



# Fine-tuning Strategies for Automatic Speech Recognition of Low-Resource Speech with Autism Spectrum Disorder

Yeseul Park, Bowon Lee

Department of Electrical and Computer Engineering, Inha University, Republic of Korea

yspark@dsp.inha.ac.kr, bowon.lee@inha.ac.kr

## Abstract

Individuals with autism spectrum disorder (ASD) exhibit unique speech patterns that challenge conventional automatic speech recognition (ASR) systems. However, research on ASD-adapted ASR models remains limited. This study explores fine-tuning strategies for ASD-specific ASR models using Whisper, comparing full fine-tuning, selective fine-tuning, adapter tuning, and LoRA-based fine-tuning. Experiments using a small-scale Korean ASD speech dataset demonstrate that adapter tuning and LoRA significantly reduce the character error rate (CER) while reducing trainable parameters. In case of Whisper-small, adapter tuning and LoRA improve the CER by 7.22% and 10.14% over full fine-tuning, respectively. Furthermore, LoRA improved CER by 10.35% and 10.15% with Whisper-base and Whisper-large-v2 models compared to full fine-tuning. These results demonstrate the adaptation efficiency and effectiveness of LoRA for low resource ASD speech dataset.

**Index Terms:** Speech recognition, accessibility, autism spectrum disorder, parameter-efficient learning

## 1. Introduction

Conventional automatic speech recognition (ASR) systems are primarily trained on speech data from neurotypical individuals making them less effective in recognizing speech patterns of individuals with autism spectrum disorder (ASD). Atypical prosody (e.g., phrasing, stress, and resonance) and articulation distortion errors [1, 2] commonly found in ASD speech contribute to this mismatch, resulting in higher error rates. This further exacerbates communication challenges for individuals with ASD, highlighting the importance for fair and inclusive ASR systems.

Despite its significance, research on ASR systems tailored for individuals with ASD remains highly limited. Most existing studies have primarily focused on predicting the severity of ASD [3, 4, 5, 6] or classifying ASD [7, 8, 9, 10], rather than on improving speech recognition performance. Although some research has explored different ASR models [11, 12] or applied data augmentation [13] techniques, fine-tuning strategies that specifically adapt ASR models for autistic speech have not been explored to the best of our knowledge. Moreover, the scarcity of autistic speech data presents a major challenge, making it difficult to train conventional ASR models effectively.

To address this, we investigate fine-tuning strategies for adapting state-of-the-art ASR models to ASD speech using Whisper models. We compare four approaches that balance the models' efficiency and performance: (1) full fine-tuning, which updates all model parameters; (2) selective fine-tuning, which updates only key layers; (3) adapter tuning, which inserts

lightweight modules for adaptation; and (4) Low-Rank Adaptation (LoRA), which efficiently fine-tunes key weight matrices.

In particular, adapter tuning on Whisper-small allowed training only with 1.48% of the total parameters while reducing character error rate (CER) by 19.06% compared to the baseline model – pre-trained Whisper without fine-tuning and by 0.57% compared to the full fine-tuning. With LoRA, only 5.36% of the parameters were trained, with the CER decrease of 19.29% from the baseline and by 0.8% compared to the full fine-tuning.

To the best of our knowledge, this is the first study to systematically compare fine-tuning techniques for improving ASR for individuals with ASD. It highlights the importance of designing ASD-specific ASR models and demonstrates the effectiveness of parameter-efficient fine-tuning techniques with low-resource ASD speech dataset. These fine-tuning strategies reduce the computational burden while maintaining low CER. Our study can thus help the development of more accessible and inclusive ASR systems for individuals with ASD.

## 2. Methods

### 2.1. Whisper

Whisper [14] is a general-purpose speech recognition model capable of performing various tasks, including multilingual speech recognition, speech translation, speaker identification, and voice activity detection. It is a multitasking model based on a transformer encoder-decoder architecture, pre-trained on 680,000 hours of speech data. This extensive pre-training allows Whisper to maintain high accuracy across different languages and environments. Whisper varies in size and is categorized into six models: tiny, base, small, medium, large, and turbo.

### 2.2. Fine-tuning approaches

Various techniques can be applied to fine-tune the Whisper models. In this study, we compare and analyze the following fine-tuning approaches.

#### 2.2.1. Full fine-tuning

Full fine-tuning updates all parameters of the model, making it the most intuitive and commonly used fine-tuning method. Since the entire model is adjusted, it can fully adapt to the target dataset. However, this approach requires significant memory and computational resources and may lead to overfitting, especially with a limited amount of data.

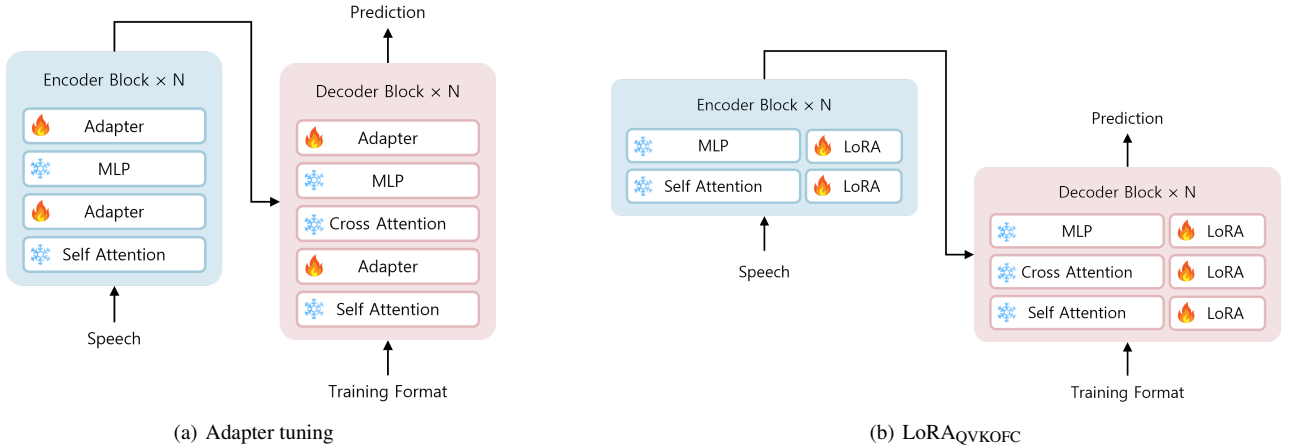


Figure 1: Architectures of adapter tuning and  $LoRA_{QV_KOFC}$  models applied to Whisper

### 2.2.2. Selective fine-tuning

Selective fine-tuning updates only specific parts of the model while keeping the remaining layers frozen. This approach allows for reduced training costs while still adapting the model effectively to the desired tasks. Our implementation focused on independently fine-tuning either the encoder or decoder components [15], allowing for a systematic evaluation of each module’s impact on overall model performance. Additionally, we incorporated the BitFit [16] method, which updates only the bias terms, to further enhance the efficiency of model adaptation.

### 2.2.3. Adapter tuning

Adapter tuning [17] involves the insertion of additional adapter layers into the model while keeping the original weights frozen, as illustrated in Fig. 1(a). Typically, these adapters follow a bottleneck structure, enabling the model to retain its original performance while adapting to a specific task [15, 18]. In this study, adapter layers with a bottleneck structure were added after the multi-head attention and output layers to fine-tune the model.

### 2.2.4. LoRA

LoRA [19] is a fine-tuning technique that introduces low-rank matrices into specific layers while keeping the original model parameters frozen, as illustrated in Fig. 1(b). This method significantly reduces memory consumption and computational costs while maintaining fine-tuning performance. LoRA is particularly useful for efficiently fine-tuning large-scale models. To evaluate its effectiveness, we applied LoRA to the transformer in two configurations:  $LoRA_{QV}$ , which adapts the attention layers by applying LoRA to  $\{W_q, W_v\}$ , and  $LoRA_{QV_KOFC}$ , which extends adaptation to both the attention and multi-layer perceptron by including  $\{W_q, W_v, W_k, W_o, W_{fc}\}$ , inspired by [20]. This allows us to analyze the impact of incorporating additional layers.

Additionally, we applied Adaptive Low-Rank Adaptation (AdaLoRA) [21], an improved version of LoRA, and conducted additional experiments. AdaLoRA dynamically updates each weight based on its importance. In this study, we applied AdaLoRA using the configuration that is identical to  $LoRA_{QV_KOFC}$  except adaptive weight updates.

## 3. Experiments

### 3.1. Data

A total of 6,699 utterances, amounting to 7.74 hours of speech data, were collected from four speakers with ASD. The dataset consists of commonly used commands in daily life, and the speakers read the sentences selected from the script from AI-Hub speech command dataset, which was collected from neurotypical male and female speakers [22]. Each speaker recorded randomly selected commands from a total of 9,000 sentences, and the speech data were collected at a 16 kHz sampling rate. All utterances were recorded in Korean using laptops in a classroom setting to minimize variability due to external factors. Each speaker’s dataset was split into training, validation, and test sets in a ratio of 8:1:1.

Because the primary objective is comparing the fine-tuning strategies to improve the performance of ASR models for ASD speakers, and since there are only four speakers in our dataset, we did not rule out any of the speakers for training except the 8:1:1 split. Table 1 shows differences in the statistics of the dataset regarding the maximum utterance duration compared to the average utterance duration, which demonstrate huge variabilities in speech characteristics of individuals with ASD.

Table 1: The first letter of the ID denotes gender (‘M’ for male, ‘F’ for female). Utt. indicates the number of utterances, while Avg., Min, and Max duration represent the average, minimum, and maximum utterance lengths (in seconds), respectively.

ID	Utt.	Avg. duration	Total duration	Min duration	Max duration
M1	1278	3.39 s	1.2 h	1.54 s	6.4 s
M2	1918	4.08 s	1.42 h	1.54 s	13.57 s
M3	1259	4.63 s	2.46 h	2.3 s	21.25 s
F1	2244	4.25 s	2.64 h	1.54 s	11.26 s

### 3.2. Metric

Phonetic boundaries in Korean are often ambiguous, and individuals with ASD may exhibit atypical pronunciation patterns. Furthermore, command-style speech data tend to be relatively short. Given these characteristics, a robust error rate evaluation

is crucial, even for small-scale datasets. To ensure reliability, we employ CER as our evaluation metric which is calculated as follows:

$$CER = \frac{D + S + I}{N}, \quad (1)$$

where Deletion ( $D$ ), Substitution ( $S$ ), and Insertion ( $I$ ) denote the number of deleted, substituted, and inserted characters, respectively, while  $N$  represents the total number of characters in the label text.

### 3.3. Training setup

We compared and evaluated the performance of various fine-tuning techniques using the Whisper model. The key hyperparameters for each method were set as follows. For the adapter, we adopted the adapter library from AdapterHub<sup>1</sup> using the adapter proposed in [23] and enabled both the multi-head adapter and output adapter. The reduction factor was set to 16, and ReLU was used as the activation function. For LoRA and AdaLoRA, we followed the prior LoRA-Whisper study [20] and set the rank to 32, the LoRA scaling factor to 64, and the dropout probability to 0.1. For full fine-tuning, the learning rate was set to  $3e-6$ , while BitFit and adapter tuning used a learning rate of  $1e-3$ . For all other fine-tuning methods, a learning rate of  $1e-4$  was applied. All methods used a warmup step of 500 and a batch size of 32. The maximum number of epochs is set to 50.

### 3.4. Hardware settings

For the experiments, we employed a workstation equipped with four NVIDIA A100 80 GB GPUs and an AMD EPYC 7543 32-Core Processor, featuring 48 CPUs with a base frequency of 2.0 GHz. The system has a total of 900 GB of RAM.

## 4. Results

### 4.1. Comparison of the fine-tuning methods

Table 2 presents various fine-tuning methods for Whisper-small, along with the number of trained parameters and their corresponding CERs. In the case of full fine-tuning, the CER significantly improved to 7.89% from 26.38% of the baseline. Meanwhile, selective fine-tuning (encoder fine-tuning, decoder fine-tuning, BitFit) also resulted in improvements over the baseline, albeit with higher CERs than that of full fine-tuning. Interestingly for BitFit, despite adjusting only 0.09% of the parameters by fine-tuning only the bias terms, it achieved better performance than fine-tuning the entire decoder. This highlights the potential of minimizing trainable parameters while still achieving effective performance gains.

Adapter tuning employed a bottleneck-structured adapter, fine-tuning only 1.48% of the overall model parameters. As a result, the CER decreased by 0.57% compared to full fine-tuning, demonstrating high performance with relatively low training costs. In the case of LoRA, LoRA<sub>QV</sub> used a similar number of parameters as adapter tuning but exhibited a CER increase of 1.66%, showing relatively lower performance. When the target module was extended to LoRA<sub>QV</sub>KOFC, the number of trainable parameters increases to 5.36%. Nevertheless, it achieved the lowest CER of 7.09% with a relatively low number of trainable parameters. This suggests that the  $\{W_k, W_o, W_{fc}\}$  module plays a crucial role in regulating the model’s key information flow. Additionally, Figure 2 compares character error

<sup>1</sup><https://adapterhub.ml/>

Table 2: Comparison of trainable parameters, trainable ratio, and CER for different methods applied to Whisper-small.

Method	# Trainable Parameters	Trainable Ratio (%)	CER (%)
Baseline	-	-	26.38
Full fine-tuning	241M	100	7.89
Encoder fine-tuning	116M	48.19	9.91
Decoder fine-tuning	153M	63.53	12.22
BitFit	224K	0.09	11.26
Adapter tuning	3M	<b>1.48</b>	<b>7.32</b>
LoRA <sub>QV</sub>	3M	1.46	8.98
LoRA <sub>QV</sub> KOFC	12M	<b>5.36</b>	<b>7.09</b>
AdaLoRA <sub>QV</sub> KOFC	12M	5.36	7.29

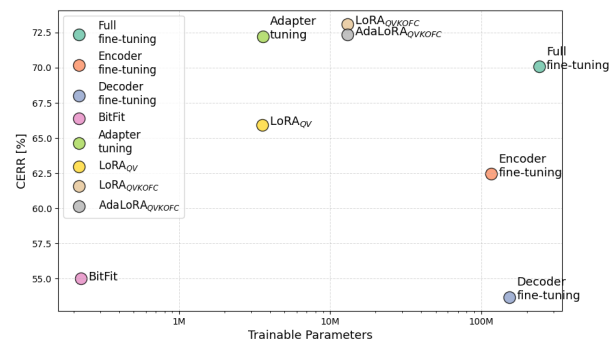


Figure 2: Comparison of CER by the number of trainable parameters: A relative comparison of how efficiently various fine-tuning strategies improve CER.

rate reduction (CERR) based on the number of trained parameters.

Figure 3 presents a comparison of CERs across different fine-tuning methods for each speaker. As shown in Fig. 3, for Speaker F1, the large number of data samples led to a significant reduction in CER, suggesting that the F1 speaker’s data contributed favorably to model training. In contrast, despite having more data than speakers M1 and M3, speaker M2 exhibited a relatively higher CER. This implies that Speaker M2’s speech may have distinct characteristics that differ from the generalized speech learned by the ASR model. These findings highlight the necessity of further research to enhance the generalization of ASR models and better accommodate diverse speaker characteristics.

The performance variance in decoder fine-tuning across speakers was the largest among all methods. This suggests that adjusting only the decoder may hinder the model from consistently capturing speech context and utterance characteristics, leading to ineffective adaptation to certain speakers or in specific speaking environments. These indicate that fine-tuning the decoder alone could reduce the model’s generalization ability, emphasizing the need for additional adjustments for fine-tuning.

Figure 4 presents the CER distributions across different fine-tuning methods, highlighting variations in performance. Despite having a similar average CER compared to adapter tuning and LoRA<sub>QV</sub>KOFC, AdaLoRA<sub>QV</sub>KOFC exhibited noticeable outliers. This result implies that AdaLoRA’s dynamic adaptation mechanism did not fully harmonize with the internal structure of the Whisper-small model. In other words, the high per-

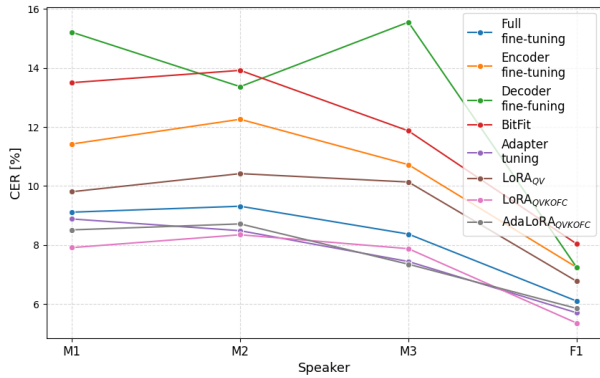


Figure 3: CER Comparison of fine-tuning methods by speaker: Evaluating the performance differences of fine-tuning methods across different speakers.

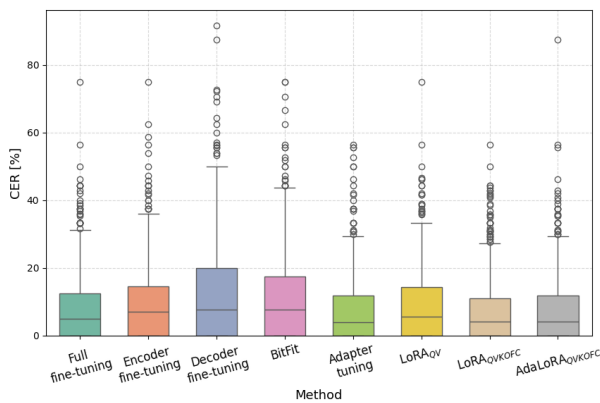


Figure 4: Comparison of CER distributions by fine-tuning Methods: A relative comparison of outliers across different fine-tuning methods to assess performance variability.

formance variation across certain data samples may act as a limitation of AdaLoRA.

Experimental results showed that LoRA<sub>QVKOFC</sub> achieved the best performance, followed by AdaLoRA<sub>QVKOFC</sub> and adapter tuning. Notably, LoRA<sub>QVKOFC</sub> consistently maintained high performance compared to other methods. Adapter tuning uses the fewest number of parameters, and AdaLoRA outperformed adapter tuning with outliers. These findings indicate that determining the optimal fine-tuning strategy is challenging based solely on performance comparisons. Nevertheless, LoRA<sub>QVKOFC</sub> demonstrated the most stable and superior performance, making it a promising candidate for the ASD-adaptive ASR model.

#### 4.2. Performance comparison by model sizes

As confirmed in subsection 4.1, LoRA<sub>QVKOFC</sub> demonstrated the best performance in the Whisper-small, even under the constrained ASD dataset environment. To verify whether the effectiveness of LoRA<sub>QVKOFC</sub> remains consistent across different model sizes, we compared the performance of Whisper-base (a smaller model than Whisper-small) and Whisper-large-v2 (a larger model than Whisper-small). When applying LoRA<sub>QVKOFC</sub>, the Whisper-base, Whisper-small, and Whisper-large-v2 models exhibited performance improvements

Table 3: Comparison of trainable parameters, trainable ratio, CER, and CERR for different fine-tuning methods based on model size.

Method	# Trainable Parameters	Trainable Ratio (%)	CER (%)	CERR (%)
<b>Whisper-base</b>				
Baseline	-	-	36.62	-
Full fine-tuning	72M	100	12.46	65.59
LoRA <sub>QVKOFC</sub>	4M	<b>5.95</b>	<b>11.17</b>	<b>69.49</b>
<b>Whisper-small</b>				
Baseline	-	-	26.38	-
Full fine-tuning	241M	100	7.89	70.09
LoRA <sub>QVKOFC</sub>	12M	<b>5.36</b>	<b>7.09</b>	<b>73.12</b>
<b>Whisper-large-v2</b>				
Baseline	-	-	19.31	-
Full fine-tuning	1543M	100	5.02	74
LoRA <sub>QVKOFC</sub>	57M	<b>3.73</b>	<b>4.51</b>	<b>76.64</b>

of 69.49%, 73.12%, and 76.64% in terms of CERR, respectively, compared to their baselines. This suggests that as model size increases, LoRA<sub>QVKOFC</sub> becomes increasingly efficient. This study compares the performance of baseline, full fine-tuning, and LoRA<sub>QVKOFC</sub>.

The experimental results indicate that LoRA<sub>QVKOFC</sub> achieved better performance than full fine-tuning in the Whisper-base and Whisper-large-v2 models, similar to its effectiveness in Whisper-small. Furthermore, the proportions of trainable parameters in Whisper-base, Whisper-small, and Whisper-large-v2 were 5.95%, 5.36%, and 3.73%, respectively. Compared to full fine-tuning, this corresponds to 18×, 20×, and 27× reductions in trainable parameters while improving the CER. These results indicate that LoRA<sub>QVKOFC</sub> is effective for low-resource datasets and that larger models benefit from relatively higher training efficiency.

In summary, LoRA<sub>QVKOFC</sub> serves as a robust fine-tuning technique with consistent performance regardless of model sizes. Particularly for large-scale models, LoRA<sub>QVKOFC</sub> offers a highly efficient alternative for fine-tuning with low-resource ASD speech dataset.

## 5. Conclusion

In this study, we explored and validated efficient and effective ASR fine-tuning strategies for ASD speakers using a small-scale Korean ASD speech dataset. By utilizing fine-tuning techniques that incorporate adapters and LoRA targeting specific modules, we demonstrated that it is possible to achieve low CER while efficiently managing the number of trainable parameters. Additionally, regardless of the model sizes, LoRA<sub>QVKOFC</sub> consistently achieves a lower CER compared to full fine-tuning, with its efficiency improving as the model size increases.

These findings suggest that a more lightweight approach could be practical for developing speech recognition technologies tailored for individuals with ASD with low-resource datasets. Future research would benefit from validating these methods across more diverse environments, datasets, and a broader range of speakers, as the current study is limited by the relatively small number of speakers. Furthermore, the proposed methodology has the potential to be adapted for users with various disabilities, contributing to the expansion of inclusive ASR technologies.

## 6. Acknowledgements

This work was supported by the Ministry of Science and ICT of the Republic of Korea and the National Research Foundation of Korea (NRF-2023R1A2C2006725).

## 7. References

- [1] I. Vogindroukas, M. Stankova, E.-N. Chelas, and A. Proedrou, "Language and speech characteristics in autism," *Neuropsychiatric Disease and Treatment*, vol. 18, pp. 2367–2377, 2022. [Online]. Available: <https://www.tandfonline.com/doi/abs/10.2147/NDT.S331987>
- [2] L. D. Shriberg, R. Paul, J. L. McSweeney, A. Klin, D. J. Cohen, and F. R. Volkmar, "Speech and prosody characteristics of adolescents and adults with high-functioning autism and asperger syndrome," 2001.
- [3] J. Mun, S. Kim, and M. Chung, "Developing an end-to-end framework for predicting the social communication severity scores of children with autism spectrum disorder," in *Interspeech 2024*, 2024, pp. 1430–1434.
- [4] S. Lee, J. Mun, S. Kim, and M. Chung, "Speech corpus for Korean children with autism spectrum disorder: Towards automatic assessment systems," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 15 160–15 170. [Online]. Available: <https://aclanthology.org/2024.lrec-main.1318/>
- [5] M. Eni, I. Dinstein, M. Ilan, I. Menashe, G. Meiri, and Y. Zigel, "Estimating autism severity in young children from speech signals using a deep neural network," *IEEE Access*, vol. 8, pp. 139 489–139 500, 2020.
- [6] A. Xu, R. Hebbur, R. Lahiri, T. Feng, L. Butler, L. Shen, H. Tager-Flusberg, and S. Narayanan, "Understanding spoken language development of children with asd using pre-trained speech embeddings," in *Interspeech 2023*, 2023, pp. 4633–4637.
- [7] N. A. Chi, P. Washington, A. Kline, A. Husic, C. Hou, C. He, K. Dunlap, and D. P. Wall, "Classifying autism from crowdsourced semistructured speech recordings: Machine learning model comparison study," *JMIR Pediatr Parent*, vol. 5, no. 2, p. e35406, Apr 2022. [Online]. Available: <https://pediatrics.jmir.org/2022/2/e35406>
- [8] A. Mohanta and V. K. Mittal, "Classifying speech of asd affected and normal children using acoustic features," in *2020 National Conference on Communications (NCC)*, 2020, pp. 1–6.
- [9] J. Li, M. Hasegawa-Johnson, and K. Karahalios, "Enhancing child vocalization classification with phonetically-tuned embeddings for assisting autism diagnosis," in *Interspeech 2024*, 2024, pp. 5163–5167.
- [10] X. Yin, C. Zhang, and W. Wang, "A discriminative multi-task learning for autism classification based on speech signals," in *2023 IEEE Symposium on Computers and Communications (ISCC)*, 2023, pp. 386–391.
- [11] A. Ashvin, R. Lahiri, A. Kommineni, S. Bishop, C. Lord, S. R. Kadiri, and S. Narayanan, "Evaluation of state-of-the-art asr models in child-adult interactions," 2024. [Online]. Available: <https://arxiv.org/abs/2409.16135>
- [12] S. Lee, J. Mun, S. Kim, H. Park, S. Yang, H. Kim, S. Noh, W. Kim, and M. Chung, "Automatic speech recognition and assessment systems incorporated into digital therapeutics for children with autism spectrum disorder," in *International Conference on Computers Helping People with Special Needs*. Springer, 2024, pp. 328–335.
- [13] R. Gale, L. Chen, J. Dolata, J. van Santen, and M. Asgari, "Improving asr systems for children with autism and language impairment using domain-focused dnn transfer techniques," in *Interspeech 2019*, 2019, pp. 11–15.
- [14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: <https://proceedings.mlr.press/v202/radford23a.html>
- [15] S. Radhakrishnan, C.-H. H. Yang, S. A. Khan, N. A. Kiani, D. Gomez-Cabrero, and J. N. Tegner, "A parameter-efficient learning approach to arabic dialect identification with pre-trained general-purpose speech model," in *Interspeech 2023*, 2023, pp. 1958–1962.
- [16] E. Ben Zaken, Y. Goldberg, and S. Ravfogel, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1–9. [Online]. Available: <https://aclanthology.org/2022.acl-short.1/>
- [17] 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 *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 2790–2799. [Online]. Available: <https://proceedings.mlr.press/v97/houlsby19a.html>
- [18] Y. Liu, X. Yang, and D. Qu, "Exploration of whisper fine-tuning strategies for low-resource asr," *EURASIP J. Audio Speech Music Process.*, vol. 2024, no. 1, Jun. 2024. [Online]. Available: <https://doi.org/10.1186/s13636-024-00349-3>
- [19] 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," 2021. [Online]. Available: <https://arxiv.org/abs/2106.09685>
- [20] Z. Song, J. Zhuo, Y. Yang, Z. Ma, S. Zhang, and X. Chen, "LoRA-Whisper: Parameter-efficient and extensible multilingual ASR," in *Interspeech 2024*, 2024, pp. 3934–3938.
- [21] Q. Zhang, M. Chen, A. Bukharin, P. He, Y. Cheng, W. Chen, and T. Zhao, "Adaptive budget allocation for parameter-efficient fine-tuning," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: <https://openreview.net/forum?id=lq62uWRJjiY>
- [22] "Command speech (general male & female) dataset," AI-Hub, 2025. [Online]. Available: <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=96>
- [23] C. Poth, H. Sterz, I. Paul, S. Purkayastha, L. Engländer, T. Imhof, I. Vulić, S. Ruder, I. Gurevych, and J. Pfeiffer, "Adapters: A unified library for parameter-efficient and modular transfer learning," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 149–160. [Online]. Available: <https://aclanthology.org/2023.emnlp-demo.13>