Adapting Whisper Models using LoHA for Robust Recognition of Children's Speech

1st Ankita
Department of ECE
National Institute of Technology Patna
ankita.ph21.ec@nitp.ac.in

2nd S. Shahnawazuddin
Department of ECE
National Institute of Technology Patna
s.syed@nitp.ac.in

***Abstract*—Fine-tuning of large pre-trained models, such as Whisper, has gained prominence in the area of speech processing. Since full fine-tuning requires a large amount of data as well as high-ended computational resources, parameter efficient fine-tuning (PEFT) has been the preferred choice among researchers. During the past few years, several PEFT techniques have been developed and have been observed to be extremely effective. Motivated by the success of PEFT, in this paper, we have investigated and documented the efficacy of Low-Rank Hadamard Product Adaptation (LoHA) of Whisper models for children's automatic speech recognition (ASR) task especially in limited data scenario. We have also compared LoHA with a few other PEFT variants. Even though LoHA has been explored for signal processing and federated learning tasks, its impact on children's ASR has not yet been studied. Experimental results presented in this paper indicate that applying LoHA significantly influences ASR performance, achieving a word error rate of 3.0%.**

***Index Terms*—Children's speech recognition, Fine-tuning, PEFT, LoRA, LoHA.**

I. Introduction

Extensive use of voice-based technologies has led to a greater demand for enhanced automatic speech recognition (ASR) systems for children. Children prefer speaking to communicate with people as well as with electronic devices since their reading and writing skills are not yet fully developed. This calls for a need to develop effective speech recognition systems tailored to the needs of children. However, creating efficient children's ASR systems is a challenging task due to the fact that children's speech exhibits a higher degree of inter- and intra-speaker variability when compared to adult speakers [1]. Scarcity of freely available children's speech data further exacerbate the problem since the modern encoder-decoder architectures require huge amounts of domain-specific data [2].

Recently, transfer learning (TL) techniques have introduced the concept of domain adaptation through fine-tuning (FT) of large pre-trained models [3]. Ready to use pre-trained models, originally trained on very large and diverse data-sets, were found to be promising for several machine learning tasks including ASR [4]–[6]. Works employing one such pre-trained model, named Whisper, demonstrated robust performance across various languages and accents for ASR [7]. Adapting pre-trained models to children's speech was also observed to be effective in enhancing the recognition performance with respect to child speakers [8], [9]. It is to note that, full FT requires a significantly large amount of domain-specific data as well as high-ended computational resources. Furthermore, full fine-tuning is also prone to over-fitting.

The issues related with full FT were overcome by a set of techniques collectively known as parameter-efficient fine-tuning (PEFT). These techniques avoid adjusting the entire set of the model parameters, making it a more resource-efficient way to adapt pre-trained models. Since its inception, several PEFT variants have been proposed such as: *low-rank adaptation* (LoRA), *adaptive* LoRA (AdaLoRA), *quantized* LoRA (qLoRA), and *low-rank Hadamard product adaptation* (LoHA) [10]–[13]. Some of the recent works have investigated the effectiveness of PEFT variants in children's ASR task [14]–[17]. These studies have shown the worth of PEFT in improving children's speech recognition without the need for extensive retraining. Motivated by the earlier findings, in this paper, we have experimentally studied and documented the effectiveness of LoHA in the context of children's ASR task in resource-constrained settings. We have also experimentally compared LoHA with some of the other PEFT variants, namely LoRA, qLoRA, and AdaLoRA. It is worth highlighting here that, to the best of our knowledge, LoHA has not yet been employed in children's ASR task. Consequently, the study presented in this paper helps in contrasting the effectiveness of LoHA with several LoRA variants. This work also provides some important insights on the suitability of LoHA for children's ASR tasks.

The remainder of this paper is organised as follows: A brief review of the existing works along with the motivation behind this study are given in Section II. In Section III, the explored variants of LoRA are discussed. An overview of LoHA is presented in Section IV. The experimental studies exploring the effectiveness of applying LoHA on different Whisper models are given in Section V. Finally, the paper is concluded in Section VI.

II. Related works and motivation

The work reported in [18] underscored the effectiveness of employing LoRA on Whisper in order to address the key challenges related to multilingual ASR. By reducing language interference and facilitating seamless language expansion, the approach was noted to enhance ASR performance, making it a promising solution for real-world multilingual applications.

In [19], the potential of LoRA in improving ASR personalization with limited speaker-specific data was highlighted. By efficiently fine-tuning large models with minimal additional parameters, the reported approach was observed to enhance adaptability of ASR system, making it a viable solution for real-world voice assistant applications. In [20], the effectiveness of few-shot learning and LLM fine-tuning through LoRA and qLoRA techniques for accented speech classification was studied. The approach outlined in that study significantly enhanced classification accuracy, making it a promising solution for handling speech variability in the context of resource-poor languages. In [21], the potential of few-shot learning and LLM fine-tuning through qLoRA was presented for improving accented speech classification. The reported approach offered a scalable solution for resource-poor languages, enhancing ASR performance in diverse linguistic settings.

As mentioned earlier, some of the recent works have also studied the role of PEFT in the context of children's ASR task. In [14], four different PEFT techniques, namely LoRA, adapter-tuning, prompt-tuning, and prefix-tuning were explored. The authors observed that the PEFT techniques were more effective when large models were adapted. In [15], the efficacy of LoRA, in combination with int8 quantization [22], was studied for adapting Whisper large-v2 model. In [16], the authors studied and documented the use of Adapters, in combination with synthetic data augmentation, for children's speech recognition. Synthetic data was, in turn, generated using a text-to-speech synthesis module. The authors came up with a two-step training procedure wherein Adapter layers where first trained using synthetically generated data. This was followed by fine-tuning of Adapters as well as the entire model using a mix of synthetic and real data. A comparative study of LoRA and AdaLoRA was done in [17]. In addition to that, a variant of AdaLoRA, referred to as sparsely shared LoRA (S2-LoRA), was also proposed. Compared to AdaLoRA, further reduction in the number of trainable parameters was achieved in the case of S2-LoRA. At the same time, recognition rates obtained using AdaLoRA and S2-LoRA were observed to be similar.

Compared to LoRA and its variants, LoHA, on the other hand, exploits Hadamard products. This, in turn, leads to a more efficient representation of weight updates. Even though LoHA has been explored for various signal processing and federated learning tasks [13], specific applications to speech signal processing were not found. Motivated by the success of LoRA and the absence of a study detailing the effectiveness of LoHA in ASR tasks, we have documented our findings in this paper that demonstrate the efficacy of LoHA especially in the context of children's speech recognition in limited data scenario. Furthermore, as stated earlier, we have also experimentally compared LoHA with LoRA, qLoRA, and AdaLoRA. The experimental evaluations presented in this paper show that LoHA is a more suitable PEFT variant in the context children's ASR task.

III. Basics of low-rank adaptation

Low-rank adaptation is a foundational PEFT method. It addresses the infeasibility of full fine-tuning of large models, such as GPT-3 with 175 billion parameters, by freezing the pre-trained weights and injecting trainable rank decomposition matrices into each layer of the Transformer architecture. This, in turn, reduces the number of trainable parameters by a factor of $10,000$ compared to full fine-tuning while GPU memory requirements are lowered by a factor of 3.

In the case of LoRA, instead of updating all the parameters of a pre-trained model, adaptation is performed by introducing a low-rank update to specific weight matrices, typically in the linear layers. Consider a pre-trained weight matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$ in a linear layer, where d is the output dimension while k is the input dimension. In standard fine-tuning, the output $\mathbf{h}$ for an input vector $\mathbf{x} \in \mathbb{R}^k$ is computed as:

$$\mathbf{h} = \mathbf{W}\mathbf{x}. \tag{1}$$

Equation 1 is modified using LoRA by introducing a low-rank adaptation term. Specifically, two trainable matrices are introduced: $\mathbf{A} \in \mathbb{R}^{r\times k}$ which maps the input into a lower-dimensional (rank-r) subspace, and $\mathbf{B} \in \mathbb{R}^{d\times r}$ which projects it back to the original output space. Here, $r \ll \min(d, k)$ is a small rank hyperparameter introduced to ensure that the adaptation term contains significantly fewer parameters than the full weight matrix. The adaptation term is now given by:

$$\Delta\mathbf{W} = \mathbf{B}\mathbf{A} \tag{2}$$

where, $\Delta\mathbf{W} \in \mathbb{R}^{d\times k}$ is a low-rank matrix with rank at most r. During fine-tuning, the original weight matrix $\mathbf{W}$ remains frozen while the matrices $\mathbf{A}$ and $\mathbf{B}$ are updated. The adapted output $\mathbf{h} \in \mathbb{R}^d$, incorporating low-rank adaptation term, is given by:

$$\mathbf{h} = \mathbf{W}\mathbf{x} + \Delta\mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{x} + \mathbf{B}\mathbf{A}\mathbf{x}. \tag{3}$$

This expression reflects the forward pass of a LoRA-based adapted layer. The term $\mathbf{B}\mathbf{A}\mathbf{x}$ can be interpreted as:

$$\mathbf{B}(\mathbf{A}\mathbf{x}),$$

where, $\mathbf{A}\mathbf{x} \in \mathbb{R}^r$ is the projection of the input onto a lower-dimensional subspace, and $\mathbf{B}(\mathbf{A}\mathbf{x}) \in \mathbb{R}^d$ projects it back to align with the dimensionality of $\mathbf{W}\mathbf{x}$. Thus, the final expression for the LoRA-adapted output is:

$$\mathbf{h} = \mathbf{W}\mathbf{x} + \mathbf{B}\mathbf{A}\mathbf{x}. \tag{4}$$

In the context of qLoRA, the matrices $\mathbf{A}$ and $\mathbf{B}$ are quantized as: $\mathbf{A} \approx \text{scale}_{\mathbf{A}} \cdot \mathbf{A}_q$ and $\mathbf{B} \approx \text{scale}_{\mathbf{B}} \cdot \mathbf{B}_q$, respectively. Here, $\mathbf{A}_q \in \mathbb{Z}^{d\times r}$ and $\mathbf{B}_q \in \mathbb{Z}^{r\times k}$ are the quantized versions of $\mathbf{A}$ and $\mathbf{B}$, respectively, while $\text{scale}_{\mathbf{A}}, \text{scale}_{\mathbf{B}} \in \mathbb{R}$ are scalar values used to de-quantize. The quantized low-rank adaptation term then becomes:

$$\Delta\mathbf{W} \approx \text{scale}_{\mathbf{A}} \cdot \text{scale}_{\mathbf{B}} \cdot (\mathbf{A}_q\mathbf{B}_q). \tag{5}$$

In traditional PEFT methods, the budget is often evenly distributed across weight matrices, overlooking varying importance. In AdaLoRA, on the other hand, this shortcoming is

addressed by allocating the budget based on importance scores. Thus, incremental updates are parameterized via singular value decomposition (SVD). This allows effective pruning of less important updates. The low-rank adaptation term $\Delta\mathbf{W}$, in the case of AdaLoRA, is given by:

$$\Delta\mathbf{W} = \mathbf{P} \cdot \mathbf{\Sigma} \cdot \mathbf{Q} \tag{6}$$

where: $\mathbf{P} \in \mathbb{R}^{d\times r}$ is an orthogonal matrix projecting to a lower-dimensional space, $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$ is a diagonal matrix of learnable singular values σ_i, and $\mathbf{Q} \in \mathbb{R}^{r\times k}$ is an orthogonal matrix projecting back to the input dimension. Therefore, the adapted weight matrix in the context of AdaLoRA is expressed as:

$$\mathbf{W}' = \mathbf{W} + \mathbf{P} \cdot \mathbf{\Sigma} \cdot \mathbf{Q}. \tag{7}$$

IV. Low-rank Hadamard product adaptation

LoHA is a PEFT variant that uses element-wise Hadamard products for weight updates. LoHA breaks down the weight updates $\Delta\mathbf{W}$ into four smaller matrices, enhancing model expressiveness while keeping the increase in trainable parameters to a minimal value. Initially developed for computer vision tasks, particularly for generating diverse images, LoHA aims to expand its application to other model types, though it currently excludes embedding layers in PEFT setups. This approach offers a novel perspective on matrix operations, potentially increasing efficiency for specific tasks.

In LoHA, the standard LoRA formulation is modified by using the Hadamard (element-wise) product to combine two low-rank matrices. The adaptation term $\Delta\mathbf{W}$ is defined as:

$$\Delta\mathbf{W} = (\mathbf{A} \circ \mathbf{B}) \tag{8}$$

where, $\circ$ denotes the Hadamard product. The matrices $\mathbf{A}$ and $\mathbf{B}$ are parameterized in a low-rank fashion given as follows:

$$\mathbf{A} = \mathbf{U}_A\mathbf{V}_A^\top \tag{9}$$

$$\mathbf{B} = \mathbf{U}_B\mathbf{V}_B^\top \tag{10}$$

where, $\mathbf{U}_A, \mathbf{U}_B \in \mathbb{R}^{d\times r}, \mathbf{V}_A, \mathbf{V}_B \in \mathbb{R}^{k\times r}$, and $r \ll \min(d, k)$ is the rank for the low-rank parameterization. Thus, the full adaptation term becomes:

$$\Delta\mathbf{W} = (\mathbf{U}_A\mathbf{V}_A^\top) \circ (\mathbf{U}_B\mathbf{V}_B^\top). \tag{11}$$

The adapted weight matrix in the case of LoHA is given by:

$$\mathbf{W}' = \mathbf{W} + (\mathbf{U}_A\mathbf{V}_A^\top) \circ (\mathbf{U}_B\mathbf{V}_B^\top). \tag{12}$$

This formulation enables richer expressivity compared to standard LoRA, while still maintaining parameter efficiency through low-rank structures.

V. Experimental evaluations

In this section, we present the experimental evaluations performed to study the efficacy of LoHA and some of the explored LoRA variants in the context of children's ASR task especially in resource constrained settings.

A. Database employed

For experimental assessments, we have chosen CMU Kids corpus [23] for adaptation as well as evaluation purposes. CMU Kids corpus primarily consists of American English read speech. A subset of the CMU Kids database, used in our work for adaptation, consisted of 6.45 hours of speech data, representing itself as a case-study for extremely limited data conditions. Experimental validations were performed using a test set derived from the CMU Kids corpus. The test set consisted of 2.66 hours of data from 75 children. The adaptation data comprised of 3641 utterances while the test set consisted of 1539 utterances.

B. Experimental set-up

We have performed the experimental evaluations on three variants of the Whisper (English) model: **Base**, **Small** and **Medium**. These models allowed us to achieve satisfactory word error rate (WER) when transcribing children's speech, despite our constrained GPU resources. Whisper is based on an encoder-decoder Transformer architecture. To fine-tune the model, we employed the PEFT module from Hugging Face, which provides the necessary classes and functions. The PEFT configuration was consistent across all three Whisper variants, ensuring uniformity during the fine-tuning process. While applying LoRA, the rank of the update matrix was set to 32 while a scaling factor of 64 was chosen. The LoRA layers were applied to the query and value projection layers. A dropout rate of 0.05 was used. Bias parameters were not retrained during fine-tuning. Query and value projections were adapted in the case qLoRA as well as AdaLoRA. For LoHA, on the other hand, target modules included the query and value projection layers, the key and output projection layers, as well as two fully connected layers (fc1 and fc2). The sequence-to-sequence model was trained using the Transformer library with a batch size of 8, a gradient accumulation step of 1, and a learning rate of 0.001. The warm-up step was set to 10, and fine-tuning was conducted over 20 epochs. The recognition performance was evaluated using WER as the primary metric.

C. Experimental results and analysis

The WERs obtained by adapting the three different Whisper models using the explored LoRA variants are enlisted in Table I. It is worth highlighting here that, the unadapted Whisper Medium model yields a WER of 12.7% with respect to the test set used in this paper (see Table II). This implies that there is an absolute reduction in WER by **9.7%** on adapting Whisper Medium model. Among the three explored variants, LoRA consistently results in lower WERs. At the same time, WERs obtained using AdaLoRA are poorer than those obtained using LoRA and qLoRA. This may be due the fact that, the total number of trainable parameters are much lower in the case of AdaLoRA (0.46% of the total parameters when Whisper Medium model is adapted). From these results, we can conclude that the use of LoRA variants is an extremely effective tool to develop robust ASR systems for children in limited data set-up.

TABLE I
WERs OBTAINED BY ADAPTING THE DIFFERENT WHISPER MODELS USING LORA, QLORA, ADALORA, AND LOHA.

Whisper model	Adaptation method	WER in %	Trainable params. in %
Base	LoRA	5.2	1.59
	qLoRA	5.0	1.59
	AdaLoRA	8.6	0.60
	LoHA	**4.2**	10.67
Small	LoRA	3.9	1.44
	qLoRA	4.3	1.44
	AdaLoRA	4.9	0.54
	LoHA	**3.4**	9.69
Medium	LoRA	3.7	1.22
	qLoRA	3.8	1.22
	AdaLoRA	4.2	0.46
	LoHA	**3.0**	8.30

It can be also observed that LoHA consistently delivers the lowest error rates across all model sizes, achieving WER of 4.2%, 3.4%, and 3.0% for Base, Small and Medium models, respectively. However, this gain in WER comes at a increased number of trainable parameters when compared to LoRA. The relative reduction in WERs on applying LoHA are shown in Figure 1. The bar charts show that there are significant relative reductions in WERs. Furthermore, on comparing the best case result obtained using LoHA with that obtained using unadapted Whisper Medium model, a relative reduction of **76.4%** in WER is obtained. Therefore, it is evident that LoHA is more suitable for children's ASR task.

In order to gain better insight, we next performed studies to determine the recognition rates on age as well as gender-dependent split of the test data. For that purpose, the test set was split into two subsets: **Age-6–7**, containing speech from children aged between 6 to 7 years, and **Age-**8**–**9, consisting of speech from children aged between 8 to 9 years. In addition to that, the test data was also divided into male and female speaker subsets to analyze gender-wise performance. The WERs for this study are given in Table II. Since the training data and model architecture remain unchanged, the total number of parameters, trainable parameters, and trainable parameter percentages are identical to the main experiment. We have also included the WERs for the case when unadapted Whisper Medium model is used for better contrast. It is to note here that, both LoRA as well as LoHA were applied to only query and value projections unlike the earlier presented experimental results. This was done to ensure a better comparison between LoRA and LoHA. As evident from the tabulated results, large reductions in WER are obtained on adapting the Whisper model using either LoRA or LoHA.

The recognition performances obtained by applying low-rank adaptation techniques can be overly good if there is a

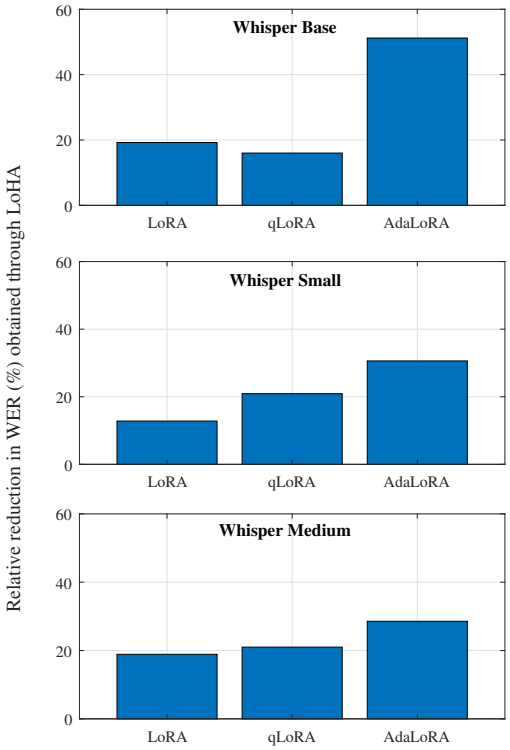

Fig. 1. Bar plots showing the relative reductions in WERs obtained on applying LoHA compared to the cases when LoRA, qLoRA and AdaLoRA are applied.

TABLE II
AGE AND GENDER-WISE WERS OBTAINED ON APPLYING LORA AND LOHA ON THE WHISPER MEDIUM MODEL. LORA/LOHA ARE APPLIED TO THE QUERY AND VALUE PROJECTIONS ONLY.

Test subset	WER (%)		
	Unadapted	LoRA	LoHA
Full test set	12.7	3.7	3.5
Age 6–7	16.7	4.4	4.5
Age 8–9	11.0	3.5	3.2
Male	14.1	4.2	3.6
Female	11.9	3.6	4.2

TABLE III
WERS OBTAINED ON APPLYING LORA AND LOHA ON THE WHISPER MEDIUM MODEL USING ADAPTATION AND TEST DATA DERIVED FROM THE CMU KIDS CORPUS WITH NO SPEAKER OVERLAP. LORA AND LOHA WERE APPLIED TO VALUE AND QUERY PROJECTIONS ONLY.

Adaptation data	Test set	WER (%)	
		LoRA	LoHA
Age 8–9	Age 6–7	13.2	11.8
Age 6–7	Age 8–9	8.5	7.9
Male	Female	8.8	8.4
Female	Male	6.6	5.9

speaker overlap across the adaptation and test data. To rule out this aspect, we used non-overlapping adaptation and test sets and re-evaluated the efficacy of LoRA and LoHA. The WERs for that study are given in Table III. On comparing the WERs enlisted in Table III with those given in Table II, it is clearly evident that, using non-overlapping adaptation and test sets leads to much higher WERs. At the same time, LoHA tends to perform much better than LoRA under such setups. This set of results reveal the potential of LoHA over LoRA despite both being applied on same layers of the Transformer model. In other words, LoHA is less affected by a mismatch in speaker characteristics when compared to LoRA. This makes LoHA a better option for adapting Whisper models in order to develop robust children's ASR systems.

VI. Conclusion

In this paper, we have documented our efforts towards developing efficient ASR systems for children by adapting pre-trained Whisper models. For that purpose, we have explored the efficacy of three variants of LoRA. In addition to those, we have also studied and documented the effectiveness of LoHA. To the best of our knowledge, it is the first study on children's speech recognition that exploits LoHA for adapting Whisper models. From the experimental evaluations presented in this paper, we can conclude that adapting Whisper models does help in reducing error rates with respect to children's speech by a large margin despite using only 6.45 hours of adaptation data. At the same time, LoHA yields significantly lower WERs when compared to those obtained using LoRA, qLoRA and AdaLoRA. Furthermore, LoHA is also observed to be more robust towards inter-speaker acoustic variations when compared to LoRA.

References

[1] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech recognition: A study of individual differences with implications for system development," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014.

[2] P. G. Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech & Language*, vol. 72, p. 101289, 2022.

[3] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," *Proceedings of ICML workshop on unsupervised and transfer learning*, 2012.

[4] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *Journal of Machine Learning Research*, 2010.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end asr: from supervised to semi-supervised learning with modern architectures," in *Workshop on Speech and Language Processing for Assistive Technologies*, 2019.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.

[8] R. Lu, M. Shahin, and B. Ahmed, "Improving children's speech recognition by fine-tuning self-supervised adult speech representations," *arXiv preprint arXiv:2211.07769*, 2022.

[9] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of whisper models to child speech recognition," in *Proc. INTERSPEECH*, 2023.

[10] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large Language models," in *Proc. ICLR*, 2022.

[11] Y. Zhang, Y. Qin, Z. Li, F. Li, Z. Liu, and Y. Hu, "Adalora: Adaptive budget allocation for parameter-efficient fine-tuning," *arXiv preprint arXiv:2303.10512*, 2023.

[12] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," *arXiv preprint arXiv:2305.14314*, 2023.

[13] N. Hyeon-Woo, M. Ye-Bin, and T. Oh, "FedPara: Low-rank Hadamard Product for Communication-Efficient Federated Learning," in *Proc. ICLR*, 2022.

[14] R. Fan, N. B. Shankar, and A. Alwan, "Benchmarking children's ASR with supervised and self-supervised speech foundation models," *arXiv preprint arXiv:2406.10507*, 2024.

[15] R. Southwell, W. Ward, V. A. Trinh, C. Clevenger, C. Clevenger, E. Watts, J. Reitman, S. D'Mello, and J. Whitehill, "Automatic speech recognition tuned for child speech in the classroom," in *Proc. ICASSP*, 2024, pp. 12 291–12 295.

[16] T. Rolland and A. Abad, "Improved children's automatic speech recognition combining adapters and synthetic data augmentation," in *Proc. ICASSP*, 2024, pp. 12 757–12 761.

[17] W. Liu, Y. Qin, Z. Peng, and T. Lee, "Sparsely shared lora on whisper for child speech recognition," in *proc. ICASSP*, 2024, pp. 11 751–11 755.

[18] Z. Song, J. Zhuo, Y. Yang, Z. Ma, S. Zhang, and X. Chen, "Lora-whisper: Parameter-efficient and extensible multilingual asr," *arXiv preprint arXiv:2406.06619*, 2024.

[19] G. Joseph and A. Baby, "Speaker personalization for automatic speech recognition using weight-decomposed low-rank adaptation," in *Proc. INTERSPEECH*, 2024, pp. 2875–2879.

[20] R. Jairam, G. Jyothish, and B. Premjith, "A few-shot multi-accented speech classification for indian languages using transformers and llm's fine-tuning approaches," in *Proc. SVLTDL*, 2024, pp. 1–9.

[21] T. Bornheim, N. Grieger, P. G. Blaneck, and S. Bialonski, "Speaker attribution in german parliamentary debates with qlora-adapted large language models," *arXiv preprint arXiv:2309.09902*, 2023.

[22] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "Llm.int8(): 8-bit matrix multiplication for transformers at scale," 2022.

[23] M. Eskenazi, J. Mostow, and D. Graff, "The CMU Kids Corpus LDC97S63," https://catalog.ldc.upenn.edu/LDC97S63, 1997.