SPARSELY SHARED LORA ON WHISPER FOR CHILD SPEECH RECOGNITION

Wei Liu[1], Ying Qin[2], Zhiyuan Peng[1], Tan Lee[1]

[1] Department of Electronic Engineering, The Chinese University of Hong Kong
[2] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China

ABSTRACT

Whisper is a powerful automatic speech recognition (ASR) model. Nevertheless, its zero-shot performance on low-resource speech requires further improvement. Child speech, as a representative type of low-resource speech, is leveraged for adaptation. Recently, parameter-efficient fine-tuning (PEFT) in NLP was shown to be comparable and even better than full fine-tuning, while only needing to tune a small set of trainable parameters. However, current PEFT methods have not been well examined for their effectiveness on Whisper. In this paper, only parameter composition types of PEFT approaches such as LoRA and Bitfit are investigated as they do not bring extra inference costs. Different popular PEFT methods are examined. Particularly, we compare LoRA and AdaLoRA and figure out the learnable rank coefficient is a good design. Inspired by the sparse rank distribution allocated by AdaLoRA, a novel PEFT approach Sparsely Shared LoRA (S2-LoRA) is proposed. The two low-rank decomposed matrices are globally shared. Each weight matrix only has to maintain its specific rank coefficients that are constrained to be sparse. Experiments on low-resource Chinese child speech show that with much fewer trainable parameters, S2-LoRA can achieve comparable in-domain adaptation performance to AdaLoRA and exhibit better generalization ability on out-of-domain data. In addition, the rank distribution automatically learned by S2-LoRA is found to have similar patterns to AdaLoRA's allocation.

Index Terms— Parameter-efficient fine-tuning, Automatic speech recognition, Child speech, Whisper, LoRA

1. INTRODUCTION

Automatic speech recognition (ASR) nowadays has become very powerful. Benefiting from large amounts of training data [1, 2] , representative ASR models, like Whisper [3], USM [4], and MMS [5], demonstrate remarkable multi-task and multi-lingual recognition capabilities. It is claimed that Whisper can perform zero-shot speech recognition across different languages and domains. Meanwhile, recognition of low-resource speech, such as child speech, disorder speech, and speech from the endangered languages, has persistently presented formidable challenges [6–8]. Consequently, it becomes imperative to examine the zero-shot performance of Whisper on these data. More importantly, should a performance gap be identified, it is essential to explore efficient adaptation strategies. Chinese child speech, as a representative type of low-resource speech, is leveraged in this work to probe the performance of Whisper and explore effective adaptation approaches.

The common practice for adaptation is to collect in-domain data and conduct model fine-tuning. However, two widely concerned issues exist. (1) The target domain data for low-resource speech is very limited. (2) With the larger model, full fine-tuning becomes increasingly computational costing, and the fine-tuned model tends to overfit. Recently, parameter-efficient fine tuning (PEFT) has attained great attention in the NLP community [9–11], mainly for large language models (LLMs). PEFT approaches can be roughly divided into three categories [9]. (a) *function composition*: like adapter [12]; (b) *input composition*: like prompt tuning and prefix tuning [13, 14]; (c) *parameter composition*: Low-rank adaptation (LoRA) [15], BitFit [16], etc [17–19]. These approaches introduce a small proportion of learnable parameters in different forms to the original model. Adapters directly insert lightweight neural network modules into the pre-trained model, which changes the model architecture. Prompt tuning and prefix tuning add leading learnable embeddings to the input sequence, resulting in the increased inference cost due to longer input sequences. Different from adapters and prompt tuning, *parameter composition*'s PEFT carries no additional inference cost once it has finished training. This type of approach enables the trainable parameters to be seamlessly integrated into the original weights, thereby promoting *parameter composition* as a prevailing fine-tuning strategy.

LoRA, as the most representative *parameter composition* approach, has several variants, particularly, AdaLoRA [17]. It is known that LoRA manually specifies a fixed rank for all the weight matrices, overlooking the varying importance between different weights. AdaLoRA addresses the issue with an importance-aware rank allocator. It maintains an overall rank budget and dynamically distributes the ranks to weight matrices according to an importance metric.

In this paper, several research questions are raised and addressed: (1) Why do we need PEFT? The effectiveness of these *parameter composition* types of PEFT methods are examined on Whisper adaptation to child speech. (2) What makes AdaLoRA superior to LoRA? A series of ablation studies are conducted to verify the importance of the rank coefficient design. (3) To what level the trainable parameters can be reduced while keeping performance? Inspired by the sparsity property of AdaLoRA's rank distribution, the Sparsely Shared LoRA (S2-LoRA) is proposed. The two low-rank matrices of LoRA are module-wise globally shared. Each weight matrix within the module only needs to adjust its specific sparse coefficient vector to control each rank's importance, achieving the same purpose of adaptive rank allocation but with much fewer parameters (only 0.02%). Experiments on low-resource Chinese child speech further demonstrate the effectiveness of S2-LoRA.

2. S2-LORA ON WHISPER MODEL

This section starts with a brief introduction of Whisper, followed by the formulations of LoRA, AdaLoRA, and the proposed S2-LoRA.

2.1. Whisper

Whisper [3] has a typical transformer-based encoder-decoder architecture that is designed for multiple speech tasks, including multilin-
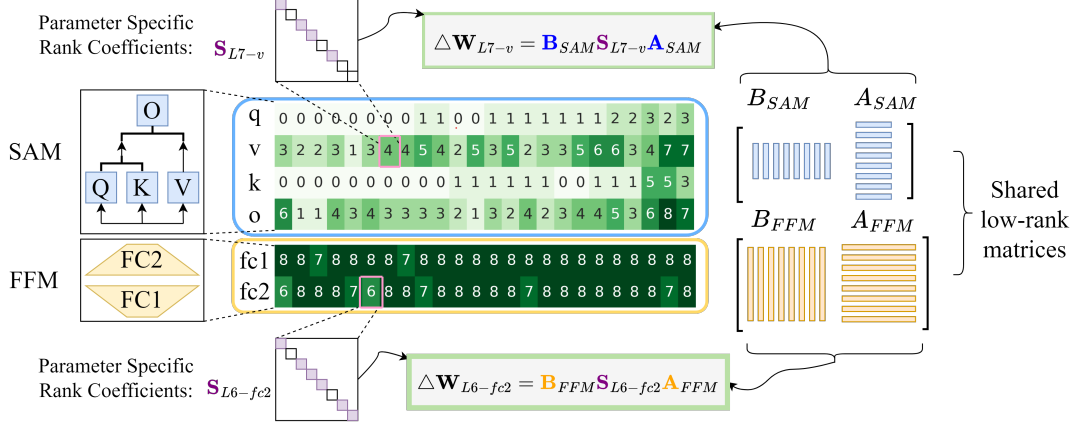
Fig. 1: The overall illustration of S2-LoRA. Numbers in the green matrices denote the ranks distributed across all weight matrices of the Whisper encoder. The maximum shared rank is 8. The Whisper *medium*'s encoder with 24 layers is shown as an example where SAM and FFM denote self-attention and feed-forward modules, the decoder operates similarly while including another cross-attention module (CAM).

gual ASR, multilingual speech translation, language identification, etc. The transformer [20] structure consists of three submodules as building blocks, namely the self-attention module (SAM), feed-forward module (FFM), and cross-attention module (CAM). The encoder layer comprises SAM and FFM, while the decoder layer includes all of the three. The learnable weight matrices in both SAM and CAM are $\{\mathbf{W_q}, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o\}$, which represents the query, key, value, and output projection matrices, respectively. In FFM, $\{\mathbf{W}_{fc1}, \mathbf{W}_{fc2}\}$ are the projection matrices of the fully-connected layers.

Whisper is trained on around 680,000 hours of weakly-supervised speech data collected from the Internet, with the performance approaching human-level accuracy and robustness.

2.2. LoRA

LoRA [15] was first proposed in NLP to efficiently adapt large language models (LLMs) to specific domains or downstream tasks. It was found that the weights of pre-trained LLMs tend to reside in a low intrinsic dimensional space [21]. Inspired by the observation, LoRA freezes the original weights and only updates the low-rank incremental weight matrices. Specifically, consider the i-th linear projection forward, $f_i(\mathbf{x}) = \mathbf{x}\mathbf{W}_i^T + \mathbf{b}_i$, where $\mathbf{W}_i \in \mathbb{R}^{d_2 \times d_1}$ and $\mathbf{b}_i \in \mathbb{R}^{d_2}$ denotes the frozen weight and bias. By applying LoRA, the forward is modified as,

$$f_i(\mathbf{x}) = \mathbf{x}(\mathbf{W}_i + \triangle\mathbf{W}_i)^T + \mathbf{b}_i; \ \ \triangle\mathbf{W}_i = \mathbf{B}_i\mathbf{A}_i, \quad (1)$$

where $\mathbf{B}_i \in \mathbb{R}^{d_2 \times r}$ and $\mathbf{A}_i \in \mathbb{R}^{r \times d_1}$ are the two trainable rank-decomposed matrices, with the rank $r \ll \min\{d_1, d_2\}$.

2.3. AdaLoRA

The importance of weight parameters to the performance varies across different layers and modules. Intuitively, some weight matrices should be allocated with higher ranks than others in adaptation. However, LoRA specifies a fixed rank for all the weight matrices, which overlooks the varying importance of weights and can be suboptimal. AdaLoRA [17] addresses this issue with an importance-aware rank allocation method. It comes with two modifications to LoRA. First, the incremental update $\triangle\mathbf{W}_i$ is parameterized in the form of singular value decomposition (SVD), i.e.,

$$\triangle\mathbf{W}_i = \bar{\mathbf{B}}_i\mathbf{\Lambda}_i\bar{\mathbf{A}}_i; \ \ \bar{\mathbf{B}}_i^T\bar{\mathbf{B}}_i = \bar{\mathbf{A}}_i\bar{\mathbf{A}}_i^T = \mathbf{I}, \quad (2)$$

where $\bar{\mathbf{B}}_i \in \mathbb{R}^{d_2 \times r}$ and $\bar{\mathbf{A}}_i \in \mathbb{R}^{r \times d_1}$ contain the left/right singular vectors of $\triangle\mathbf{W}_i$. $\mathbf{I}$ denotes the identity matrix. and the $\{\bar{\mathbf{B}}_i, \bar{\mathbf{A}}_i\}$ are enforced to be orthogonal. The diagonal matrix $\mathbf{\Lambda}_i \in \mathbb{R}^{r \times r}$ contains r singular values, with $r \ll \min\{d_1, d_2\}$. Only $\bar{\mathbf{B}}_i, \bar{\mathbf{A}}_i, \mathbf{\Lambda}_i$ are trainable.

Second, in adaptation, AdaLoRA dynamically allocates an overall rank budget to its update matrices $\{\triangle\mathbf{W}_i\}$. This is achieved by iteratively masking out less important singular values after every gradient update step. A sensitivity-based importance metric [22] is utilized to measure and sort the importance of the k-th triplet $\{\mathbf{\Lambda}_i^{k,k}, \bar{\mathbf{B}}_i^{*,k}, \bar{\mathbf{A}}_i^{k,*}\}$ of the i-th weight matrix $\mathbf{W}_i$, which takes account of both singular values and vectors. The non-zero $\mathbf{\Lambda}_i^{k,k}$ acts as the rank coefficient to control the allocated rank budget.

2.4. Proposed S2-LoRA

Our preliminary experiments on AdaLoRA (See Fig. 3) show that (1) it tends to allocate significantly fewer rank budgets to $\{\mathbf{W}_q, \mathbf{W}_k\}$ than those to $\{\mathbf{W}_v, \mathbf{W}_o\}$ in both SAM and CAM; (2) the sparsity of rank budgets among layers exists. In this regard, parameters for adaptation can be further reduced. It may not be necessary to learn the rank-decomposed matrices $\mathbf{B}_i, \mathbf{A}_i$ separately for each weight matrix $\mathbf{W}_i$. They can be tied and shared across layers and modules. A novel PEFT approach, **Sparsely Shared Low-rank Adaptation** (S2-LoRA), is therefore proposed in this work. S2-LoRA has $\mathbf{B}$ and $\mathbf{A}$ globally shared. The i-th weight matrix only stores a single trainable rank coefficient vector $\mathbf{s}_i$. In addition, S2-LoRA introduces an L1 constraint on the rank vector for sparsity. The orthogonal constraint proposed in Eq. 2 is discarded. To this end, the incremental update for each weight matrix $\mathbf{W}_i$ is given as:

$$\triangle\mathbf{W}_i = \mathbf{B}\mathbf{S}_i\mathbf{A}; \ \ \text{s.t.} \ ||\mathbf{S}_i||_1 < \epsilon, \quad (3)$$

where $\mathbf{S}_i = diag(\mathbf{s}_i)$ and ϵ denotes an empirically small value. The objective function of S2-LoRA can be written as:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{CE}}(\mathcal{D}|\mathbf{B}, \mathbf{A}, \{\mathbf{S}_i\}_{i=1}^N) + \alpha_1 * \frac{1}{N}\sum_{i=1}^{N}||\mathbf{S}_i||_1$$
$$+ \alpha_2 * \frac{1}{2r}\sum_{k=1}^{r}(||\mathbf{B}^{*,k}||_2 + ||\mathbf{A}^{k,*}||_2), \quad (4)$$

where $\mathcal{L}_{CE}$ represents the original cross-entropy loss for fine-tuning Whisper. $\mathcal{D}$ is the adaptation data. N denotes the number of trainable weight matrices. α_1 controls the sparsity penalty on $\mathbf{S}_i$. α_2 regularizes $\{\mathbf{B}, \mathbf{A}\}$ by L2 norm, avoiding the numeric scaling issue in training.

In practice, as shown in Fig. 1, the $\{\mathbf{B}, \mathbf{A}\}$ of S2-LoRA are not globally shared across the whole Whisper model. The functionality of the encoder and decoder in Whisper are different. The three basic blocks, SAM, CAM, and FFM, act as various modeling roles. Consequently, we split Whisper into 5 modules, namely, Enc-SAM, Enc-FFM, Dec-SAM, Dec-CAM, and Dec-FFM. $\{\mathbf{B}, \mathbf{A}\}$ are tied only within each module. Fig. 1 gives the details of how S2-LoRA is applied to Enc-SAM and Enc-FFM.

3. EXPERIMENTS SETUP

3.1. Datasets

The Chinese child speech for Whisper adaptation comes from the subsets of CSRC-2021 dataset [1] [23]. CSRC-2021 contains 28.6 hours of child-read speech (zh-C1) and 29.5 hours of child conversational speech (zh-C2). The zh-C1 is the training set in the target domain, which has around 30K utterances. To investigate the effect of different amounts of adaptation data, the zh-C1 was split into three sets: namely zh-C1-1K, zh-C1-10K, and the original zh-C1-30K. In the evaluation stage, three Chinese Mandarin test sets were prepared. They are child read/conversational speech from the evaluation sets of CSRC-2021 and adult read speech from Aishell1 [24] test set (zh-A1). In addition, Cantonese (zh-HK) and English (en) from Common Voice 11.0 [25] test sets are utilized to evaluate the cross-lingual generalization. Each of the above test sets contains randomly selected 1,000 utterances for fast evaluation.

3.2. PEFT Configurations

The PEFT methods we compared belong to the category of *parameter composition*. (1) BitFit [16]: update all the bias parameters of the model. (2) IA3 [19]: introduce three learned vectors per layer to elementwisely multiply with the projection output from $\{\mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_{fc2}\}$. (3) LoRA [15]: model incremental updates of $\{\mathbf{W}_q, \mathbf{W}_v\}$ with the low-rank matrices where rank $r = 8$. (4) GLoRA [18]: a generalized LoRA that gives a unified formulation encompassing all tunable dimensions. In GLoRA, $\Delta\mathbf{W}_i = \mathbf{W}_i\mathbf{A}_i + \mathbf{B}_i$, $\Delta\mathbf{b}_i = \mathbf{C}_i\mathbf{W}_i + \mathbf{D}_i\mathbf{b}_i + \mathbf{E}_i$, where $\{\mathbf{A}_i, \mathbf{B}_i, \mathbf{C}_i, \mathbf{D}_i, \mathbf{E}_i\}$ are trainable parameters that can be further low-rank decomposed for matrices ($r = 8$). Note that GLoRA becomes LoRA if only having $\mathbf{B}$ and becomes BitFit if only $\mathbf{E}$ exists. (5) AdaLoRA [17]: dynamically allocate the rank budget to weight matrices of $\{\mathbf{W}_q, \mathbf{W}_v\}$, where the initial rank is set to 12 and target rank is 8. (6) S2-LoRA: propose to sparsely share the low-rank components ($r = 8$), every weight matrix only need to maintain a rank coefficient vector $\mathbf{s}$. In Eq. 4, the sparsity weight α_1 and regularization weight α_2 is set to 0.05 and 0.1 respectively. All the above methods are implemented with the open-sourced Huggingface PEFT library [2].

3.3. Whisper Tuning

Whisper has five different sizes, namely *tiny* (39M), *base* (74M), *small* (244M), *medium* (769M), and *large* (1.5B). The Huggingface
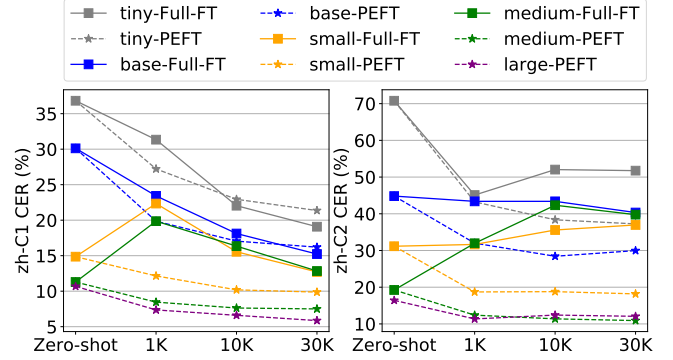

Fig. 2: Full-FT vs. PEFT on Whisper with different amount of adaptation data zh-C1. Full fine-tuning on Whisper *large* (large-Full-FT) is not conducted due to limited GPU memory.

trainer is used to fine-tune the model. Since the speech input to Whisper would be padded to 30 seconds, $batch\_size=2$ is used with 8 gradient accumulations. Mixed precision training is enabled. The number of epochs is set to 3. The learning rate is 1e-3 in PEFT mode while using 1e-4 under full fine-tuning. Greedy decoding is applied for ASR inference. Character error rate (CER) is used to evaluate Chinese languages and word error rate (WER) is used for English.

4. RESULTS AND ANAYSIS

4.1. Why we need PEFT on Whisper?

To prove the necessity of PEFT for Whisper, the performance comparison between Full-FT and PEFT is carried out. AdaLoRA ($r = 8$) is used as a representative PEFT method in this experiment.

As shown in Fig. 2, five different sizes of Whisper are evaluated for Chinese child speech recognition. The x-axis represents the varying amount of adaptation data from zh-C1-1K to zh-C1-30K. The CER in the y-axis indicates the recognition performance on evaluation sets of zh-C1 (left) and zh-C2 (right). Full-FT is plotted by solid curves and PEFT use the dashed ones.

We can notice that Whisper with PEFT clearly outperforms that of using Full-FT on the zh-C1 in-domain evaluation and keeps a similar error reduction trend on zh-C2. Particularly for Full-FT, it can be observed only the *tiny* and *base* Whispers show the improvements after adaptation, while the *small* and *medium* all made a large performance degradation, especially for the case of 1K utterances. However, PEFT gives consistent improvements across different model sizes and different amounts of adaptation data. In addition, PEFT on *medium* (a half size of *large*) surpasses a lot over those with smaller model sizes, approaching the level of *large*. If not specified otherwise, Whisper *medium* is utilized in the following experiments.

4.2. What makes AdaLoRA superior to LoRA?

AdaLoRA always gives significantly better PEFT performance than LoRA in our experiments, which is in line with previous findings [17]. As mentioned in section 2.3, AdaLoRA mainly comes with two modifications to the vanilla one: (1) orthogonal regularization and learnable rank coefficients via SVD; (2) importance-aware rank allocation. To investigate the main reason that makes AdaLoRA superior to LoRA, the following ablation experiments are performed:(1) remove orthogonal regularization ($orth$) from AdaLoRA; (2) remove importance-aware rank allocation ($alloc$) from AdaLoRA; (3) remove $orth$ and $alloc$ from AdaLoRA, only learnable rank coeffi-

Table 1: The CER(%) performance comparison between AdaLoRA and LoRA on the evaluation sets of zh-C1 and zh-C2, where the adaptation data is zh-C1-1K. r denotes the rank in LoRA or target rank in AdaLoRA. Initial rank of AdaLoRA is set to 48 for $r = 32$, otherwise being 12. The *orth* and *alloc* represent the orthogonal regularization and rank allocation in AdaLoRA, respectively.

PEFT	zh-C1			zh-C2		
	$r=1$	$r=8$	$r=32$	$r=1$	$r=8$	$r=32$
Zero-shot		11.29			19.25	
LoRA	10.63	9.32	9.22	17.87	14.73	14.45
AdaLoRA	8.93	**8.46**	8.79	14.26	13.73	13.48
w/o orth	8.81	8.62	8.78	14.46	12.56	13.35
w/o alloc	8.53	8.53	8.78	13.80	13.80	13.38
w/o both	8.64	8.64	8.76	12.52	12.52	13.41
α-LoRA	8.86	**8.77**	8.47	13.75	13.68	12.70

cients left; (4) add rank coefficients to LoRA, since rank coefficients used for rank allocation can not be removed from AdaLoRA. As illustrated in Tab. 1, Whisper *medium* was used to adapt on zh-C1-1K by different ablation variants of AdaLoRA.

It can be seen that three different rank values $\{1, 8, 32\}$ were tested and AdaLoRA all obviously surpasses LoRA. By removing *orth*, *alloc* and both of them from AdaLoRA, clear performance degradation does not happen, even bringing some performance gains. This shows that *orth* and *alloc* are not the key factors for the performance improvement of AdaLoRA compared to LoRA.

To verify the benefits of learnable rank coefficients, a trainable scalar α was introduced to LoRA, named α-LoRA for experiments where $\Delta\mathbf{W}_i = \alpha\mathbf{B}_i\mathbf{A}_i$. For vanilla LoRA, $\mathbf{B}_i$ has to be initialized as zero to make sure $\Delta\mathbf{W}_i = \mathbf{0}$ at the start. With $\alpha = 0$, $\{\mathbf{B}_i, \mathbf{A}_i\}$ can be initialized by a normal distribution, similar to the initialization way of AdaLoRA. It can be seen from Tab. 1, α-LoRA largely improves the vanilla LoRA across all ranks, illustrating the learnable rank coefficients are the most important design.

4.3. Comparing S2-LoRA with other PEFTs

The proposed S2-LoRA is inspired by the design of rank coefficients (See section 4.2) and the sparse rank distribution of AdaLoRA. Tab. 2 evaluates S2-LoRA with several other *parameter composition* types of PEFT methods.

AdaLoRA with 0.46% trainable parameters achieves the best in-domain performance while GLoRA with 0.81% ones exhibits the best generalization ability. IA3 with 0.03% trainable parameters performs the worst among PEFT approaches. Compared to AdaLoRA, the proposed S2-LoRA is competitive on in-domain data and performs better under out-of-domain conditions, while using 20x fewer trainable parameters (0.02%). Notably, S2-LoRA has trainable parameters even fewer than IA3 and BitFit.

For cross-lingual evaluation, all PEFT methods achieved positive performance gains in Mandarin (zh-C1, zh-C2, and zh-A1), while suffering from increased CERs (around 2-3%) in Cantonese (zh-HK). This may be due to the conflict in the token modeling of Whisper where the same language ID *zh* is used for both Mandarin and Cantonese. Performance improvement in Mandarin would cause a performance reduction in Cantonese. With different language IDs (*en* vs. *zh*), all PEFT methods are found to be able to bring somewhat improvements in English while adapting to Mandarin.

Table 2: The CER(%) performance comparison between S2-LoRA and other PEFT methods where rank=8. Whisper *medium* with zh-C1-1K as the adaptation data was tested on in-domain data (zh-C1) and out-of-domain data (others). WER(%) is used for English (en).

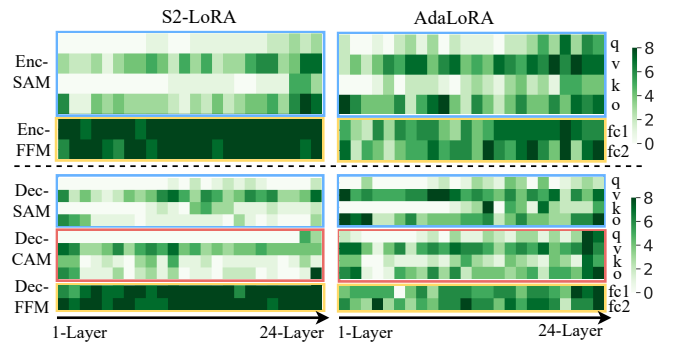| Methods	#Trainable Params	in-domain zh-C1	out-of-domain			
			zh-C2	zh-A1	zh-HK	en
Zero-shot	-	11.29	19.25	9.79	31.14	7.93
Full-FT	100%	19.86	31.97	28.90	57.70	11.47
GLoRA	0.81%	8.85	**13.04**	6.71	**31.79**	6.69
AdaLoRA	**0.46%**	**8.46**	13.73	7.65	34.79	7.57
LoRA	0.31%	9.32	14.73	7.52	33.30	7.35
BitFit	0.08%	8.72	13.38	7.85	34.77	7.50
IA3	0.03%	10.24	15.73	9.55	34.26	**6.62**
α-LoRA	0.31%	8.77	13.68	**6.37**	33.01	6.73
S2-LoRA	**0.02%**	**8.67**	13.19	6.81	33.71	6.72

Fig. 3: The rank distribution of S2-LoRA ($r = 8$) and AdaLoRA. For a fair comparison, the initial rank of AdaLoRA is set to 8, and the target rank is 4. In S2-LoRA, the rank coefficients less than $1e - 4$ are masked out. The row order in SAM/CAM follows $\{\mathbf{W}_q, \mathbf{W}_v, \mathbf{W}_k, \mathbf{W}_o\}$, while in FFM follows $\{\mathbf{W}_{fc1}, \mathbf{W}_{fc2}\}$.

Furthermore, the rank distribution of S2-LoRA is found to have similar patterns with AdaLoRA as shown in Fig. 3. Each element of the distribution matrix denotes the allocated rank of the incremental update matrix $\Delta\mathbf{W}_i$. The elements with brighter colors represent lower ranks allocated, thus being less important. In the SAM/CAM of both encoder and decoder, the rank allocated to $\{\mathbf{W}_q \mathbf{W}_k\}$ are mostly zero, suggesting that they are much less important than $\{\mathbf{W}_v, \mathbf{W}_o\}$. This phenomenon exists in AdaLoRA and S2-LoRA, indicating that S2-LoRA learns to identify important blocks and allocate ranks to them for improving adaptation performance.

5. CONCLUSIONS

This paper presents a novel S2-LoRA approach to adapting Whisper. S2-LoRA improves AdaLoRA with much fewer trainable parameters than most existing PEFT methods such as GLoRA, BitFit, etc. It benefits from the design of the sparse learnable rank coefficients and shared rank-decomposed matrices. Experiments carried out on low-resource Chinese child speech demonstrate the effectiveness of our proposed approach, showing that S2-LoRA achieves adaptation performance comparable to AdaLoRA and noticeably better cross-domain generalizability while retaining only 0.02% trainable parameters. Though using child speech as a study case, the proposed S2-LoRA is general and can benefit adaptation in other low-resource speech recognition scenarios.

6. REFERENCES

[1] 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.

[2] 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, "XLS-R: self-supervised cross-lingual speech representation learning at scale," in *Proc. of Interspeech*. 2022, pp. 2278–2282, ISCA.

[3] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[4] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al., "Google USM: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[5] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al., "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[6] Prashanth Gurunath Shivakumar and Shrikanth Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech & Language*, vol. 72, pp. 101289, 2022.

[7] Zi Ye, Shoukang Hu, Jinchao Li, Xurong Xie, Mengzhe Geng, Jianwei Yu, Junhao Xu, Boyang Xue, Shansong Liu, Xunying Liu, et al., "Development of the cuhk elderly speech recognition system for neurocognitive disorder detection using the dementiabank corpus," in *Proc. of ICASSP*. IEEE, 2021, pp. 6433–6437.

[8] Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020.

[9] Sebastian Ruder, Jonas Pfeiffer, and Ivan Vulić, "Modular and parameter-efficient fine-tuning for nlp models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, 2022, pp. 23–29.

[10] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al., "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220–235, 2023.

[11] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig, "Towards a unified view of parameter-efficient transfer learning," in *Proc. of ICLR*. 2022, OpenReview.net.

[12] 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.

[13] Brian Lester, Rami Al-Rfou, and Noah Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. of EMNLP*. 2021, pp. 3045–3059, Association for Computational Linguistics.

[14] Xiang Lisa Li and Percy Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proc. of ACL*. 2021, pp. 4582–4597, Association for Computational Linguistics.

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," in *Proc. of ICLR*. 2022, OpenReview.net.

[16] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," in *Proc. of ACL*. 2022, pp. 1–9, Association for Computational Linguistics.

[17] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao, "Adaptive budget allocation for parameter-efficient fine-tuning," in *Proc. of ICLR*. 2023, OpenReview.net.

[18] Arnav Chavan, Zhuang Liu, Deepak Gupta, Eric Xing, and Zhiqiang Shen, "One-for-all: Generalized lora for parameter-efficient fine-tuning," *arXiv preprint arXiv:2306.07967*, 2023.

[19] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1950–1965, 2022.

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

[21] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer, "Intrinsic dimensionality explains the effectiveness of language model fine-tuning," in *Proc. of ACL*. 2021, pp. 7319–7328, Association for Computational Linguistics.

[22] Qingru Zhang, Simiao Zuo, Chen Liang, Alexander Bukharin, Pengcheng He, Weizhu Chen, and Tuo Zhao, "Platon: Pruning large transformer models with upper confidence bound of weight importance," in *International Conference on Machine Learning*. PMLR, 2022, pp. 26809–26823.

[23] Fan Yu, Zhuoyuan Yao, Xiong Wang, Keyu An, Lei Xie, Zhijian Ou, Bo Liu, Xiulin Li, and Guanqiong Miao, "The slt 2021 children speech recognition challenge: Open datasets, rules and baselines," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 1117–1123.

[24] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proc. of O-COCOSDA*. IEEE, 2017, pp. 1–5.

[25] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. of LREC*. 2020, pp. 4218–4222, European Language Resources Association.