



Thinking Fast and Slow: Robust Speech Recognition via Deep Filter-Tuning

Dianwen Ng^{1,2}, Kun Zhou³, Bin Ma³, Eng Siong Chng¹

¹College of Computing and Data Science, Nanyang Technological University, Singapore

²MiroMind, Singapore

³Tongyi Speech Lab, Alibaba Group, Singapore

dianwen001@e.ntu.edu.sg

Abstract

Self-supervised learning (SSL) models have revolutionized speech representation by extracting rich acoustic and phonetic features with minimal labeled data. However, their computational demands during fine-tuning and vulnerability to catastrophic forgetting pose challenges for practical deployment. Parameter-Efficient Fine-Tuning (PEFT) methods, such as prompt tuning, are often employed to address these challenges. While prompt tuning has been successful with large language models in natural language processing (NLP), it struggles to learn effective instructional signals when adapting to speech SSL models, likely due to insufficient a priori knowledge that hinders soft token learning during fine-tuning. We introduce Deep Filter Tuning (DFT), a soft-token adaptation strategy that selectively filters semantic information from noise-distorted representations. By modifying only 0.38% of model weights, DFT achieves a 12% performance gain in noisy environments, offering an efficient solution for robust speech recognition under challenging conditions such as noise adaptation.

Index Terms: Noise Robust Adaptation, Parameter Efficient Fine-tuning, Speech Recognition

1. Introduction

Self-supervised learning (SSL) models have made significant strides in speech representation, effectively extracting expressive acoustic and phonetic features with minimal reliance on labeled data [1, 2, 3, 4, 5, 6]. These models have catalyzed the development of robust applications in traditionally data-scarce environments. However, their extensive computational requirements for downstream task fine-tuning pose challenges in terms of cost-effectiveness and feasibility in resource-constrained settings [7, 8, 9, 4]. SSL models are also particularly susceptible to catastrophic forgetting when operating with limited data resources, which can severely impair their performance [10].

In response to these challenges, there is increasing interest in parameter-efficient fine-tuning (PEFT) techniques. These methods aim to refine large pre-trained SSL models by making minimal updates to their parameters. This approach helps conserve computational resources and retains much of the pre-trained a priori knowledge, effectively balancing cost-efficiency with knowledge retention [11]. Among various PEFT strategies, prompt tuning emerges as a particularly promising method due to its adaptability and efficiency. Unlike traditional fine-tuning, which updates all model parameters, prompt tuning introduces either instructional vectors or trainable soft tokens. These elements guide the model, leveraging its pre-existing knowledge [12] to shape latent features to achieve specific downstream objectives. This method makes prompt tuning a versatile tool for adapting SSL models to various applications

[13, 14]. However, applying prompt tuning to speech representation models presents unique challenges. Speech signals are inherently more complex than text [3], given their continuous nature and the dynamic features they encompass, such as prosody, speaker emotion, and ambient noise. These characteristics require more nuanced adaptations than those typically needed for natural language processing (NLP) [7]. In this paper, we will detail these technical challenges to provide a clearer understanding of the issues at hand.

To address these challenges, our research introduces Deep Filter Tuning (DFT), a novel downstream adaptation strategy that adapts the soft-token-based tuning algorithm to the acoustic domain, specifically designed to manage the dynamic complexities of real-world speech scenarios. DFT is inspired by speech extraction techniques that use filters [15] and Feature-wise Linear Modulation (FiLM) [16], adapted in our approach to utilize initialized soft tokens. This method enables selective filtering of crucial semantic information from continuous noise-distorted latent representations, effectively targeting the speech content for robust speech recognition. In particular, the proposed module develops a sophisticated filtering system comprising **static filters sourced from soft tokens** and **adaptive filters instantiated through a linear bottleneck**, each being specifically designed to counteract noise in the targeted domain. During inference, these filters proficiently regulate both the temporal and channel-wise dimensions of the intermediate latent representations. Utilizing external parameter modulators in conjunction with a filtering mask on frozen pre-trained representations, our method effectively mitigates adverse distortions. This approach marks a substantial deviation from conventional prompt tuning, which typically centers on acquiring an instruction set and exploiting pre-trained models to refine latent features.

Through comprehensive empirical evaluations, we have demonstrated that DFT significantly enhances automatic speech recognition (ASR) performance in noisy conditions. It consistently achieves over a 12% gain in various domain environments compared to the standard prompt tuning model. These improvements are realized using the robust WavLM+ speech encoder while utilizing only 0.38% of the full model's weights, illustrating DFT's efficiency and effectiveness in managing challenging acoustic scenarios. Additionally, we have verified consistent performance gains on larger models, which further supports the effectiveness of our approach.

2. Related Work

Prompt (soft-token based) tuning has proven to be a versatile adaptation strategy across various machine learning domains, including visual question answering [17, 18] and vision-language models [19]. In speech processing, its potential

for rapid adaptation to new speakers was highlighted by [20], who explored its use in end-to-end speech recognition systems. However, applying prompt tuning to ASR presents unique challenges. While this method has proven effective in NLP, its efficacy in ASR is less certain. Chen et al. [7] highlighted significant challenges in initializing embedding tokens that accurately capture the continuous nature of speech signals. This difficulty is further exacerbated by the temporal dynamics inherent in speech variability, such as differences in speaking rate and accent. These factors significantly impact the effectiveness of prompt tuning, underscoring the need for more adaptive methods in this domain, an area that remains underexplored.

Recent theoretical advances have illuminated the mechanisms of prompt tuning. Oymak et al. (2023) [21] explored the role of attention mechanisms, proposing that soft prompts may function as learned key vectors in attention calculations. While Bailey et al. (2023) [22] revealed that soft prompts occupy distinct regions in the embedding space, with geometrical properties such as magnitude and direction, significantly differ from those of natural language prompts. This distinction highlights gaps in interpretable systems, which are crucial for improving control and design, as well as mitigating potential malicious attacks and biases. Our work empirically evaluated various prompt (soft-token-based) tuning methods for noise robust ASR, confirming these challenges. We aimed to overcome the limitations identified in earlier studies by developing more robust adaptation techniques for pre-trained SSL speech models. Our objective is to make these powerful models more accessible, even in settings with limited resources.

3. Methodology

3.1. Preliminaries – Soft-prompt (Token) Based Tuning

To simplify the task of soft prompt tuning without the loss of generality, we consider a single-head self-attention layer,

$$\mathcal{O}_{\text{frozen}} = \varphi\left(\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^T\mathbf{X}^T\right)\mathbf{X}\mathbf{W}_V \quad (1)$$

with input $X \in \mathbb{R}^{T \times d}$ consisting of T utterance frames of dimension d each. \mathbf{W}_Q , \mathbf{W}_K and \mathbf{W}_V are the frozen pre-trained weights for *query*, *key* and *value*. φ denotes the softmax nonlinearity function that acts row-wise for a $T \times T$ matrix. To incorporate trainable prompt tokens as instructional signals for PEFT on the downstream task, a series of soft embeddings, denoted as $\mathbf{P} \in \mathbb{R}^{m \times d}$, is prepended to X , resulting in the augmented matrix of $\mathbf{X}_P := [\mathbf{P} \ X] \in \mathbb{R}^{(m+T) \times d}$, serving as the latent input to the transformer blocks. The output of the attention-layer, as introduced in [13, 21] is thus in the form

$$\mathcal{O} = \varphi\left(\frac{1}{\sqrt{d}}\mathbf{X}_P\mathbf{W}_Q\mathbf{W}_K^T\mathbf{X}_P^T\right)\mathbf{X}_P\mathbf{W}_V \quad (2)$$

Note that this is slightly different from (1) in that now the layer computes a cross-attention between the augmented inputs \mathbf{X}_P and the original inputs X . We can rearrange this to

$$\begin{aligned} \mathcal{O} &= \varphi\left(\frac{1}{\sqrt{d}}[\mathbf{P} \ \mathbf{X}] \mathbf{W}_Q \mathbf{W}_K^T \begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix}\right) \begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix} \mathbf{W}_V \\ &= \underbrace{\varphi\left(\frac{1}{\sqrt{d}}\mathbf{P}\mathbf{W}_Q\mathbf{W}_K^T\mathbf{P}^T\right)}_{\text{prompt-attention, } \mathcal{O}_{\text{prompt}}} \begin{bmatrix} \mathbf{P} \\ \mathbf{X} \end{bmatrix} \mathbf{W}_V + \underbrace{\varphi\left(\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^T\mathbf{X}^T\right)}_{\text{frozen self-attention, } \mathcal{O}_{\text{frozen}}} \mathbf{X} \mathbf{W}_V \end{aligned} \quad (3)$$

Here, we observe that the adaptation stems from the additive component of prompt-attention on otherwise static latent speech representations. The efficacy of the output hinges on the quality

of the prompt signals, which interact with the pre-trained network to extract relevant knowledge for handling the information required by the downstream task. Therefore, achieving an optimal solution requires a robust pre-trained model capable of managing diverse scenarios and effective prompts that guide the model in addressing complex noise distortions and variations in speaker attributes.

It is important to recognize that although pre-trained models have been exposed to a substantial amount of data during training, they cannot account for all possible variations of speakers and specific noise distortions. Consequently, the effectiveness of this approach is limited by the generalizability of the pre-trained knowledge and the quality of the soft prompts. Rather than relying solely on the pre-trained model, we propose an alternative strategy that uses soft tokens as feature-modulating filters.

3.2. Deep Filter Tuning (DFT)

To address the challenge of inadequate a priori knowledge, which hinders the effective learning of soft tokens during fine-tuning, we propose a more deliberate application of soft tokens to modulate information specifically for noisy speech recognition. This approach involves constructing a static bias from the soft tokens and scaling the intermediate representations to deemphasize noise signals from speech content, akin to the FiLM operation used in speaker extraction, as referenced in [23, 24]. Specifically, we aim to develop a filtering mask derived from the soft tokens to selectively attenuate noisy signals within the frozen features.

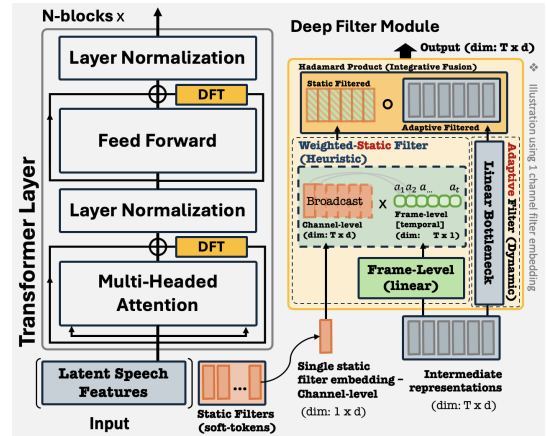


Figure 1: An illustration of the Deep Filter Tuning module operating on a single functional channel filter embedding.

As shown in Figure 1, we initialize a set of m soft tokens $\mathbf{S} \in \mathbb{R}^{m \times d}$ that act as static filtering tokens in handling noisy latent speech features. Each soft token functions as a slightly different masking filter, which is broadcast to match the sequential length T of the frozen representations. To weigh the importance of each masking token on every frame, we introduce a linear module that computes the temporal weights of the token conditional on the input, denoted as $W = \delta(\mathbf{X}\mathbf{W}_s)^T \in \mathbb{R}^{m \times T}$, where $\mathbf{W}_s \in \mathbb{R}^{d \times m}$, and δ represents \tanh . The weighted-static filter is calculated as $W^T \mathbf{S} \in \mathbb{R}^{T \times d}$, which serves as the scaling mask. Soft tokens, which are trainable parameters, are optimized through the noise adaptation objective of ASR. This forces them to learn a masking pattern that extracts speech content from noisy signals, mirroring the scaling factor γ in FiLM of speaker extraction.

Nevertheless, just as cognitive processing encompasses both ‘fast’ and ‘slow’ thinking, static filters in our model rely on predefined implicit biases (akin to heuristics) to swiftly manage noise, paralleling fast cognitive processing. Conversely, we have introduced a branch specifically designed for slow cognitive assessment, which features a bottleneck linear architecture to facilitate rapid assimilation of sampled utterances. This process is deliberate and incorporates conscious (dynamic) analysis. The modulated output is then represented by the Hadamard product of the heuristic bias and the bottleneck-adapted representations. As a whole, the system adopts cognitive biases similar to human processing, functioning as an adaptive information bottleneck that selectively permits relevant information to pass, embodying principles of both thinking fast and slow.

4. Experiment Setup

We insert DFT module in parallel with the main layers, positioned adjacent to the multi-headed attention and feed-forward layers, following the typical PEFT plug-in framework [25]. Notably, DFT module is shared between the two parallel insertions within the same block, as illustrated in Figure 1. This configuration minimizes the introduction of additional parameters and ensures efficient computational resource usage. Furthermore, we enhance the functionality of the DFT module by optimizing soft tokens using the vanilla prompt attention mechanism, a combination we have named DFT++. This hybrid approach allows soft tokens to be processed through the prompt attention mechanism, leveraging optimization gradients to facilitate communication between the soft tokens and the pre-trained model. This interaction builds a stronger heuristic bias, enabling a deeper understanding of the intrinsic knowledge, which significantly improves the modulation of masking information for enhanced signal processing.

The training dataset consists of a 100-hour subset from LS combined with the full 10-hour ESD [26]. To create a noisy corpus, we corrupted the speech data with the FreeSound noise dataset, which includes both stationary (Type A) and non-stationary (Type B) noises. Type A noises consist of sounds from cars, metros, and traffic, while Type B includes babble, airport/station, café, and AC/vacuum noises. Each noise type comprises 10 audio streams for training and 8 for testing, totaling approximately 2 hours of noise data. During testing, we evaluated performance across specific environments categorized into clean, noisy, and emotional domains. Clean testing was conducted on the official LS testing set. For noisy ASR testing, we used the noise pre-mixed testing set [27], selecting 120 sub-files from the LibriSpeech test-clean set and corrupting them with test noise at various signal-to-noise ratios (SNRs) ranging from 0 to 20 dB, resulting in 4,200 instances of noisy test data. We benchmark the performance of prompt-based tuning against other commonly adopted PEFT methods to provide a more comprehensive understanding. These methods include Adapter-Tuning, LoRA Tuning, and fully frozen network tuning. Our implementation setup closely follows the approach outlined in [7] for PEFT optimization. In Adapter-Tuning, we use a reduction factor of 4 with one adapter module at the feed-forward layer, following Houlsby [25]. For Prompt-Tuning, we prepend 300 trainable prompts to the input utterance—both setups, which match the trainable size of DFT for comparable complexity. In the case of DFT, we employ 10 trainable soft tokens alongside the filtering modules. We use an RNN decoder [28], specifically from [29], to decode the features from the encoder.

5. Results

Table 1 illustrates the speech recognition performance in a noisy environment, where we compare our proposed DFT method to other approaches using a synthetic in-domain noisy corpus with noise levels ranging from 0 to 20dB. DFT consistently outperforms prompt tuning, reducing the Word Error Rate (WER) by 13.7% under noisy conditions on the base WavLM+ [3]. This improvement highlights the efficacy of the ‘fast’ and ‘slow’ filtering units in our model, which effectively modulate information flow and enhance content representations, adapting latent representations to various noisy environments. This performance trend is consistent across other SSL-based speech encoders, notably HuBERT [2] and WavLM (Large) [3].

Additionally, we observed an unexpected performance decline in LoRA, evidenced by an increased error rate compared to the fully frozen model, which only adapts a priori knowledge. We hypothesize that the assumption underlying LoRA—that fine-tuned model weights remain low-rank—may not hold in contexts involving continuous speech representations and the cross-modality demands of speech recognition systems. These systems transition from processing wave signals to generating discrete text output, a complex process not fully learned during upstream pre-training and affected by variations in speaker and background noises, potentially leading to higher-rank fine-tuned parameters. As such, the learning outcome for LoRA is suboptimal, which generated poorer performance. However, further investigation into this aspect was beyond the scope of our study.

Out-of-domain testing on the CHiME-4 [30] real noisy ASR dataset was conducted to evaluate the generalizability of our method. While fine-tuning the full WavLM+ model demonstrated strong in-domain noise performance, it struggled with generalization to out-of-domain (OOD) cases. This underscores a potential pitfall of fine-tuning the full model—prone to catastrophic forgetting and overfitting, especially in low-resource scenarios common with SSL models. In contrast, our DFT++ approach outperformed all other PEFT methods, demonstrating superior efficacy and consistency across various SSL models.

5.1. Ablation Studies of Deep Filter-Tuning

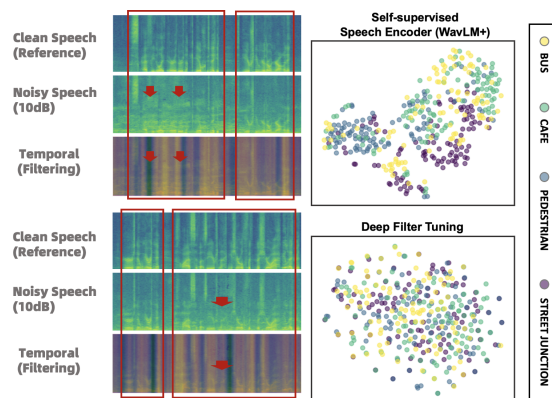


Figure 2: Results of ablation studies performed using DFT. **Left:** shows the action heatmap of frame-level temporal weights activating static filters on noisy utterances. **Right:** displays the t-SNE plot for OOD CHiME4 (real noise). Vector representations were obtained by average pooling 100 sampled utterances from each category.

To observe the static filtering operation in action, we illus-

Table 1: The table shows the word error rate, WER (%) (\downarrow) of the ASR system on noisy speech recognition at SNRs of (0 - 20)dB for the synthesized noisy in-domain LS (FreeSound) and real noisy speech CHiME-4 (OOD) dataset

Models	Params (M)	Non-Stationary (Type-B) Noise				Stationary (Type-A) Noise			Avg. (Noisy)	CHiME-4 (Real)
		Babble	Airport/ Station	AC/ Vacuum	Cafe	Traffic	Metro	Car		
HuBERT (Base)	94.70	25.87	19.93	17.78	14.17	13.58	13.22	8.83	16.20	31.22
Frozen	0	49.98	39.32	31.98	26.49	23.15	20.83	11.53	29.04	48.56
Adapter	3.54	33.74	25.23	23.57	16.58	15.19	15.87	8.17	19.76	38.43
LoRA (R=16)	0.59	54.43	42.94	36.88	28.87	24.51	23.17	11.14	31.71	54.08
Prompt Tuning	0.23	47.51	37.45	30.32	24.33	20.64	19.45	9.62	27.05	44.38
DFT (Ours)	0.36	34.07	25.42	22.77	14.91	15.04	14.85	7.80	19.27	38.62
DFT++ (Ours)	0.36	33.28	23.89	22.05	13.88	14.52	13.97	7.23	18.40	37.83
WavLM+ (Full)	94.70	15.13	11.35	11.29	8.71	9.01	8.64	6.01	10.02	20.58
Frozen Model	0	26.65	18.39	16.70	13.27	12.63	11.90	7.38	15.27	26.77
Adapter Tuning	3.54	16.00	11.60	11.80	8.95	8.73	8.88	6.04	10.29	19.46
LoRA (R=16)	0.59	35.71	26.17	21.59	15.58	14.28	14.04	7.86	19.32	30.98
Prompt Tuning	0.23	18.32	13.68	13.26	10.01	10.17	9.49	6.62	11.65	22.57
DFT (Ours)	0.36	15.70	11.28	11.33	8.65	8.74	8.72	5.94	10.05	19.67
DFT++ (Ours)	0.36	15.39	11.13	11.19	8.46	8.65	8.61	5.86	9.90	19.49
WavLM+ (Large)	315	7.97	6.38	6.30	5.01	5.28	4.96	3.65	5.65	12.13
Frozen Model	0	13.72	10.98	9.21	7.14	6.94	6.83	4.28	8.44	16.24
Adapter Tuning	12.6	8.51	6.55	6.47	5.12	5.23	5.03	3.71	5.80	11.59
LoRA (R=16)	2.10	16.59	12.68	11.81	8.83	8.77	8.42	6.53	10.52	20.87
Prompt Tuning	0.31	9.32	7.18	7.20	5.82	5.90	5.39	3.82	6.38	13.07
DFT (Ours)	1.90	8.70	6.41	6.38	5.05	5.14	4.81	3.49	5.71	11.87
DFT++ (Ours)	1.90	8.51	6.28	6.07	4.99	5.07	4.59	3.48	5.57	11.56

trate a heatmap of frame-level temporal weights superimposed on the spectrogram of a randomly selected utterance from the LS Test Clean subset (Figure 2). Background noise was artificially introduced into the speech sample by adding 10 dB of babble noise, simulating the presence of another person speaking in the background. We computed the mean of temporal weights across all layers to demonstrate how heuristic biases are generally applied at the frame level. To enhance clarity and focus on relevant spectral features, soft tokens (representing channel-wise heuristic biases) were omitted. These tokens embody high-dimensional latent information, which does not correspond to the y-axis of the spectrogram and could potentially obscure interpretative accuracy.

In the heatmap, speech segments are marked with boxes, while noise-distorted regions are indicated by red arrows. Both extremes of the tanh weights (-1 and 1, shown as darker and brighter shades respectively) indicate high filter activation, while values near 0 represent minimal processing. This bipolar activation pattern demonstrates the filter’s ability to both enhance speech content and suppress noise across different frames, validating the choice of tanh activation.

The t-SNE distributions, comparing the frozen WavLM+ model with our proposed DFT approach on OOD CHiME-4 speech, demonstrate the filter’s effectiveness. Our approach shows less prominent noise domain clusters and more random arrangement, indicating successful noise attenuation and improved content preservation.

We evaluated the relative importance of ‘Fast’ versus ‘Slow’ thinking by removing each sub-branch and finetuning under identical conditions, testing only with unprocessed real speech. To distinguish between the adaptive slow conscious branch and the adapter module, we used a smaller reduction factor of 64, resulting in an intermediate channel dimension of 12.

Table 2: WERs on experiments using only the single cognitive branch of the filtering module from DFT using WavLM+.

Sub-models	Clean (LS)	Other (LS)	CHiME4
Adapter Tuning	4.47	9.83	19.46
DFT (Ours-Default)	4.56	9.73	19.67
Adaptive (Slow-Conscious)	4.73	10.47	21.13
Static (Fast-Heuristic)	4.85	10.12	20.01

This bottleneck forces the slow conscious branch to compress more information and identify the most important features. Results in Table 2 show that static filters alone achieve effective domain adaptation, providing good calibration with fewer trainable parameters than the adapter. The static filters perform comparably to the adaptive slow conscious branch to enforce selective feature processing. Notably, the slow conscious module slightly underperforms compared to fast-heuristic, plausibly caused by over-filtering content while overthinking noise distortion. However, when combined with fast-heuristic biases, the slow-conscious module achieves optimal performance by relaxing its filtering criteria.

6. Conclusion

We present DFT, a novel approach that enhances soft-token PEFT for robust ASR by incorporating dual-process filtering. The system combines fast and slow thinking mechanisms to manage environmental distortions and improve speech recognition under challenging conditions. Our empirical evaluation shows that DFT achieves a 12% relative gain in ASR performance over conventional prompt tuning methods while modifying only 0.38% of WavLM+’s parameters, demonstrating significant efficiency across diverse domains.

7. References

- [1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.
- [2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [3] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.
- [5] D. Ng, R. Zhang, J. Q. Yip, Z. Yang, J. Ni, C. Zhang, Y. Ma, C. Ni, E. S. Chng, and B. Ma, “De’hubert: Disentangling noise in a self-supervised model for robust speech recognition,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [6] D. Ng, K. Zhou, Y.-W. Chao, Z. Xiong, B. Ma, and E. S. Chng, “Multi-band frequency reconstruction for neural psychoacoustic coding,” *arXiv preprint arXiv:2505.07235*, 2025.
- [7] Z.-C. Chen, C.-L. Fu, C.-Y. Liu, S.-W. D. Li, and H.-y. Lee, “Exploring efficient-tuning methods in self-supervised speech models,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1120–1127.
- [8] R. Fan, Y. Zhu, J. Wang, and A. Alwan, “Towards better domain adaptation for self-supervised models: A case study of child asr,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1242–1252, 2022.
- [9] B. Thomas, S. Kessler, and S. Karout, “Efficient adapter transfer of self-supervised speech models for automatic speech recognition,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7102–7106.
- [10] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, “Towards a unified view of parameter-efficient transfer learning,” in *International Conference on Learning Representations*, 2021.
- [11] D. Ng, C. Zhang, R. Zhang, Y. Ma, T. H. Nguyen, C. Ni, S. Zhao, Q. Chen, W. Wang, E. S. Chng *et al.*, “Adapter-tuning with effective token-dependent representation shift for automatic speech recognition,” in *In Proc. Interspeech 2023*, 2023, pp. 1319–1323.
- [12] D. Ng, C. Zhang, R. Zhang, Y. Ma, F. Ritter-Gutierrez, T. H. Nguyen, C. Ni, S. Zhao, E. S. Chng, and B. Ma, “Are soft prompts good zero-shot learners for speech recognition?” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 366–10 370.
- [13] B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efficient prompt tuning,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3045–3059.
- [14] X. L. Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 4582–4597.
- [15] S. Zhao, Y. Ma, C. Ni, C. Zhang, H. Wang, T. H. Nguyen, K. Zhou, J. Q. Yip, D. Ng, and B. Ma, “Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 356–10 360.
- [16] M. O. Turkoglu, A. Becker, H. A. Gündüz, M. Rezaei, B. Bischl, R. C. Daudt, S. D’Aronco, J. Wegner, and K. Schindler, “Film-ensemble: Probabilistic deep learning via feature-wise linear modulation,” *Advances in neural information processing systems*, vol. 35, pp. 22 229–22 242, 2022.
- [17] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. Selvan Dhanasekaran, A. Naik, D. Stap *et al.*, “Benchmarking generalization via in-context instructions on 1,600+ language tasks,” *arXiv e-prints*, pp. arXiv–2204, 2022.
- [18] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.
- [19] M. Kim, H.-I. Kim, and Y. M. Ro, “Prompt tuning of deep neural networks for speaker-adaptive visual speech recognition,” *arXiv preprint arXiv:2302.08102*, 2023.
- [20] K. Goswami, L. Lange, J. Araki, and H. Adel, “Switchprompt: Learning domain-specific gated soft prompts for classification in low-resource domains,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 2681–2687.
- [21] S. Oymak, A. S. Rawat, M. Soltanolkotabi, and C. Thrampoulidis, “On the role of attention in prompt-tuning,” in *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.
- [22] L. Bailey, G. Ahdriz, A. Kleiman, S. Swaroop, F. Doshi-Velez, and W. Pan, “Soft prompting might be a bug, not a feature,” in *ICML 2023 Challenges of Deploying Generative AI Workshop*, 2023.
- [23] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu, “Neural target speech extraction: An overview,” *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.
- [24] Z. Pan, M. Borsdorf, S. Cai, T. Schultz, and H. Li, “Neuroheed: Neuro-steered speaker extraction using eeg signals,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [25] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, “Parameter-efficient transfer learning for nlp,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.
- [26] K. Zhou, B. Sisman, R. Liu, and H. Li, “Emotional voice conversion: Theory, databases and esd,” *Speech Communication*, vol. 137, pp. 1–18, 2022.
- [27] A. Prasad, P. Jyothi, and R. Velmurugan, “An investigation of end-to-end models for robust speech recognition,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6893–6897.
- [28] A. Sherstinsky, “Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network,” *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.
- [29] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhota, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, “Superb: Speech processing universal performance benchmark,” in *Proc. Interspeech 2021*, 2021.
- [30] E. Vincent, S. Watanabe, J. Barker, and R. Marxer, “The 4th chime speech separation and recognition challenge,” *URL: http://spandh.dcs.shef.ac.uk/chime_challenge/* (last accessed on 1 August, 2018), 2016.