Mixture of Low-Rank Adapter Experts in Generalizable Audio Deepfake Detection

Janne Laakkonen*, Ivan Kukanov† and Ville Hautamäki*
* University of Eastern Finland, Joensuu, Finland
E-mail: janne.laakkonen@uef.fi
† KLASS Engineering and Solutions, Singapore

***Abstract*—Foundation models such as Wav2Vec2 excel at representation learning in speech tasks, including audio deepfake detection. However, after being fine-tuned on a fixed set of bonafide and spoofed audio clips, they often fail to generalize to novel deepfake methods not represented in training. To address this, we propose a mixture-of-LoRA-experts approach that integrates multiple low-rank adapters (LoRA) into the model's attention layers. A routing mechanism selectively activates specialized experts, enhancing adaptability to evolving deepfake attacks. Experimental results show that our method outperforms standard fine-tuning in both in-domain and out-of-domain scenarios, reducing equal error rates relative to baseline models. Notably, our best MoE-LoRA model lowers the average out-of-domain EER from 8.55% to 6.08%, demonstrating its effectiveness in achieving generalizable audio deepfake detection.**

I. Introduction

Significant advances in speech synthesis technology have enabled Text-to-Speech (TTS) [1] and Voice Conversion (VC) [2] systems to produce audio indistinguishable from genuine human speech. Malicious actors can exploit synthetic speech to deceive Automatic Speaker Verification (ASV) systems [3] or commit fraud [4], thereby reducing trust in voice-based authentication platforms. Furthermore, the technology can be used to spread misinformation or impersonate public figures in political and social discourse. Ongoing community efforts, such as ASVspoof challenges [5]–[8], underscore that *audio deepfake detection* (ADD), often termed speech anti-spoofing, has become a significant research focus. Although notable progress has been made [9], detection models must generalize effectively to out-of-domain or previously unseen attack types. This is a fundamental requirement given the continuous evolution of deepfake generation methods and the difficulty of generalizing detection models across diverse real-world acoustic conditions.

ADD aims to distinguish between genuine (bonafide) and artificially generated (spoofed) audio. Early studies often relied on handcrafted acoustic features such as LFCCs [10] and CQCCs [11], but recent efforts have shifted toward self-supervised learning (SSL) frameworks, including Wav2Vec2 [12] and HuBERT [13], which can learn generalized acoustic representations from large-scale unlabeled data. Beyond general representations, Graph Neural Networks (GNNs) [14] have also shown promising results in ADD, modeling complex relationships between different parts of the audio signal. The spectrotemporal graph attention network AASIST [15], designed to capture local spoofing artifacts, has become an effective GNN-based architecture. Tak et al. [16] were the first to combine Wav2Vec2 and AASIST, achieving strong results in in-domain evaluations. Despite this progress, current approaches often exhibit a notable performance decrease when faced with unseen attacks or novel acoustic conditions [17], [18], highlighting a key vulnerability: the reliance on fixed, domain-specific cues, which allows more sophisticated or out-of-distribution spoofing to slip past detection.

Parameter-efficient and adaptive fine-tuning strategies offer a promising approach to address generalization challenges in audio deepfake detection. Techniques such as Low-Rank Adapters (LoRA) [19] and Mixture-of-Experts (MoE) [20], [21] have shown promise in adapting large pre-trained models to new tasks or domains with limited data. LoRA achieves this by updating only a small subset of model parameters, while MoE dynamically combines the output of multiple specialized "expert" networks. Recent work has explored applying these techniques to audio deepfake detection, with promising results [18], [22]–[24]. For instance, [24], [25] have demonstrated the effectiveness of applying adapters in Wav2Vec2 for improved performance, while [26] introduced a MoE-based architecture for enhanced generalization across datasets.

Although Tak et al. [16] demonstrated the effectiveness of combining Wav2Vec2 and AASIST for the ADD task, most existing approaches still struggle to adapt when confronted with unseen or evolving spoofing techniques [27]. Their reliance on fixed feature extraction or limited fine-tuning strategies often leads to domain overfitting, making it difficult to generalize beyond the conditions or attack types observed during training. Recent findings [22] indicate that LoRA-integrated models can surpass full fine-tuning in out-of-domain evaluations. Motivated by these results, we propose a sparse mixture-of-LoRA-expert framework that builds on the strong Wav2Vec2 + AASIST baseline. By integrating multiple LoRA experts within the attention layers of Wav2Vec2, our method employs a sparsely gated mechanism that dynamically selects and combines the outputs of a subset of these experts. This design allows the model to specialize in different aspects of the audio signal and to adapt to a wide range of spoofing cues. As a result, our framework improves generalization to out-of-domain attacks by leveraging both the parameter efficiency of

979-8-3315-7206-8/25/$31.00

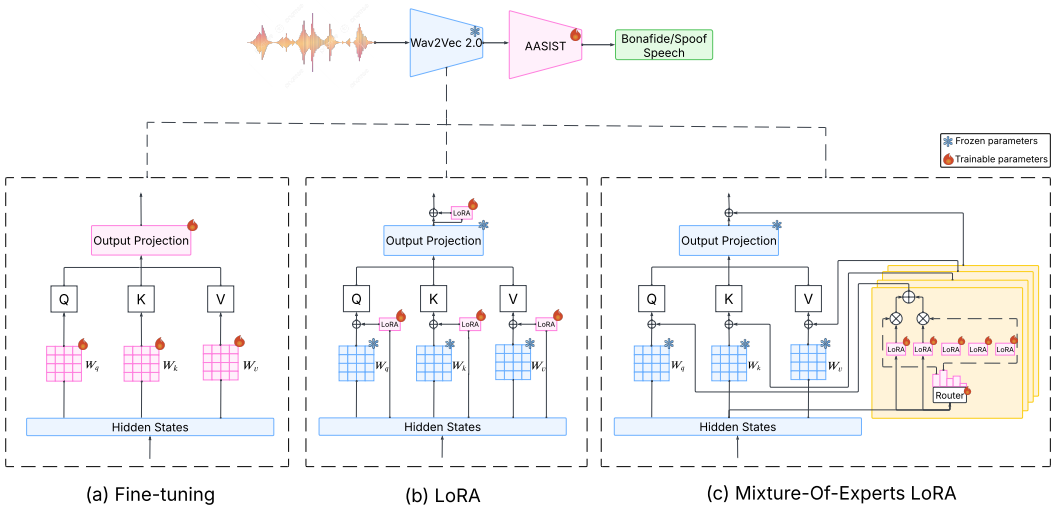

Fig. 1. Overall scheme of the audio deepfake detection system used in the present work (Wav2Vec2 + AASIST). We contrast the proposed (c) to baseline fine-tuning (a) and adapting only the LoRA [22] (b).

LoRA and the adaptability of MoE.

The remainder of this paper is organized as follows. Section 2 details our proposed approach, introducing the underlying Wav2Vec2 + AASIST baseline and outlining how Mixture-of-LoRA Experts is integrated into the attention layers to improve out-of-domain generalization. In Section 3, we describe the experimental setup, including descriptions of the datasets used for both in-domain and out-of-domain evaluations, as well as training protocols and the evaluation metrics used for comparing performance across multiple datasets. Section 4 presents the results and discussion, comparing our approach with the baseline systems and conducting ablation studies to highlight the significance of each component in improving generalization performance. Finally, Section 5 concludes the paper by summarizing our key findings.

II. Mixture-of-LoRA Experts

Recent advancements in large language models have led to efficient techniques for scalability and generalization. Among them, the mixture of experts (MoE) [20] and low-rank adaptation (LoRA) [19] have gained popularity. Originally introduced in [20], MoE has been widely used in speech processing [28], natural language understanding [29], and other applications. Specifically, it was explored for speech deepfake detection in [26].

Initially, low-rank adaptation (LoRA) was designed to efficiently fine-tune large language models [19]. In Fig. 1, we see how LoRA can be applied to a transformer-based neural model. That specific model was used in [22] for generalizable audio deepfake detection.

In the case that one LoRA is not enough, one can add more, where each one is an *expert*. Then, a routing mechanism is needed to select an appropriate LoRA expert or subset of experts for a given input. This system is called the MoE-LoRA technique and has been applied in the context of large language models (LLM), AdaMoLE [30]. In this work, we investigate the fusion of MoE-LoRA for potential improvements in audio deepfake detection.

Mixture-of-Experts. The mixture of experts (MoE) [21] utilizes a framework of specialized models (experts) that collaboratively solve complex tasks based on the input features, dynamically selecting a subset of experts. Formally, a standard MoE module consists of a set of N experts, $\{E_i(\mathbf{x})\}_{i=1}^{N}$, and a gating function $G_i(\mathbf{x})$ that dynamically coordinates the contribution of each expert. For each input $\mathbf{x}$, the gating function $G_i(\mathbf{x})$ has a trainable matrix W_g to distribute the input $\mathbf{x}$ among the experts

$$G_i(\mathbf{x}) = \mathrm{Softmax}(\mathrm{W}_g\,\mathbf{x} + \epsilon)_i, \tag{1}$$

where Gaussian noise $\epsilon \sim \mathcal{N}(\mu, \sigma^2 I)$ with learnable mean μ and variance σ^2 encourages an exploration–exploitation trade-off; it promotes load balancing and helps avoid collapsing to a single most probable expert over time. Only the top-k experts are selected $\mathcal{S}(\mathbf{x}) = \mathrm{TopK}\{G_i(\mathbf{x})\}$, i.e., *sparse selection*. If $k = N$, it is a *dense* MoE variation, which is also explored in experiments. The output from the MoE layer is a weighted sum of the top-k experts

$$\mathbf{y} = \sum_{i \in \mathcal{S}(\mathbf{x})} G_i(\mathbf{x})\, E_i(\mathbf{x}). \tag{2}$$

Low-Rank Adapters. The primary idea behind LoRA is to reduce the number of parameters needed for fine-tuning by approximating weight updates as low-rank matrices rather than updating the entire model's parameters. The general weight update in a neural network is defined as

$$\mathrm{W}' = \mathrm{W}_0 + \Delta\mathrm{W}, \tag{3}$$

where W_0 represents the pre-trained weights of the backbone model, and $\Delta\mathrm{W}$ represents the change introduced by fine-tuning. In LoRA, $\Delta\mathrm{W}$ is parameterized as the product of two

low-rank matrices:

$$\Delta \mathrm{W} = \mathrm{A}\,\mathrm{B}, \tag{4}$$

where $\mathrm{A} \in \mathbb{R}^{d\times r}$ and $\mathrm{B} \in \mathbb{R}^{r\times m}$ are low-rank matrices with rank r, much smaller than the dimensions of W_0: $r \ll d, m$. Then, the output h of the linear layer of the backbone model with fine-tuned LoRA is

$$\mathbf{h} = \mathrm{W}_0\,\mathbf{x} + \Delta \mathrm{W}\,\mathbf{x} = \mathrm{W}_0\,\mathbf{x} + \mathrm{A}\,\mathrm{B}\,\mathbf{x}. \tag{5}$$

This low-rank approximation drastically reduces the number of parameters that need to be learned, improving both the efficiency and flexibility of the fine-tuning process. LoRAs are typically added as side modules to the attention weights or feed-forward layers in the transformer. This allows the pre-trained model to retain its general knowledge while adapting to specific task requirements with minimal computational overhead.

An additional benefit of LoRAs is that the A and B matrices can be stored separately from the backbone model. If the fine-tuning dataset is partitioned into segments, we can even train a separate set of A and B matrices for each segment. This idea then naturally leads to our contribution to the MoE-LoRA.

MoE-LoRA. The fusion of these approaches, termed MoE-LoRA, aims to enhance model efficiency and performance further; see Fig. 1. MoE enables the dynamic selection of experts, where specific LoRA experts detect different types of deepfake artifacts. Combining (2) and (5), the fusion output is

$$\mathbf{h} = \mathrm{W}_0\,\mathbf{x} + \sum_{i\in\mathcal{S}(\mathbf{x})} G_i(\mathbf{x})\left(\mathrm{A}_i\,\mathrm{B}_i\,\mathbf{x}\right), \tag{6}$$

where each pair $(\mathrm{A}_i, \mathrm{B}_i)$ corresponds to a LoRA expert. We incorporate MoE-LoRA modules in each layer of the Wav2Vec2 backbone to explore the contribution of features in each layer.

In Fig. 2, we can see a visualization of fine-tuned MoE-LoRA experts. The maximal singular value of each backbone layer–LoRA expert pair is denoted in the corresponding matrix entry. We observe that, for Q and K transformer matrices, only the last layers are significantly adapted. On the other hand, V and P (multi-head attention output projection) matrices see activity throughout the backbone layers.

III. Experimental Setup

Datasets and evaluation metric: We utilize the ASVspoof 2019 [6] Logical Access (LA) dataset for both training and validation, using its official training and development partitions. To assess the generalizability of our proposed method, we evaluate the models on several datasets:

- **ASVspoof 2019 LA (evaluation split)** [6]: The official evaluation partition from the same 2019 challenge is used to test performance consistency relative to the training domain.
- **ASVspoof 2021 LA and DF** [7]: This comprises Logical Access (LA) and Deepfake (DF) attacks, offering a more diverse range of synthetic speech generation techniques.
- **ASVspoof 5 LA** [8]: A recently released, crowd-sourced dataset of ∼2,000 speakers recorded in diverse acoustic conditions, featuring 32 attack algorithms (including adversarial attacks).
- **In-The-Wild** [27]: A curated 37.9-hour dataset of real and clearly faked audio featuring celebrities and politicians under varying conditions.
- **FakeAVCeleb** [31]: A deepfake dataset derived from 500 celebrity videos in VoxCeleb2 [32]; only the extracted audio is used.

Baseline models. For our primary baseline, we employ the Wav2Vec2 + AASIST system, inspired by previous advancements in speech deepfake detection [16], [18]. Specifically, we utilize Wav2Vec2 XLSR-53 (output dimension 1024) as the front end, coupled with AASIST—a spectrotemporal graph attention network—serving as the back-end classifier. In the fully fine-tuned variant, all parameters in both the SSL front end and AASIST are trainable. In contrast, we define Wav2Vec2 + AASIST* as a partially fine-tuned baseline, where the Wav2Vec2 front end remains frozen, and only AASIST is updated during training.

LoRA models. To explore parameter-efficient adaptations, we integrate Low-Rank Adapters (LoRA) into the Wav2Vec2 encoder's self-attention modules. In these models, only the LoRA parameters and the AASIST back end are trainable, while the rest of Wav2Vec2 remains frozen. For each self-attention block, LoRA matrices are inserted at the query, key, value, and output projections. We study single-LoRA configurations with rank $r \in \{4, 8\}$.

Mixture-of-LoRA-Experts (MoE-LoRA) models. We extend the single-LoRA approach by introducing a mixture-of-experts mechanism within each self-attention block. Each block contains a set of LoRA experts—with ranks $r \in \{4, 8\}$ and a gating router—and we vary the number of experts among $\{3, 5, 7\}$. During forward propagation, a sparse gating strategy selects the top-k experts (with $k \in \{2, \text{num\_experts}\}$), providing a sparse or dense combination of experts. In MoE-LoRA models, the trainable parameters include the router parameters, the LoRA expert parameters, and AASIST.

Training strategy. All variants are trained using the AdamW optimizer [33] with a cyclic learning-rate scheduler that varies the learning rate with a minimum of 1×10^{-7} and a maximum of 1×10^{-5} per cycle. The models are optimized to minimize the negative log-likelihood loss over two-class (bonafide vs. spoof) log-softmax outputs. Models are validated on the ASVspoof 2019 LA development set. Training terminates if no improvement is detected for a fixed number of epochs (10), retaining the best checkpoint.

IV. Results

To investigate the stability of our LoRA-based models, we evaluated both single-LoRA (rank=8) and MoE-LoRA (3 experts, top-k=3, rank=8) configurations under five different random seeds. Fig. 3 summarizes the equal error rates (EER) for six test sets, namely ASVspoof 2019 LA, ASVspoof 2021

TABLE I

COMPARISON OF MODELS TRAINED WITH A SINGLE LORA PER LAYER VS. MODELS TRAINED WITH A MIXTURE OF LORA EXPERTS (MOE). PERFORMANCE IS REPORTED IN TERMS OF EER (%), WHERE BOLDED NUMBERS ARE THE BEST IN EACH COLUMN AND UNDERLINED ARE THE SECOND BEST. SPARSE MODELS USE TOP-k=2, WHILE DENSE MODELS HAVE TOP-k EQUAL TO THE NUMBER OF EXPERTS.

Model	Trainable Params.	MoE Experts	LoRA Rank	Performance (EER %)						
				ASV19:LA	ASV21:LA	ASV21:DF	In-The-Wild	FakeAVCeleb	ASV5	Avg.
Wav2Vec-AASIST	317.8M	–	–	0.28	5.84	5.29	14.03	7.98	23.88	8.55
Wav2Vec-AASIST*	447K	–	–	0.36	4.29	7.97	19.41	4.84	17.14	9.00
LoRA	1.23M	–	4	0.41	10.50	4.37	10.69	9.17	25.97	10.18
	2.02M	–	8	0.61	5.50	5.02	13.15	3.97	21.05	8.22
Sparse MoE	3.40M	3	4	0.71	6.54	5.86	13.11	8.46	19.91	9.10
	5.36M	5	4	0.50	5.33	3.89	11.32	1.81	17.19	6.67
	7.33M	7	4	0.48	4.48	5.84	14.72	2.95	19.11	7.93
	5.76M	3	8	**0.26**	5.73	6.81	9.75	6.96	22.39	8.65
	9.30M	5	8	0.34	6.06	4.69	10.59	8.77	23.66	9.02
	12.83M	7	8	0.35	5.19	5.63	**8.38**	3.71	21.65	7.49
Dense MoE	3.40M	3	4	0.38	5.95	6.78	10.60	6.30	22.49	8.75
	5.36M	5	4	0.42	6.37	4.18	9.11	4.05	20.80	7.42
	7.33M	7	4	**0.26**	**3.70**	4.01	15.59	1.96	18.41	7.31
	5.76M	3	8	0.29	4.24	3.70	9.77	**1.77**	**16.75**	**6.08**
	9.30M	5	8	0.27	4.57	4.16	11.89	3.50	20.12	7.42
	12.83M	7	8	0.69	5.35	**3.25**	9.06	5.02	19.73	7.21

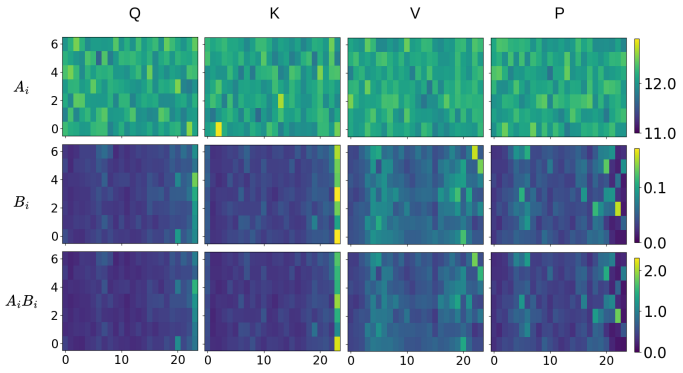

Fig. 2. The maximal singular values of trained MoE-LoRA experts across 24 layers of the Wav2Vec2 backbone; the case of seven experts, top-k is 7 (dense selection), and rank 8. Experts are indexed on the y-axis and backbone layers on the x-axis.

LA, ASVspoof 2021 DF, In-The-Wild, FakeAVCeleb, and ASVspoof 5. The single-LoRA setup has a notably high standard deviation on the FakeAVCeleb set (3.69%), highlighting the impact of random initialization on performance, especially in challenging out-of-domain conditions. By contrast, the MoE-LoRA variant often achieved slightly lower average EER values than single-LoRA, though it also exhibited variability across seeds (for instance, standard deviations reached 3.37% on the In-The-Wild corpus). On the ASVspoof 5 corpus, the single-LoRA configuration achieved an average EER of 16.17% $\pm$ 2.60%, while the MoE-LoRA model obtained 18.81% $\pm$ 2.68%, indicating no gain from the dense three-expert setup on this crowd-sourced, highly heterogeneous dataset. The variability in the evaluation EERs suggests that expert selection and gating can be sensitive to initialization, emphasizing the importance of aggregating or repeating trials when comparing approaches.

Table I compares the performance of three model types: fully or partially fine-tuned baselines (Wav2Vec2 + AASIST), single Low-Rank Adaptation (LoRA) configurations with varying ranks, and Mixture-of-LoRA-Experts (MoE-LoRA) variants. The fully fine-tuned Wav2Vec2 + AASIST baseline achieves an average EER of 8.55% across all test sets; the partially fine-tuned version (frozen Wav2Vec2) yields a slightly higher EER of 9.00%. While superior overall, the fully fine-tuned model struggles with out-of-domain generalization, achieving an EER exceeding 19% on challenging datasets like ASVspoof 5.

Replacing full fine-tuning with a single LoRA layer within Wav2Vec2 demonstrates the effectiveness of parameter-efficient adaptation. A rank-4 LoRA narrows the performance gap. A rank-8 LoRA, however, achieves an average EER of 8.22%, outperforming the partially fine-tuned approach and nearing the fully fine-tuned baseline. This highlights LoRA's ability to effectively adapt the model while keeping most Wav2Vec2 parameters frozen.

The MoE-LoRA framework further improves detection accuracy by utilizing multiple LoRA experts, each potentially specializing in different signal aspects. Each self-attention layer can employ either a sparse set of experts (a subset active at a time) or a dense set (all contributing). Sparse gating reduces computational load; dense gating can, in some cases, yield better performance. Notably, a dense MoE-LoRA configuration with three rank-8 experts achieves an average EER of 6.08%, significantly outperforming both single-LoRA

and the fine-tuned baselines. Adding more than three experts offers diminishing returns, with only marginal gains relative to the increased computational cost.

In summary, MoE-LoRA offers a compelling balance of flexibility and efficiency. By allowing expert specialization in detecting diverse spoofing cues, it achieves significantly lower error rates than single-LoRA or baseline fine-tuning. These results suggest that increased model capacity, combined with strategic gating and parameter-efficient adaptation, can markedly improve generalization to unseen deepfake attacks without a large increase in model size or computational demands.

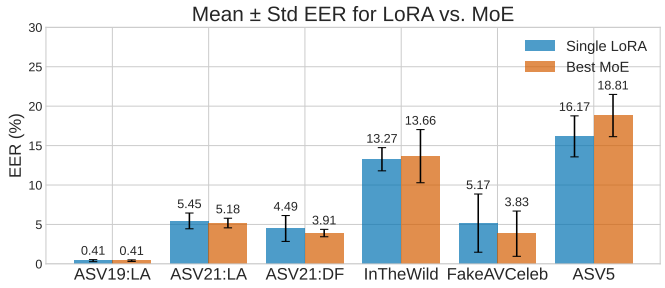

Fig. 3. Mean ± std EER (%) for LoRA and MoE models across five different seeds. The LoRA models use rank 8, while the MoE models have three experts, are dense (top-k=3), and also use rank 8. Error bars represent standard deviation across seeds.

V. Conclusions

Detecting audio deepfakes is challenging, as new generation techniques constantly outpace detection systems. We aim to improve the adaptability of audio foundation models to address this challenge. While Wav2Vec2 excels at audio representation learning, fine-tuned versions struggle with new deepfake types. Our results show a fully fine-tuned model averaging an 8.55% EER across several out-of-domain sets. LoRA provides a partial solution, achieving 8.22% EER. For significant generalization gains, we introduce a second representation-learning layer: a Mixture-of-Experts (MoE) approach. Combining multiple LoRA experts with strategic routing lowers the EER to 6.08%. MoE-LoRA's adaptability offers a practical path towards reliable deepfake detection.

Acknowledgment

This work was supported by the Finnish Doctoral Program Network in Artificial Intelligence, AI-DOC (decision number VN/3137/2024-OKM-6). The authors also wish to acknowledge CSC – IT Center for Science, Finland, for computational resources. Additionally, Ville Hautamäki thanks the Jane and Aatos Erkko Foundation for partial funding.

References

[1] C. Zhang, C. Zhang, S. Zheng, *et al.*, *A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai*, 2023. arXiv: 2303.13336 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2303.13336.

[2] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022.

[3] M. Todisco, M. Panariello, X. Wang, *et al.*, "Malacopula: Adversarial automatic speaker verification attacks using a neural-based generalised hammerstein model," *ArXiv*, vol. abs/2408.09300, 2024.

[4] N. Robins-Early, *Ceo of world's biggest ad firm targeted by deepfake scam*, https://www.theguardian.com/technology/article/2024/may/10/ceo-wpp-deepfake-scam [Accessed: (2025-02-10)], 2024.

[5] Z. Wu, J. Yamagishi, T. Kinnunen, *et al.*, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[6] M. Todisco, X. Wang, V. Vestman, *et al.*, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," English, in *Proceedings Interspeech 2019*, Interspeech 2019 ; Conference date: 15-09-2019 Through 19-09-2019, International Speech Communication Association, Sep. 2019, pp. 1008–1012. DOI: 10.21437/Interspeech.2019-2249.

[7] J. Yamagishi, X. Wang, M. Todisco, *et al.*, "Asvspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

[8] X. Wang, H. Delgado, H. Tak, *et al.*, "Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," *arXiv preprint arXiv:2408.08739*, 2024.

[9] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "Audio anti-spoofing detection: A survey," *arXiv preprint arXiv:2404.13914*, 2024.

[10] F. Alegre, A. Amehraye-Fillatre, and N. Evans, *A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns*, Oct. 2013. DOI: 10.1109/BTAS.2013.6712706.

[11] H. Tak, J. Patino, A. Nautsch, *et al.*, "An explainability study of the constant q cepstral coefficient spoofing countermeasure for automatic speaker verification," in *The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 333–340. DOI: 10.21437/Odyssey.2020-47.

[12] A. Baevski, Y. Zhou, A. Mohamed, *et al.*, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 12 449–12 460.

[13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio,*

Speech and Lang. Proc., vol. 29, pp. 3451–3460, Oct. 2021, ISSN: 2329-9290.

[14] F. Scarselli, M. Gori, A. C. Tsoi, *et al.*, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, pp. 61–80, 2009.

[15] J.-w. Jung, H.-S. Heo, H. Tak, *et al.*, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371.

[16] H. Tak, M. Todisco, X. Wang, *et al.*, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112–119.

[17] N. M. Müller, N. Evans, H. Tak, *et al.*, "Harder or different? understanding generalization of audio deepfake detection," in *Interspeech 2024*, 2024, pp. 2705–2709. DOI: 10.21437/Interspeech.2024-247.

[18] I. Kukanov, J. Laakkonen, T. Kinnunen, and V. Hautamäki, "Meta-learning approaches for improving detection of unseen speech deepfakes," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 1173–1178. DOI: 10.1109/SLT61566.2024.10832350.

[19] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[20] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, pp. 79–87, Mar. 1991. DOI: 10.1162/neco.1991.3.1.79.

[21] N. Shazeer, A. Mirhoseini, K. Maziarz, *et al.*, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[22] J. Laakkonen, I. Kukanov, and V. Hautamäki, *Generalizable speech deepfake detection via meta-learned lora*, 2025. arXiv: 2502.10838 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2502.10838.

[23] X. Zhang, J. Yi, J. Tao, *et al.*, *Adaptive Fake Audio Detection with Low-Rank Model Squeezing*, 2023. arXiv: 2306.04956 [cs.SD].

[24] C. Wang, J. Yi, X. Zhang, *et al.*, "Low-rank adaptation method for wav2vec2-based fake audio detection," in *DADA@IJCAI*, 2023.

[25] H. Wu, W. Guo, S. Peng, *et al.*, "Adapter learning from pre-trained model for robust spoof speech detection," in *Proc. INTERSPEECH 2024*, Sep. 2024, pp. 2095–2099. DOI: 10.21437/Interspeech.2024-253.

[26] V. Negroni *et al.*, "Leveraging mixture of experts for improved speech deepfake detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[27] N. Müller, P. Czempin, F. Diekmann, *et al.*, "Does audio deepfake detection generalize?" In *Interspeech 2022*, 2022, pp. 2783–2787. DOI: 10.21437/Interspeech.2022-108.

[28] Z. You, S. Feng, D. Su, *et al.*, "Speechmoe: Scaling to large acoustic models with dynamic routing mixture of experts," in *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, ISCA, 2021, pp. 2077–2081.

[29] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *CoRR*, vol. abs/2101.03961, 2021. eprint: 2101.03961.

[30] Z. Liu and J. Luo, "AdamoLE: Fine-tuning large language models with adaptive mixture of low-rank adaptation experts," in *First Conference on Language Modeling*, 2024. [Online]. Available: https://openreview.net/forum?id=ndY9qFf9Sa.

[31] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[32] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech 2018*, 2018, pp. 1086–1090. DOI: 10.21437/Interspeech.2018-1929.

[33] I. Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, 2019. arXiv: 1711.05101.