Mixture of LoRA Experts for Low-Resourced Multi-Accent Automatic Speech Recognition

Raphaël Bagat[1], *Irina Illina*[1], *Emmanuel Vincent*[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

raphael.bagat@loria.fr, irina.illina@loria.fr, emmanuel.vincent@inria.fr

Abstract

We aim to improve the robustness of Automatic Speech Recognition (ASR) systems against non-native speech, particularly in low-resourced multi-accent settings. We introduce Mixture of Accent-Specific LoRAs (MAS-LoRA), a fine-tuning method that leverages a mixture of Low-Rank Adaptation (LoRA) experts, each specialized in a specific accent. This method can be used when the accent is known or unknown at inference time, without the need to fine-tune the model again. Our experiments, conducted using Whisper on the L2-ARCTIC corpus, demonstrate significant improvements in Word Error Rate compared to regular LoRA and full fine-tuning when the accent is unknown. When the accent is known, the results further improve. Furthermore, MAS-LoRA shows less catastrophic forgetting than the other fine-tuning methods. To the best of our knowledge, this is the first use of a mixture of LoRA experts for non-native multi-accent ASR.

Index Terms: multi-accent automatic speech recognition, Whisper, LoRA, low-resourced, non-native speech

1. Introduction

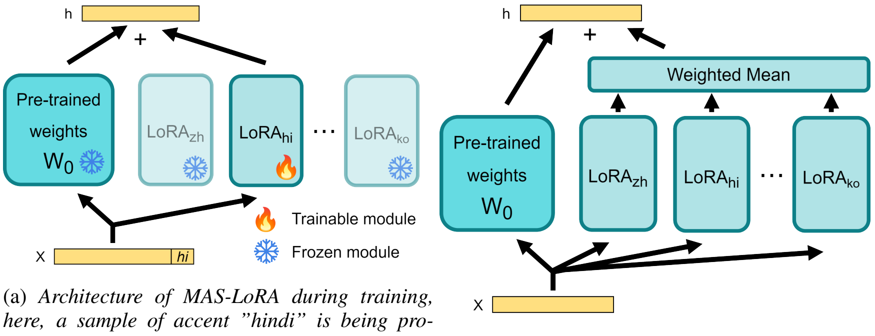
Automatic Speech Recognition (ASR) systems have reached human-like performance in many domains [1]. End-to-end systems such as Whisper [2], a multilingual ASR model, work very well when the speakers talk in their native language. However, their performance drops on non-native, accented speech. Indeed, non-native speech often involves specific pronunciations of certain phonemes borrowed from the speaker's mother tongue (L1) [3], which induce ASR errors. Non-native accent can also affect the prosody of the utterance to resemble the speaker's L1, leading to an even greater mismatch with native speech [4]. In the context of *multi-accent* ASR, when the systems are used to transcribe utterances from different accents, these phenomena are exacerbated by the larger number of accents. When facing accented speech, ASR systems can either be *accent-agnostic*, i.e. have no information about the speaker's accent, or on the contrary, be *accent-aware*. For a system to be used in an accent-agnostic setting, transcribed training data that cover a wide variety of accents are needed. Such data are rare, thus ASR systems must be trained on low-resourced data which makes the problem even more challenging. Improvements in non-native multi-accent ASR would make these systems usable in contexts where people have to speak a different language than their mother tongue, e.g., in Air Traffic Communications where pilots from all over the world have to speak English, or in international commerce.

Initial approaches explored the adaptation of Gaussian mixture model - hidden Markov model (GMM-HMM) based acoustic models for accented ASR in both accent-aware and accent-agnostic settings [5, 6]. More recently, deep learning based models have been studied to improve accented ASR. Especially in the case of multi-accent ASR, prior works proposed to improve ASR by using accent recognition in a multi-task setting to learn accent specific features along the ASR training [7, 8]. Methods based on adding one-hot representations of dialects to the model's input also showed promising improvements [9]. However, these methods considered native accents only.

To bridge the gap with non-native accents, [10] used various transfer learning methods to improve non-native multi-accent ASR, exhibiting the importance of a multilingual model to handle pronunciation differences across accents. This method is based on full fine-tuning, which is computationally expensive. Parameter-efficient fine-tuning methods have emerged, starting with Adapters [11] which consist of training small neural modules inserted in between a model's pre-existing layers while keeping these layers frozen. [12] used Adapters to fine-tune Whisper with different native English accents, leading to similar results to full fine-tuning and even improvements for the African-American accent. [13] used multiple (non-linear) Adapters to improve non-native multi-accent ASR, but this method relies on an external accent identification model. Following Adapters, *Low-Rank Adaptation* (LoRA) [14] and its many variants [15, 16, 17] have been proposed to further improve parameter-efficient fine-tuning. LoRA has been used to improve ASR systems on specific languages [18]. In order to use LoRA on data coming from different domains, many methods proposed to jointly use multiple LoRAs as a mixture of experts (MoE) [19, 20, 21] and [22] use them to improve Whisper's multilingual ASR. To the best of our knowledge, mixture of LoRA expert methods have not yet been used for non-native multi-accent ASR, which remains an understudied problem due in particular to its low-resourced nature.

In this paper, we leverage Whisper's multilingual knowledge via a mixture of LoRA experts to improve non-native multi-accent ASR. Each expert specializes in a single accent and their combined knowledge is used at inference time, either with equal weights for all accents in an accent-agnostic setting or with a fixed, higher weight for the target accent and a lower equal weight for the remaining accents in an accent-aware setting. We show that both approaches decrease the WER on the L2-ARCTIC corpus compared to using a single LoRA corresponding to the ground truth accent, which validates the MoE approach. Thanks to the linear nature of LoRA, the weights of the LoRA experts can be merged with those of the original model, leading to non-native multi-accent ASR at no extra computational cost.

This paper is organized as follows. Section 2 introduces the proposed method. Section 3 presents our experiments. Section 4 describes our results. We conclude in Section 5.



(a) *Architecture of MAS-LoRA during training, here, a sample of accent "hindi" is being processed.*

(b) *Architecture of MAS-LoRA during inference.*

Figure 1: *Architecture of MAS-LoRA.*

2. Proposed methodology

2.1. Classical LoRA

LoRA aims to approximate weight updates $\Delta W \in \mathbb{R}^{d\times k}$ of the frozen pre-trained weights $W_0 \in \mathbb{R}^{d\times k}$ during fine-tuning by the product of two low-rank matrices $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$ with rank $r \ll \min(d,k)$ scaled by a factor α:

$$\begin{aligned} W &= W_0 + \alpha\, \Delta W \\ &= W_0 + \alpha\, BA. \end{aligned} \tag{1}$$

While LoRA can be applied to any types of pre-trained weights, it is extensively used in attention modules by applying it to some or all of the attention matrices. By construction, its weights can be merged with the pre-trained weights and do not increase the computational cost at inference time.

2.2. Mixture of Accent-Specific LoRAs (MAS-LoRA)

We propose MAS-LoRA, a MoE method using LoRA experts trained on single-accent data and combined at inference time to process multi-accent data. In detail, if the training data contains n accents, we instantiate n LoRA experts, one per accent, which were used at the same time as W_0. This way, each expert is adapted to a specific accent, making it easier to learn each accent's unique characteristics.

2.2.1. *Accent-specific fine-tuning*

During fine-tuning, samples with a given accent will only pass through the pre-trained weights and the expert specialized in this accent (see Fig. 1a). As with LoRA, the experts use low-rank-parametrized update matrices. For a sample of hidden representation x and accent j, the output h of one MAS-LoRA layer during training is

$$\begin{aligned} h &= \text{MAS-LoRA}_j(x) \\ &= (W_0 + \alpha\, \Delta W_j)\, x \\ &= (W_0 + \alpha\, B_j A_j)\, x. \end{aligned} \tag{2}$$

This allows each expert to be trained separately. Similarly to other parameter-efficient fine-tuning methods, the pre-trained weights W_0 remain frozen throughout the entire fine-tuning. As shown in [13], in an encoder-decoder architecture, accent-related adaptation for the encoder leads to a systematic improvement. Thus, we chose to always use MAS-LoRA to fine-tune the encoder. Its use in the decoder is discussed in Section 4.2.

As opposed to regular MoEs, we do not learn routers, as we shall see in Section 4 that assigning all weight to the expert corresponding to the ground truth accent, what we could make routers learn, is suboptimal w.r.t. sharing weight with other experts. We have tried learning routers, but it did not show any improvement.

2.2.2. *Accent-agnostic inference*

At inference (see Fig. 1b), when the accent of the sample is unknown, we average the outputs of all experts W_i with equal $\frac{1}{n}$ weights before adding them to the output of the pre-trained weights W_0:

$$\begin{aligned} W &= W_0 + \frac{1}{n}\sum_{i=1}^{n} \alpha\, W_i \\ &= W_0 + \frac{1}{n}\sum_{i=1}^{n} \alpha\, B_i A_i. \end{aligned} \tag{3}$$

This mixture method allows us to merge experts with pre-trained weights, preserving the original inference cost.

2.2.3. *Accent-aware inference*

When the accent label is available at inference time, instead of using an equal weight of $\frac{1}{n}$ for every accent, it is possible to give a higher weight to the expert corresponding to that accent. We parameterize that weight as $\frac{1}{\beta}$, with $\beta \in [1, n]$. The residual $1 - \frac{1}{\beta}$ weight is shared equally among all other accents. Denoting as j the accent label of the current sample, the experts are used as follows:

$$W = W_0 + \frac{1}{\beta}\alpha\, B_j A_j + \frac{1-\frac{1}{\beta}}{n-1}\sum_{\substack{i=1\\ i\neq j}}^{n} \alpha\, B_i A_i. \tag{4}$$

We do not merge experts, causing only a slight increase in inference cost.

3. Experimental settings

3.1. Datasets

Our experiments are conducted on the L2-ARCTIC dataset [23]. This dataset contains speech utterances in English spoken by non-native speakers with different accents. The accents (L1) are the following: *Arabic*, *Chinese*, *Hindi*, *Korean*, *Spanish*, and *Vietnamese*. Each accent class has 4 different speakers, thus totaling 24 speakers, with 1 h of data per speaker. Every speaker reads the same phonetically-balanced sentences originating from Project Gutenberg [24].

Table 1: *WER (%) obtained with different fine-tuning methods on L2-ARCTIC and LibriSpeech test-clean. The Encoder and Decoder columns indicate the fine-tuning method used in the encoder and the decoder. The percentage of trained parameters is with respect to the total model size. Bold numbers indicate the best result for each corpus and those results which are statistically equivalent to it.*

	Encoder	**Decoder**	**Trained params. (%)**	**WER L2-ARCTIC (%)**	**WER LibriSpeech (%)**
1	No FT	No FT	0	13.77	**5.78**
2	Full FT	Full FT	100	12.21	7.90
3	LoRA-qv	LoRA-qv	0.73	12.32	6.32
4	MAS-LoRA-qv	No FT	1.44	14.08	**5.94**
5		LoRA-qv	1.91	**11.77**	**5.81**
6		MAS-LoRA-qv	4.21	**11.78**	**5.91**
7	LoRA-qkvo	LoRA-qkvo	1.44	13.48	7.16
8	MAS-LoRA-qkvo	No FT	2.84	12.14	**5.95**
9		LoRA-qkvo	3.76	**11.77**	**5.95**
10		MAS-LoRA-qkvo	8.07	**11.90**	6.27

To avoid evaluation biases, it is important that the sentences and speakers in the test set are disjoint from those in the training and validation sets. Ideally, the sentences in the training and validation sets should also be disjoint. Due to the small amount of data, we run 8-fold cross-validation. For a given fold and accent, the training set contains 80% of the unique sentences spoken by 3 speakers, the validation set contains 10% other sentences spoken by the same 3 speakers, and the test set contains the 10% remaining sentences spoken by the remaining speaker. Thus, each speaker is part of two test folds. Each method is fine-tuned and tested using the same 8 folds. Table 2 shows the split for a single accent across all folds.

Table 2: *Quantity of audio and words per accent across all folds. Each accent follows the same split.*

	Training	**Valid.**	**Test**
Audio duration	8x 2 h 48 min	8x 18 min	8x 6 min
# of words	8x 144,154	8x 17,028	8x 6,027

To evaluate the effect of non-native multi-accent fine-tuning on native English speech, we also use the *test-clean* subset of LibriSpeech [25], a well-known ASR corpus made of recordings of native English speakers who read books, for testing purposes only. This subset contains 5 h 48 min of audio data.

3.2. General parameters

Our experiments were carried out using the Whisper small model [2], which has encoder-decoder architecture and is of reasonable size (244M parameters) to run on many types of devices, such as on-board devices, and has proven to be an already highly capable ASR model. It can be found on *Hugging Face*[1]. The fact that this model was trained on multilingual data is an important feature as multilingual features have proven to be useful to improve accented ASR [10, 26]. In order to match Whisper's expected input, all audio files have been resampled from 44.1 kHz to 16 kHz. Models are trained for 3 epochs, with a batch size of 16. Parameter-efficient fine-tunings were made by applying LoRA and MAS-LoRA to attention modules in the Query and Value matrices (LoRA-qv and MAS-LoRA-qv) or the Query, Value, Key and Output matrices (LoRA-qkvo and MAS-LoRA-qkvo), where r was set to 16 and α to 1. These settings were chosen because they have proven to be effective [14, 21]. The learning rate is set to start at 1e-5 for full fine-tuning and 5e-5 for parameter-efficient fine-tuning methods, and decreases linearly to its half throughout the fine-tuning. Fine-tunings have been conducted on NVIDIA A100 GPUs and tests on NVIDIA V100 GPUs. For decoding, greedy search is used for computational reasons. Our code is publicly available[2].

3.3. Evaluation metric

The results are reported in terms of the Word Error Rate (WER). Early stopping is made using the WER on the validation set. The statistical significance of the results has been validated using the Matched Pair Sentence Segment test with SCTK [27].

4. Results and discussions

4.1. Baselines

We consider three baselines: pre-trained model without fine-tuning (referred to as No FT), full model fine-tuning (Full FT) and parameter-efficient fine-tuning using LoRA-qv or LoRA-qkvo. The obtained WERs can be found in rows 1, 2, 3 and 7 of Table 1, respectively. It can be seen that full fine-tuning improves the performance compared to No FT. LoRA applied to the Q, K, V, O matrices shows performance equivalent to No FT, while LoRA-qv shows significant improvements in performance compared to No FT, getting a WER of 12.32%.

4.2. Accent-agnostic MAS-LoRA

Impact of MAS-LoRA in the encoder — Accent-agnostic MAS-LoRA was studied under 3 conditions. As previously stated, it is always applied to the encoder. For the decoder, we either used no fine-tuning or applied LoRA or MAS-LoRA. This allows us to see the effect of accent-related fine-tuning on the decoder which, we believe, should contain less accent-related features. Results in Table 1 (rows 4-6, 8-10) show that MAS-LoRA-qkvo significantly outperforms LoRA when applied to the encoder with LoRA-qkvo in the decoder, achieving

[1] `https://huggingface.co/openai/whisper-small`

[2] `https://gitlab.inria.fr/rbagat/mas-lora`

Table 3: *Zero-shot WER (%) on test accents unseen during training. AR, ZH, HI, KR, SP and VI mean Arabic, Chinese, Hindi, Korean, Spanish, and Vietnamese accents, respectively. Bold numbers indicate the best result for each accent and those results which are statistically equivalent to it.*

Encoder	**Decoder**	**AR**	**ZH**	**HI**	**KR**	**SP**	**VI**	**Mean**
No FT	No FT	13.18	16.04	**7.64**	10.40	13.72	22.01	13.77
Full FT	Full FT	15.50	20.89	11.06	14.72	17.11	23.42	17.12
LoRA-qkvo	LoRA-qkvo	**11.44**	**15.70**	**7.36**	**9.54**	**12.46**	**19.80**	**12.72**
MAS-LoRA-qkvo	LoRA-qkvo	**11.43**	**14.96**	**7.19**	**8.65**	**12.64**	**20.41**	**12.55**

a WER of 11.77% compared to 13.48% for LoRA-qkvo alone. For the Q, V matrices, LoRA-qv applied to the encoder and decoder yields a WER of 12.32% and MAS-LoRA-qv paired with LoRA-qv in the decoder 11.77%. MAS-LoRA also significantly outperforms full fine-tuning, when applied to the encoder with LoRA in the decoder on both sets of matrices (11.77% versus 12.21%).

Impact of MAS-LoRA in the decoder — When MAS-LoRA is used in the encoder, the results obtained by applying LoRA or MAS-LoRA to the decoder (rows 5-6, 9-10) are similar to each other, with WERs of 11.77% and 11.90%, respectively. This indicates that accent-related fine-tuning isn't necessarily the best choice for the decoder. Instead, using an accent-independent method, here LoRA, is as effective. Though, it is important to note that when MAS-LoRA is used in the encoder, the decoder has to be fine-tuned. Not fine-tuning the decoder degrades the results, especially when MAS-LoRA-qv is used in the encoder.

Performance on native speech — After fine-tuning the models on non-native speech, we tested them on native speech to evaluate the extent of performance degradation. The results are shown in the last column of Table 1. It can be seen that, compared to LoRA and Full FT, MAS-LoRA yields results that are equivalent to those of the model before fine-tuning (rows 4-6, 8-9), except when MAS-LoRA-qkvo is both applied to the encoder and the decoder. This shows that MAS-LoRA is less prone to catastrophic forgetting unlike full fine-tuning and LoRA. In the following sections, when MAS-LoRA is applied, we therefore use MAS-LoRA in the encoder and LoRA in the decoder.

Performance on unseen accents — To assess the robustness of the method against new accents, we have conducted a zero-shot experiment by removing one accent from the training set and testing on that accent. This was conducted for Full FT, LoRA-qkvo and MAS-LoRA-qkvo. According to Table 3, Full FT shows performance degradation on unseen accents compared to No FT. On the other end, MAS-LoRA remains as robust as LoRA in front of new accents. Moreover, except in the case of the Hindi accent, both MAS-LoRA and LoRA achieve significantly improved results compared to No FT, highlighting that multi-accent fine-tuning is important even if the training data does not cover test accents.

4.3. Accent-aware MAS-LoRA

Using only the specialized expert — Accent-aware inference indicates that the accent label is known at inference. One could then think that instead of using all the experts at inference, it would be better to use only the expert specialized in the sample's accent. The results can be found in Fig. 2, where $\beta = 6$ indicates that all experts get equal weights and $\beta = 1$ that only the expert specialized in the sample's accent is used. It can be seen that using only 1 expert degrades the results compared to using all the experts. This demonstrates the importance of combining knowledge from all experts, and can be interpreted as a form of regularization.

Effect of β — The results obtained using accent-aware inference are shown in Fig. 2. The values of $\beta = 1$ or 6 were already discussed in the previous paragraph. For both MAS-LoRA-qv and MAS-LoRA-qkvo using $\beta = 5$ already gives significantly better results than $\beta = 6$, and performance keeps getting better as β decreases until $\beta = 2$. Though, as it can be seen, decreasing β further has a negative effect on the WER. This means that each expert has to contribute enough for MAS-LoRA to be effective.

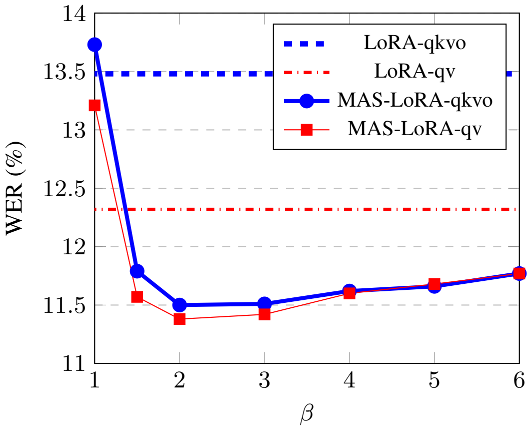


Figure 2: *Effect of β on the WER when the accent label is known at inference. LoRA is applied to both encoder and decoder and MAS-LoRA is applied to the encoder with LoRA in the decoder.*

5. Conclusion

In this article, we focused on the task of improving ASR when facing multiple non-native accents. We introduced Mixture of Accent-Specific LoRAs, a fine-tuning method based on a mixture of LoRA experts. Each expert specializes in a specific accent, and their combined knowledge is used at inference. We showed that, when the accent is unknown at inference, MAS-LoRA significantly improves the WER compared to full fine-tuning and regular LoRA, provided that it is used in the encoder at least. MAS-LoRA also shows a similar generalization capability as LoRA when facing new accents and avoids catastrophic forgetting issues. Moreover, when the accent is known at inference, MAS-LoRA obtains further improved results.

6. Acknowledgments

This work was funded by the DeepMAUVES project supported by DGA and CNRS, and granted access to the HPC resources of IDRIS under the allocation 2024-AD011015024 made by GENCI.

7. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, 2017.

[2] 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*, 2023, pp. 28 492–28 518.

[3] M. Zampini, "L2 speech production research," *Phonology and Second Language Acquisition*, vol. 36, pp. 219–249, 2008.

[4] M. G. Busà, "Effects of l1 on l2 pronunciation: Italian prosody in English," *EIL, ELF, Global English: Teaching and Learning Processes*, pp. 207–228, 2010.

[5] D. Vergyri, L. Lamel, and J.-L. Gauvain, "Automatic speech recognition of multiple accented English data," in *Interspeech*, 2010, pp. 1652–1655.

[6] H. Kamper and T. Niesler, "Multi-accent speech recognition of Afrikaans, Black and White varieties of South African English," in *Interspeech*, 2011, pp. 3189–3192.

[7] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning." in *Interspeech*, 2018, pp. 2454–2458.

[8] T. Viglino, P. Motlicek, and M. Cernak, "End-to-end accented speech recognition." in *Interspeech*, 2019, pp. 2140–2144.

[9] B. Li, T. N. Sainath, K. C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, and K. Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4749–4753.

[10] M. Matassoni, R. Gretter, D. Falavigna, and D. Giuliani, "Non-native children speech recognition through transfer learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6229–6233.

[11] 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 *36th International Conference on Machine Learning*, 2019, pp. 2790–2799.

[12] Z. Huang, H. Xing, and M. Liu, "Adapter integration: Mitigating catastrophic forgetting in multi-language and multi-accent Whisper ASR model fine-tuning," https://www.researchgate.net/publication/374867801_Adapter_Integration_Mitigating_Catastrophic_Forgetting_in_Multi-Language_and_Multi-Accent_Whisper_ASR_Model_Fine-tuning, 2023.

[13] Y. Qian, X. Gong, and H. Huang, "Layer-wise fast adaptation for end-to-end multi-accent speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2842–2853, 2022.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[15] Q. Zhang, M. Chen, A. Bukharin, N. Karampatziakis, P. He, Y. Cheng, W. Chen, and T. Zhao, "AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning," in *International Conference on Learning Representations*, 2023.

[16] S. Hayou, N. Ghosh, and B. Yu, "LoRa+: Efficient low rank adaptation of large models," in *International Conference on Machine Learning*, 2024, pp. 17 783–17 806.

[17] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "DoRA: Weight-decomposed low-rank adaptation," in *International Conference on Machine Learning*, 2024, pp. 32 100–32 121.

[18] Y. Li, Y. Wang, L. M. Hoi, D. Yang, and S.-K. Im, "A review on speech recognition approaches and challenges for Portuguese: exploring the feasibility of fine-tuning large-scale end-to-end models," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2025, no. 1, p. 3, 2025.

[19] T. Luo, J. Lei, F. Lei, W. Liu, S. He, J. Zhao, and K. Liu, "MoELoRa: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models," *arXiv preprint arXiv:2402.12851*, 2024.

[20] C. Gao, K. Chen, J. Rao, B. Sun, R. Liu, D. Peng, Y. Zhang, X. Guo, J. Yang, and V. Subrahmanian, "Higher layers need more LoRa experts," *arXiv preprint arXiv:2402.08562*, 2024.

[21] D. Li, Y. Ma, N. Wang, Z. Cheng, L. Duan, J. Zuo, C. Yang, and M. Tang, "MixLoRa: Enhancing large language models fine-tuning with LoRa based mixture of experts," *arXiv preprint arXiv:2404.15159*, 2024.

[22] 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.

[23] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A non-native English speech corpus," in *Interspeech*, 2018, pp. 2783–2787.

[24] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *5th ISCA workshop on speech synthesis*, 2004, pp. 223–224.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[26] N. T. Vu, Y. Wang, M. Klose, Z. Mihaylova, and T. Schultz, "Improving ASR performance on non-native speech using multilingual and crosslingual information," in *Interspeech*, 2014, pp. 11–15.

[27] NIST, "SCTK," https://github.com/usnistgov/SCTK.git, 2024.