EFFICIENT ADAPTER TUNING OF PRE-TRAINED SPEECH MODELS FOR AUTOMATIC SPEAKER VERIFICATION

Mufan Sang, John H.L. Hansen

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, TX, USA
{mufan.sang, john.hansen}@utdallas.edu

ABSTRACT

With excellent generalization ability, self-supervised speech models have shown impressive performance on various downstream speech tasks in the pre-training and fine-tuning paradigm. However, as the growing size of pre-trained models, fine-tuning becomes practically unfeasible due to heavy computation and storage overhead, as well as the risk of overfitting. Adapters are lightweight modules inserted into pre-trained models to facilitate parameter-efficient adaptation. In this paper, we propose an effective adapter framework designed for adapting self-supervised speech models to the speaker verification task. With a parallel adapter design, our proposed framework inserts two types of adapters into the pre-trained model, allowing the adaptation of latent features within intermediate Transformer layers and output embeddings from all Transformer layers. We conduct comprehensive experiments to validate the efficiency and effectiveness of the proposed framework. Experimental results on the VoxCeleb1 dataset demonstrate that the proposed adapters surpass fine-tuning and other parameter-efficient transfer learning methods, achieving superior performance while updating only 5% of the parameters.

Index Terms— Speaker verification, pre-trained model, adapter, transfer learning, parameter-efficiency

1. INTRODUCTION

In recent years, we have seen the rapid development of speaker verification (SV) driven by deep learning. Various models and methods for SV have been introduced, encompassing different deep neural network (DNN) architectures [1, 2, 3], attention mechanisms [4], Transformer-based architectures [5, 6], and self-supervised SV systems [7, 8, 9]. Most of these works focus on utilizing task-specific datasets to train SV systems from scratch. Recently, the emergence of large-scaled pre-trained speech models has propelled the research in the field of speech processing. Taking the advantages of Transformer architecture, self-supervised learning (SSL), and increasingly large amounts of unlabeled data, pre-trained models exhibit strong generalization capabilities across various downstream speech tasks. Applying large-scale pre-trained speech models (e.g., HuBERT [10], WavLM [11]) to downstream tasks has remarkably improved performance over conventional models. The question of how to more efficiently utilize pre-trained models to improve the performance of downstream tasks remains an open area for investigation.

The pre-training and fine-tuning paradigm has become the most common approach for adapting pre-trained models to downstream tasks [11, 12, 13]. However, fine-tuning has drawn some issues due to two main reasons. Firstly, fine-tuning requires one to update all the model parameters, store and deploy a separate copy of the model parameters for each individual downstream task. As the size of pre-trained SSL models increases, fine-tuning becomes prohibitively costly in terms of training, storage, and deployment, rendering it practically infeasible. Secondly, pre-trained models are prone to overfitting when fine-tuned on limited amounts of data for downstream tasks, which degrades their generalization abilities. Therefore, parameter-efficient fine-tuning methods are crucial for large-scale pre-trained model adaptation. A simple and straightforward approach is linear probing, where the pre-trained model remains fixed and the stacked classification head is fine-tuned for each downstream task. However, it often results in unsatisfactory performance compared to full fine-tuning. More recently, adapters have drawn more and more attention for transferring knowledge from pre-trained models to downstream tasks. Adapters [14] were first proposed in the Natural Language Processing (NLP) field for model adaptation, which inserts lightweight modules with bottleneck architecture into Transformer layers after multi-head self-attention (MHSA) and feed-forward network (FFN) modules. A bottleneck layer consists of a down and up projection pair that shrinks and recovers the size of token hidden states. During fine-tuning, only the inserted adapters get updated and other parts of the model keep frozen. Some studies [15, 16, 17] explored the use of adapters to adapt pre-trained models to diverse speech processing tasks. However, most of them do not adequately utilize the information embedded in different layers of pre-trained models. Efficient methods for adapting pre-trained models to speaker verification are not well-studied.

In this paper, we propose an effective adapter framework that consists of two modules: the Inner-layer Adapter and the Inter-layer Adapter, aiming to efficiently transfer the universal knowledge of pre-trained SSL model to the speaker verification task. The proposed adapters learn task-specific knowledge for speaker verification by adapting latent features within intermediate Transformer layers and output embeddings from all Transformer layers of pre-trained model. Moreover, we introduce a parallel adapter design that inserts and sets adapters in parallel to the FFN of Transformer layers. A scaling operation is introduced to control adapter outputs, and balance task-agnostic and task-specific features learned from original FFN branches and adapter branches within Transformer blocks. Experimental results demonstrate that our proposed adapter-tuning method significantly outperforms other transfer learning methods and full fine-tuning while updating only 5.0% extra parameters. The primary contributions of this paper can be summarized as follows: (1) We propose an effective adapter framework that fully leverages speaker-related information embedded in different layers of pre-trained model. (2) We propose a parallel adapter design that helps the pre-trained model learn complementary task-specific knowledge. (3) We conduct comprehensive experiments to validate the efficiency and effectiveness of the proposed adapter framework.

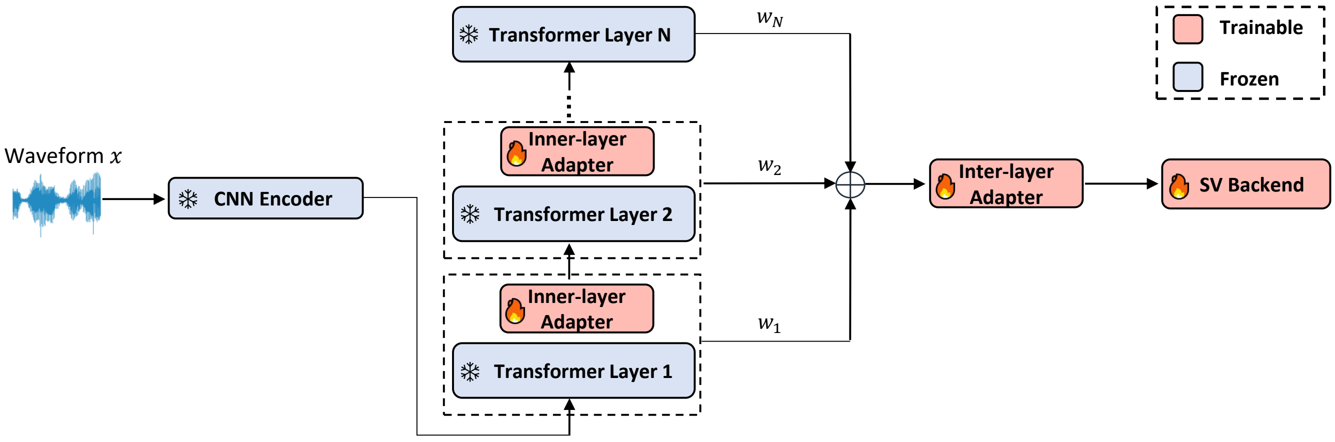


Fig. 1: Overview of the pre-trained model and the proposed adapter framework. During fine-tuning, the pre-trained model is frozen, only the Inner-layer Adapter, Inter-layer Adapter, and the SV backend are updated.

2. RELATED WORKS

2.1. Self-supervised Pre-trained Speech Models

Self-supervised learning (SSL) can utilize large amounts of unlabeled data to help models learn generic representations, thereby it has received increasing attention in the speech field. Recently, various SSL based pre-trained models and methods, including wav2vec [18], wav2vec 2.0 [19], HuBERT [10], and WavLM [11] have been proven effective on certain speech tasks [15, 17, 20]. Among them, WavLM was proposed to explore a full stack of speech tasks instead of focusing on specific tasks. It combines masked speech prediction and denoising during pre-training to learn not only knowledge related to automatic speech recognition (ASR) but also information about other non-ASR tasks. The pre-train and fine-tune paradigm has shown its success in adapting pre-trained models to downstream speech tasks. However, fine-tuning large-scale pre-trained models remains data-dependent and computationally expensive, limiting the broader application of SSL pre-trained models. Therefore, it is worthwhile to explore how to transfer the knowledge of pre-trained models to downstream tasks with lower computation and storage costs.

2.2. Adapter-based Tuning

Adapters [14] were initially introduced as an alternative approach for adapting large-scale pre-trained language models in NLP. Adapters modify the feature extractors by inserting some lightweight bottleneck modules without changing the parameters of pre-trained models. Adapter-based methods have proven to be comparable with full fine-tuning with much higher parameter efficiency, and sometimes perform slightly better in low-resource settings [21]. With the advantage, adapters have also been applied to computer vision tasks [22, 23].

Recently, adapters have also been introduced in speech processing tasks. In [24], researchers applied adapters to the RNN-T model for multilingual ASR. The work [25] proposed employing adapters to a speech Transformer to address the long-tail problem of multilingual ASR. In [15], adapters were applied to wav2vec 2.0 to increase the model's scalability to multiple languages. In [16], adapters were utilized to improve the domain adaptation of SSL models, including wav2vec 2.0 and HuBERT for child ASR. SimAdapter [26] was proposed for cross-lingual low-resource ASR. In [17, 27, 28], researchers explored the effectiveness of adapters for different downstream speech tasks beyond ASR (e.g., emotion recognition, speaker verification, intent classification). However, most of these studies do not sufficiently leverage the information embedded in different layers of pre-trained models. Thus, the goal of this study is to design an efficient and effective adapter framework for speaker verification task.

3. METHOD

We propose a novel adapter framework to efficiently transfer the knowledge of large pre-trained speech models to speaker verification task. In our framework, we insert two types of adapters into the pre-trained backbone model: (1) Inner-layer Adapter, inserted within the intermediate Transformer layers. (2) Inter-layer Adapter, inserted after the weighed-sum operation between the backbone and speaker verification backend. The overall framework is illustrated in Fig. 1.

3.1. Inner-layer Adapter and Inter-layer Adapter

Adapters are lightweight modules inserted into the Transformer layers of pre-trained models for adaptation. To preserve the generalization ability of pre-trained models, only adapters are fine-tuned, and the pre-trained model keeps frozen during training. To better utilize output representations from all intermediate layers, we propose the Inner-layer Adapter and Inter-layer Adapter, allowing the adaptation of latent features within intermediate Transformer layers and output embeddings from all Transformer layers.

In many studies [14, 15], adapters are inserted after both multi-head self-attention and feed-forward network. To improve parameter efficiency, we insert the Inner-layer Adapter after FFN only. The Inner-layer Adapter has a bottleneck structure consisting of a down-projection to reduce the hidden dimension d to bottleneck dimension $\hat{d}$ with parameter $\boldsymbol{W}_{\text{down}} \in R^{d\times\hat{d}}$, an up-projection with parameter $\boldsymbol{W}_{\text{up}} \in R^{\hat{d}\times d}$, a non-linear activation function between them, layer normalization (LN), and a residual connection. Given $\boldsymbol{x_i}$ as the input feature of FFN, the output of Inner-layer Adapter can be formulated as:

$$\tilde{\boldsymbol{z}}_i^s = \mathrm{FFN}\left(\boldsymbol{x_i}\right) + \mathrm{LN}(\boldsymbol{W}_{\text{up}} f\left(\boldsymbol{W}_{\text{down}}\,\mathrm{FFN}\left(\boldsymbol{x_i}\right)\right)) \qquad (1)$$

where f denotes the ReLU activation function.

The previous study [29] indicates that the output representations from lower layers of pre-trained models can contribute to better performance on various downstream speech tasks. Consequently, we add a group of trainable weights to average the output representations from all layers. The Inner-layer Adapters are integrated into intermediate Transformer layers for adapting latent features within layers explicitly. However, the interaction among all layers is ignored. To better adapt the pre-trained model and fully leverage the speaker-related information embedded in all layers, we propose the Inter-layer Adapter. As shown in Fig. 1, we insert the Inter-layer Adapter after the weighted sum operation to facilitate the model adaptation. The Inter-layer Adapter consists of a fully connected (FC) layer and a non-linear activation function with LN. Given the output representation from the i-th layer as $\boldsymbol{H}_i$, the output of the Inter-layer Adapter is computed as:

$$\tilde{\boldsymbol{H}} = \mathrm{LN}(f(\boldsymbol{W}_{\text{inter}}(\sum_{i=1}^{N} w_i \boldsymbol{H}_i))) \tag{2}$$

where $\boldsymbol{W}_{\text{inter}} \in R^{d\times e}$ denotes the FC layer of Inter-layer Adapter, d is the hidden dimension and e is the speaker embedding dimension, f denotes the ReLU activation function and w_i denotes the trainable weight for the i-th layer.

3.2. Parallel Adapter Design

Adapters are usually inserted sequentially after MHSA and FFN, and take their outputs as inputs for further computing. Inspired by [23, 21], we propose a parallel design for our adapters and illustrate it in Fig. 2. Unlike sequential design, the parallel adapter is integrated into an additional sub-branch for task-specific fine-tuning. The output of the parallel adapter is rescaled by a factor s and then added to the original branch through a residual connection. The scaling factor s is proposed to control the balance between the task-agnostic features obtained from the original frozen branch and the task-specific features obtained from the tunable adapter branch. This parallel design enables the pre-trained model to preserve its generalization capability, while the domain-specific features learned from the adapters can serve as a valuable complement for feature ensemble. For a specific input feature of FFN $\boldsymbol{x_i}$, the output of the parallel adapter is formulated as:

$$\tilde{\boldsymbol{z}_i^p} = \mathrm{LN}(\boldsymbol{W}_{\text{up}} f\left(\boldsymbol{W}_{\text{down}}\, \boldsymbol{x_i}\right)) \tag{3}$$

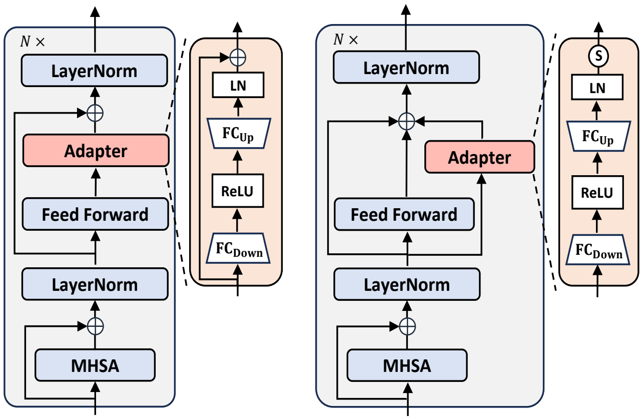
Accordingly, features from the adapter branch, FFN branch and the residual connection are fused, and the final output of the i-th Transformer layer is shown as:

$$\boldsymbol{H}_i = \mathrm{LN}\left(\mathrm{FFN}\left(\boldsymbol{x_i}\right) + s \cdot \tilde{\boldsymbol{z}_i^p} + \boldsymbol{x_i}\right) \tag{4}$$

4. EXPERIMENTS

4.1. Datasets

The speaker verification systems are trained and evaluated on the VoxCeleb1 [30] dataset, which contains 148,642 utterances from 1,211 speakers in the development set and 4,874 utterances from 40 speakers in the test set. We report the performance of systems on the VoxCeleb1-O evaluation trial.



(a) Sequential Adapter (b) Parallel Adapter

Fig. 2: Detailed architectures of (a) Sequential Adapter and (b) Parallel Adapter.

Moreover, we evaluate the proposed adapter framework in more challenging scenarios on the naturalistic 1^{st}48-UTD forensic corpus [31], where the total duration is 3.5 hours and more than 50% of utterances are shorter than 2 seconds. The training set consists of 3,755 utterances from 228 speakers, and the test set contains 882 utterances from 39 speakers. More details of the corpus can be found in [31].

4.2. Implementation Details

In this study, we employ the pre-trained WavLM Base+ as the backbone model. It comprises a convolutional feature encoder and 12 Transformer blocks equipped with gated relative position bias. Within the Transformer blocks, there are 8 attention heads, each with 768-dimensional hidden states. The WavLM Base+ has 94.70 million parameters. The speaker verification backend is composed of an average time pooling layer and two FC layers, with an embedding size of 512. The Inner-layer Adapter consists of two FC layers with a bottleneck dimension of 256, a ReLU activation function between them, LN and a residual connection. The Inter-layer Adapter consists of a FC layer with a hidden dimension of 512, followed by a ReLU activation function and LN.

The models are trained with the cross-entropy loss. We use the Adam [32] optimizer with an initial learning rate of 5e-4 for SV backend and 1e-5 for all other parameters. We apply a warm-up strategy at the first 38k steps and the learning rates decrease to 2.5e-5 for SV backend and 5e-7 for all other parameters in the remaining steps.

We compare our method with several transfer learning approaches, including full fine-tuning, linear probing, weighted sum, and two adapter-based methods: Houlsby adapter [14] and E-adapter and L-adapter (E+L adapter) [27]. In this study, in the case of full fine-tuning, we update all the parameters of WavLM Base+ but keep its convolutional encoder frozen. The weighted sum method is implemented as in [29], which has been demonstrated to be effective for speaker verification. To ensure a fair comparison, we reimplement the Houlsby adapter and the E+L adapter, and apply the same training configurations for all these methods.

We report the system performance using two evaluation metrics: Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) with $p_{target} = 0.05$.

Table 1: Performance of our method on VoxCeleb1. The second column represents the number of trainable parameters in the pre-trained model. Upper block: different tuning methods; Lower block: our proposed method and variants; FT: full fine-tuning; LP: linear probing.

Method	**# Params**	**VoxCeleb1-O**	
		EER (%)	**minDCF**
FT	85.1M (90.0%)	3.69	0.265
LP	0.0M (0.0%)	8.58	0.622
Weighted Sum	0.03M (0.03%)	4.81	0.324
Houlsby Adapter [14]	9.5M (10.1%)	3.67	0.244
E+L Adapter [27]	9.1M (9.6%)	2.77	0.195
Ours (Inner-layer)	4.4M (4.6%)	3.24	0.242
Ours (Inter-layer)	0.4M (0.4%)	3.01	0.215
Ours (Inner+Inter)	4.8M (5.0%)	**2.58**	**0.187**

4.3. Comparison among Transfer Learning Methods

In the experiments, we investigate the performance of our proposed adapter framework and evaluate it on the VoxCeleb1 dataset. We compare our method with other transfer learning methods, including fine-tuning, linear-probing, weighted sum, and two adapter-based methods: Houlsby adapter and E+L adapter. From Table 1, we can observe that the proposed Inner+Inter Adapter achieves the best performance, and outperforms fine-tuning and all other methods. Notably, compared to fine-tuning, our method improves the performance with relative 30.1% and 29.4% reductions in EER and minDCF respectively, by introducing only 5.0% of the pre-trained model parameters. The weighted sum fails to attain similar performance as fine-tuning, and linear probing performs significantly worse than fine-tuning. Compared to the other two adapter-based methods, our approach remarkably outperforms Houlsby adapter and E+L adapter while saving approximately 50% of the parameters, which sufficiently demonstrates the effectiveness and parameter efficiency of the proposed adapter framework.

Furthermore, in the lower section of Table 1, we can observe that the two variants, which insert the Inner-layer Adapter and Inter-layer Adapter separately, outperform both fine-tuning and Houlsby adapter. Compared to the weighted sum method, the Inter-layer Adapter inserts a single adapter after the weighted sum operation, achieving significantly better performance with a 37.4% reduction in EER on VoxCeleb1-O. Specifically, our Inter-layer Adapter surpasses even the Houlsby adapter and the other variant (Inner-layer) in terms of performance while saving 22× and 10× parameters, respectively. Consequently, experimental results indicate that the Inter-layer Adapter can serve as an essential module for adapter-based methods in speaker verification.

4.4. Ablation Study

We further conduct experiments to study the effectiveness of the parallel design. We compare the performance of our adapter framework using sequential and parallel insertion formulation. As presented in Table 2, the parallel adapter yields better performance than the sequential counterpart when using learnable and fixed scaling factors ($s \geq 0.5$). Additionally, we explicitly study the impact of scaling factor on the parallel adapter. In Table 2, we observe that our parallel adapter achieves the best performance with the fixed scale at 0.5, and using a learnable scale factor results in a slightly worse but comparable performance. Increasing or decreasing the value of s brings a performance drop. The reason could be that a smaller s might diminish the impact of task-specific features learned from adapters, and a larger s might weaken the contribution of task-agnostic features learned from the frozen pre-trained backbone. Based on the experimental results, we infer that the parallel adapter proves to be a more suitable choice for speaker verification.

Table 2: Performance of sequential adapter and parallel adapter with learnable and fixed scaling factors.

Scales	**EER** (%)	**minDCF**
Sequential	3.05	0.208
Learnable	2.76	0.190
0.05	3.45	0.236
0.1	3.32	0.224
0.5	**2.58**	**0.187**
1.0	2.69	0.195
1.5	2.79	0.194
2.0	2.82	0.202

Table 3: Performance of different transfer learning methods on 1^{st}48-UTD forensic dataset.

Systems	**EER** (%)	**minDCF**
FT	19.13	0.842
LP	18.52	0.829
Weighted Sum	18.10	0.819
Ours (Inner+Inter)	**16.75**	**0.734**

4.5. Evaluation in More Challenging Scenarios

In this section, we evaluate the performance of the proposed method on a more challenging dataset for forensic speaker recognition. As shown in Table 3, our method consistently outperforms fine-tuning, linear probing, and weighted sum, which illustrates the effectiveness and robustness of the proposed method even in more complex scenarios.

5. CONCLUSIONS

In this paper, we propose a parameter-efficient adapter-tuning framework aimed at effectively transferring the knowledge of pre-trained self-supervised speech models to speaker verification task. To sufficiently leverage the information embedded in all intermediate layers, our framework incorporates Inner-layer Adapters after the feed-forward network to adapt latent features within Transformer blocks. Additionally, it inserts an Inter-layer Adapter after the weighted sum operation to adapt the aggregated hidden representations extracted from all layers. The parallel design further improves model performance. Experimental results show that the proposed adapter outperforms fine-tuning and other transfer learning methods while updating only 5% extra parameters. The proposed framework can efficiently adapt the pre-trained model to the speaker verification task, leading to substantial reductions in computational and storage costs. We hope this work will inspire future research on parameter-efficient transfer learning of large-scale pre-trained speech models for speaker verification.

6. REFERENCES

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *ICASSP*, 2018, pp. 5329–5333.

[2] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot, “But system description to voxceleb speaker recognition challenge 2019,” *arXiv preprint arXiv:1910.12592*, 2019.

[3] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *arXiv preprint arXiv:2005.07143*, 2020.

[4] Jianfeng Zhou, Tao Jiang, Zheng Li, Lin Li, and Qingyang Hong, “Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function.,” in *Interspeech*, 2019, pp. 2883–2887.

[5] Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung yi Lee, and Helen Meng, “MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification,” in *Proc. Interspeech 2022*, 2022, pp. 306–310.

[6] Mufan Sang, Yong Zhao, Gang Liu, John HL Hansen, and Jian Wu, “Improving transformer-based networks with locality for automatic speaker verification,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, et al., “Unispeech-sat: Universal speech representation learning with speaker aware pre-training,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6152–6156.

[8] Mufan Sang, Haoqi Li, Fang Liu, Andrew O Arnold, and Li Wan, “Self-supervised speaker verification with simple siamese network and self-supervised regularization,” in *ICASSP*. IEEE, 2022, pp. 6127–6131.

[9] Haoran Zhang, Yuexian Zou, and Helin Wang, “Contrastive self-supervised learning for text-independent speaker verification,” in *ICASSP*. IEEE, 2021, pp. 6713–6717.

[10] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman 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.

[11] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong 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.

[12] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL*, 2019, pp. 4171–4186.

[13] Yigong Wang, Zhuoyi Wang, Yu Lin, Jinghui Guo, Sadaf Halim, and Latifur Khan, “Dual contrastive learning framework for incremental text classification,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 194–206.

[14] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly, “Parameter-efficient transfer learning for nlp,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[15] Bethan Thomas, Samuel Kessler, and Salah 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.

[16] Ruchao Fan and Abeer Alwan, “Draft: A novel framework to reduce domain shifting in self-supervised learning and its application to children’s asr,” *arXiv preprint arXiv:2206.07931*, 2022.

[17] Zih-Ching Chen, Yu-Shun Sung, and Hung-yi Lee, “Chapter: Exploiting convolutional neural network adapters for self-supervised speech models,” in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2023, pp. 1–5.

[18] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, “wav2vec: Unsupervised pre-training for speech recognition,” *Proc. Interspeech 2019*, pp. 3465–3469, 2019.

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

[20] Siyuan Shan, Yang Li, Amartya Banerjee, and Junier B Oliva, “Phoneme hallucinator: One-shot voice conversion via set expansion,” *arXiv preprint arXiv:2308.06382*, 2023.

[21] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig, “Towards a unified view of parameter-efficient transfer learning,” *International Conference on Learning Representations*, 2022.

[22] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal, “Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5227–5237.

[23] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo, “Adaptformer: Adapting vision transformers for scalable visual recognition,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 16664–16678, 2022.

[24] Anjuli Kannan, Arindrima Datta, Tara N Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Yonghui Wu, Ankur Bapna, Zhifeng Chen, and Seungji Lee, “Large-scale multilingual speech recognition with a streaming end-to-end model,” *arXiv preprint arXiv:1909.05330*, 2019.

[25] Genta Indra Winata, Guangsen Wang, Caiming Xiong, and Steven Hoi, “Adapt-and-adjust: Overcoming the long-tail problem of multilingual speech recognition,” *arXiv preprint arXiv:2012.01687*, 2020.

[26] Wenxin Hou, Han Zhu, Yidong Wang, Jindong Wang, Tao Qin, Renjun Xu, and Takahiro Shinozaki, “Exploiting adapters for cross-lingual low-resource speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 317–329, 2021.

[27] Shinta Otake, Rei Kawakami, and Nakamasa Inoue, “Parameter efficient transfer learning for various speech processing tasks,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[28] Junyi Peng, Themos Stafylakis, Rongzhi Gu, Oldřich Plchot, Ladislav Mošner, Lukáš Burget, and Jan Černockỳ, “Parameter-efficient transfer learning of pre-trained transformer models for speaker verification using adapters,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[29] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee, “SUPERB: Speech Processing Universal PERformance Benchmark,” in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[30] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, “Voxceleb: A large-scale speaker identification dataset,” *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[31] Mufan Sang, Wei Xia, and John H.L. Hansen, “Open-set short utterance forensic speaker verification using teacher-student network with explicit inductive bias,” *Proc. Interspeech 2020*, pp. 2262–2266, 2020.

[32] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.