	14.24
2 Blocks + Angular $\alpha = 0.5$																
$\langle\mathcal{T}_1\rangle$	479.11K	0.07	2.16	1.92	13.51	16.96	3.53	0.71	12.36	1.58	22.44	2.01	5.70	13.93	7.45	13.75
$\langle\mathcal{T}_2\rangle$	826.75K	0.07	2.16	1.92	13.46	16.85	3.56	0.71	12.29	1.56	22.47	2.01	5.71	13.93	7.44	13.76
3 blocks																
$\mathcal{T}_1$	479.11K	0.15	1.02	2.52	13.16	18.47	4.75	1.81	9.00	3.72	26.10	2.20	11.89	18.58	8.72	15.43
$\mathcal{T}_2$	826.75K	0.22	0.64	2.44	13.01	19.37	4.47	1.02	5.32	4.37	29.23	2.23	12.69	18.12	8.70	15.21
$\mathcal{T}_3$	1.17M	0.22	0.63	2.77	13.28	19.30	4.46	1.94	4.98	3.83	29.74	3.07	13.41	18.39	8.92	14.92
3 Blocks + Angular $\alpha = 0.1$																
$\langle\mathcal{T}_1\rangle$	479.11K	0.16	2.72	1.80	12.78	15.55	4.20	0.63	9.71	2.68	20.67	1.70	7.97	15.41	7.38	13.84
$\langle\mathcal{T}_2\rangle$	826.75K	0.15	1.84	1.86	12.88	15.54	4.18	0.58	8.00	2.53	22.09	2.00	6.76	15.79	7.25	13.73
$\langle\mathcal{T}_3\rangle$	1.17M	0.16	1.20	2.13	12.92	16.33	4.35	0.93	7.38	2.48	23.04	2.23	6.74	16.64	7.43	13.38
3 Blocks + Angular $\alpha = 0.3$																
$\langle\mathcal{T}_1\rangle$	479.11K	9.68	4.10	3.02	14.01	17.52	5.14	3.17	13.90	18.98	31.33	2.04	4.63	17.26	11.14	18.65
$\langle\mathcal{T}_2\rangle$	826.75K	0.19	3.82	2.14	13.70	17.18	4.41	3.26	12.08	1.50	24.63	1.51	5.70	17.19	8.25	14.46
$\langle\mathcal{T}_3\rangle$	1.17M	0.17	3.58	2.18	13.45	17.29	4.53	3.26	11.72	1.53	25.95	1.58	5.98	17.25	8.34	15.09
4 blocks																
$\mathcal{T}_1$	479.11K	9.75	6.59	3.48	13.19	19.10	4.12	16.83	31.82	35.02	25.22	2.05	40.13	28.73	18.16	14.66
$\mathcal{T}_2$	826.75K	0.15	1.34	2.74	13.25	18.18	4.42	7.64	11.71	2.44	22.52	1.30	8.34	17.72	8.60	14.57
$\mathcal{T}_3$	1.17M	0.16	1.40	2.79	13.16	18.02	4.10	7.64	10.89	2.16	22.69	1.25	8.34	17.12	8.44	14.62
$\mathcal{T}_4$	1.52M	0.19	1.31	2.80	13.49	18.14	4.15	5.44	10.32	2.18	23.17	1.33	13.67	17.44	8.74	14.46
4 Blocks + Angular $\alpha = 0.1$																
$\langle\mathcal{T}_1\rangle$	479.11K	2.80	5.23	2.86	12.94	15.99	3.83	1.68	15.67	7.95	23.80	1.51	37.34	13.90	11.19	14.11
$\langle\mathcal{T}_2\rangle$	826.75K	0.14	3.58	2.76	13.14	16.01	3.72	1.64	13.69	1.97	22.92	1.97	11.93	12.91	8.15	13.99
$\langle\mathcal{T}_3\rangle$	1.17M	0.14	3.24	2.75	13.26	16.01	3.72	1.64	13.18	1.91	22.90	1.51	11.25	12.91	8.03	13.98
$\langle\mathcal{T}_4\rangle$	1.52M	0.13	2.27	2.72	13.38	16.03	3.71	1.55	11.79	1.95	23.07	1.51	10.30	13.26	7.82	13.98

Table 5: Performance comparison on non-English datasets. Reported parameter counts exclude the *XLS-R* (315M) backbone.

Layer	#Params	Dataset												Average	Pooled
		ADD-R1	ADD-R2	D-CH	HABLA	M-DE	M-ES	M-FR	M-IT	M-PL	M-RU	M-UK	SF-CH		
1 Block															
$\mathcal{T}_1$	479.11K	22.98	20.83	14.67	2.00	16.51	8.07	6.62	4.89	6.60	6.10	10.12	30.78	12.52	22.75
2 Blocks															
$\mathcal{T}_1$	479.11K	20.60	21.95	13.36	2.55	17.50	9.14	6.57	6.56	8.69	7.80	11.94	32.71	13.28	23.87
$\mathcal{T}_2$	826.75K	19.57	19.79	13.11	2.21	17.48	8.55	6.70	6.21	8.29	7.84	10.50	28.34	12.38	23.35
2 Blocks + Angular $\alpha = 0.1$															
$\langle\mathcal{T}_1\rangle$	479.11K	23.49	22.43	13.61	2.75	25.39	10.79	9.49	5.69	8.38	7.18	13.04	27.58	14.15	22.90
$\langle\mathcal{T}_2\rangle$	826.75K	21.67	19.86	13.37	2.43	23.21	9.51	8.21	5.64	7.90	6.72	11.06	25.50	12.92	22.19
3 Blocks															
$\mathcal{T}_1$	479.11K	28.89	28.53	17.89	3.20	28.22	13.00	11.31	6.59	12.95	8.20	15.68	38.86	17.78	28.64
$\mathcal{T}_2$	826.75K	26.45	25.92	16.71	3.50	21.93	8.97	8.38	5.68	8.87	6.92	9.72	33.05	14.67	27.18
$\mathcal{T}_3$	1.17M	25.96	24.75	17.09	3.18	20.41	7.74	7.31	5.71	8.45	6.08	9.22	33.11	14.08	27.00
3 Blocks + Angular $\alpha = 0.1$															
$\langle\mathcal{T}_1\rangle$	479.11K	22.46	20.47	15.22	1.65	19.29	10.17	6.66	7.15	8.23	7.52	12.28	31.38	13.54	22.81
$\langle\mathcal{T}_2\rangle$	826.75K	22.55	19.57	15.19	1.84	18.22	9.07	6.14	6.45	7.73	7.08	10.58	30.35	12.90	23.13
$\langle\mathcal{T}_3\rangle$	1.17M	22.70	18.32	15.22	1.78	18.43	8.87	6.08	6.05	7.65	7.10	9.64	27.62	12.45	24.19
4 Blocks															
$\mathcal{T}_1$	479.11K	26.45	26.80	16.58	9.48	36.35	23.24	32.25	14.09	17.13	20.90	20.38	33.58	23.10	25.99
$\mathcal{T}_2$	826.75K	26.03	23.73	14.50	3.20	20.58	10.14	8.33	7.39	9.33	8.54	13.16	30.93	14.65	25.90
$\mathcal{T}_3$	1.17M	24.71	22.88	14.34	2.95	19.70	9.00	7.35	6.87	9.00	8.18	12.58	31.42	14.08	25.09
$\mathcal{T}_4$	1.52M	23.92	22.57	14.01	2.93	19.41	8.71	7.57	6.64	9.05	8.18	11.68	29.90	13.71	24.87
4 Blocks + Angular $\alpha = 0.1$															
$\langle\mathcal{T}_1\rangle$	479.11K	20.51	20.90	13.49	1.84	18.86	10.23	9.45	8.12	11.80	8.78	17.66	30.09	14.31	21.43
$\langle\mathcal{T}_2\rangle$	826.75K	20.65	20.59	13.11	2.16	17.48	9.17	8.22	7.73	10.93	8.40	16.04	29.46	13.66	21.62
$\langle\mathcal{T}_3\rangle$	1.17M	20.66	20.57	13.16	2.22	17.35	9.00	7.94	7.58	10.65	8.14	15.82	29.47	13.55	21.65
$\langle\mathcal{T}_4\rangle$	1.52M	20.51	20.55	13.26	2.31	17.13	8.20	7.24	7.05	9.80	7.44	14.76	29.12	13.12	21.57