Benchmarking Children's ASR with Supervised and Self-supervised Speech Foundation Models

Ruchao Fan[1]*, Natarajan Balaji Shankar*[1]*, Abeer Alwan*[1]

[1]Dept. of Electrical and Computer Engineering, University of California, Los Angeles, U.S.A

fanruchao,balaji1312@g.ucla.edu, alwan@ee.ucla.edu

Abstract

Speech foundation models (SFMs) have achieved state-of-the-art results for various speech tasks in supervised (e.g. Whisper) or self-supervised systems (e.g. WavLM). However, the performance of SFMs for child ASR has not been systematically studied. In addition, there is no benchmark for child ASR with standard evaluations, making the comparisons of novel ideas difficult. In this paper, we initiate and present a comprehensive benchmark on several child speech databases based on various SFMs (Whisper, Wav2vec2.0, HuBERT, and WavLM). Moreover, we investigate finetuning strategies by comparing various data augmentation and parameter-efficient finetuning (PEFT) methods. We observe that the behaviors of these methods are different when the model size increases. For example, PEFT matches the performance of full finetuning for large models but worse for small models. To stabilize finetuning using augmented data, we propose a perturbation invariant finetuning (PIF) loss as a regularization.[1].

Index Terms: Children's Speech Recognition, Speech Foundation Model, Whisper, Data Augmentation, PEFT

1. Introduction

Large foundation models have increasingly gained attention in the research community because of their impressive zero-shot and in-context learning ability [1, 2, 3]. Specifically in the speech area, the Whisper-large model [4] has shown great robustness to diverse domains of speech data by learning from large-scale supervised data in a multi-task training setting. In addition to Whisper, another type of speech foundation models (SFM) is obtained through self-supervised learning, e.g. Wav2vec2.0 [5], HuBERT [6], WavLM [7], and W2VBERT2.0 [8]. Such models do not require annotations and learn to extract contextual representations based on data patterns in the speech signals [9]. State-of-the-art results of various speech recognition tasks can be achieved by finetuning these models or using them as feature extractors.

With the increasing use of voice-based educational technology, better child ASR systems are needed because speech is one of the mechanisms young children use to interact with devices due to their limited reading and writing abilities. However, child ASR is difficult due to, in part, the lack of large child speech databases. To address this issue, researchers have developed a variety of data augmentation methods by perturbation [10, 11, 11, 12] or voice conversion [13, 14]. Another direction is to adopt the pretraining finetuning paradigm, which utilizes the un-annotated data with self-supervised learning [15, 16] or the annotated adult data with transfer learning [17] in the pretraining stage. The knowledge learned in the pretrained models can greatly improve the performance for child ASR. With the recent advances of SFMs, several studies have compared the performance of widely used SFMs on child speech [18, 19, 20]. However, these studies provide only full finetuning experiments on SFMs. In addition, the speech corpora used in these studies are partitioned differently in each one, making direct performance comparisons difficult.

In this paper, we initiate and present a comprehensive benchmark on the OGI [21] read and MyST spontaneous [22] child speech corpora, studying the performance of various SFMs. More importantly, we investigate finetuning strategies for child speech by comparing various data augmentation and parameter-efficient finetuning (PEFT) methods, which are not discussed in previous works. We observe many interesting behaviors of finetuning different speech foundation models. For example, adapter finetuning [23] is better than full finetuning for large models but vice versa for small models. Our benchmarking study may offer guidance in selecting appropriate models, data augmentation and PEFT strategies to develop robust and accurate child ASR systems. We hope the standard evaluation can lead to fairer comparisons for child ASR research. We also welcome new evaluation sets and new algorithms to include in the benchmark.

2. Methods in the Benchmark

In this section, we briefly introduce the methods that are compared in the benchmark, including SFMs, data augmentation and parameter-efficient finetuning (PEFT) techniques.

2.1. Speech Foundation Models

Large models trained with large amounts of data have shown great potential to improve the performance of speech recognition tasks. There are two types of SFMs that are: 1) trained with supervised speech-text pairs, such as Whisper [4] and Parakeet [24], and 2) trained with unannotated speech data using self-supervised learning, e.g. Wav2vec2.0 [5], HuBERT [6], and WavLM [7]. For supervised SFMs, the zero-shot ability of these models is compared as they can directly perform speech recognition tasks. We then conduct in-depth finetuning experiments on the Whisper series (tiny, base, small, medium, large and largeV3). For self-supervised SFMs, finetuning experiments are conducted. The models used in the benchmark are listed in Table 1 along with details including model architecture, input features, model size and training data. The open-sourced

This work was supported in part by the NSF.

[1]Our code is available at https://github.com/Diamondfan/SPAPL_KidsASR

Table 1: *Details of the speech foundation models used in the benchmark.*

Model	Model Architecture	Input Features	Model Size	Sup./Self-sup.	Training Data (hours)
Whisper-{tiny-large}	Encoder-Decoder	Fbank	39M-1550M	Supervised	680K
Whisper-largeV3 [4]	Encoder-Decoder	Fbank	1550M	Supervised	1M
Canary [24]	Encoder-Decoder	Fbank	1B	Supervised	85K
Parakeet-TDT [24]	Transducer	Fbank	1.1B	Supervised	64K
Wav2vec2-Large [5]	Encoder	Waveform	311M	Self-supervised	60k
HuBERT-Large [6]	Encoder	Waveform	311M	Self-supervised	60k
WavLM-Large [7]	Encoder	Waveform	311M	Self-supervised	94k

models can be accessed in the OpenASR leaderboard[2].

2.2. Data Augmentation

Data augmentation methods are commonly used in child ASR systems for alleviating the data scarcity problem, but no systematic comparison between them has been conducted before with all SFM models. Based on the Whisper-small model, we compare several widely-used methods including pitch perturbation (PP) [25], speed perturbation (SP) [10], vocal tract length perturbation (VTLP) [26], and SpecAugment (SA) [27]. Two augmented utterances were created for each utterance as has commonly been done in the literature.

- **Pitch perturbation (PP)** involves altering the fundamental frequency of speech signals while preserving other temporal /spectral features. The pitch is shifted to $n/12$ octave higher or lower for each utterance, where n is randomly sampled from 1 to 12 twice.
- **Speed perturbation (SP)** modifies the speed of speech signals. Two copies of each utterance are created with the perturbation rate of 0.9 and 1.1.
- **Vocal tract length perturbation (VTLP)** involves simulating the effects of variations in vocal tract length by applying frequency warping. The perturbation rate used (0.9 and 1.1) is the same as that used in speed perturbation.
- **SpecAugment (SA)** randomly masks consecutive frequency bands and time frames, which effectively increases the robustness of the model to time-frequency variations. We use the default SpecAug settings in the Whisper model.

We look forward to incorporating new augmentation methods in the benchmark in the future.

2.3. Parameter Efficient Finetuning (PEFT)

Parameter-efficient finetuning techniques have become increasingly important when large foundation models are used as model initialization for various tasks [28]. These techniques aim to adapt pretrained models to new tasks or domains while minimizing the computational resources required for training. We compare four widely-used PEFT techniques, which are Low Rank Adaptation (LoRA) [29], adapter tuning [30, 23], prompt tuning [31], and prefix tuning [32].

- LoRA leverages the observation that the matrices of model layers often exhibit low-rank structures. By decomposing weight matrices into low-rank factors and updating only the low-rank factors during fine-tuning, LoRA reduces computational overhead while preserving model performance. We apply LoRA weights to both the query and value-related parameters in each attention layer, with a rank of eight [29].
- Adapter tuning introduces lightweight adapter modules, which are small neural network components inserted between layers of the foundation models. By finetuning only the parameters within the adapters, efficient adaptation is achieved. We used the residual adapters with the bottleneck dimension of 32, which is similar to [30]. The residual adapters are inserted after each block in both the encoder and decoder.
- Prompt Tuning prepends randomly initialized prompt vectors to the input sequence and the prompts are optimized through gradient-based methods, allowing the model to directly learn task-specific input representations during fine-tuning. 100 and 20 prompts are inserted in the encoder and decoder inputs, respectively, in our experiments.
- Prefix tuning is similar to prompt tuning but prepends the prompts at each layer instead of at the input, bringing more flexibility during finetuning. In our experiments, 50 and 10 prompts are inserted at the beginning of each layer input to both the encoder and decoder modules, respectively.

The number of prompts used in prompt tuning and prefix tuning are chosen empirically. By comparing various PEFT methods to full finetuning, we will discover the best finetuning strategy for child ASR when using speech foundation models.

3. Experiments

In this section, we present the speech datasets used, experimental setup and results..

3.1. Child Speech Datasets

The experiments are conducted on two child speech databases: My Science Tutor (MyST) spontaneous speech corpus [22], and CSLU OGI scripted read speech corpus [21].

The MyST corpus consists of around 240 hours of transcribed conversational children's speech (from grade 3 to grade 5), recorded from virtual tutoring sessions in physics, geography, biology, and other topics. Similar to [19], we identify and filter low quality audio samples by passing the transcribed dataset through the Whisper-largeV2 model. Utterances with WER larger than 50% or with less than 3 words are removed, resulting in a 133 hours training set. When evaluating the Whisper model, we find that the results are unstable for the test samples that are longer than 30s (the maximum length for training Whisper). Hence, utterances longer than 30s are also removed in both the training and test sets. As a result, the original data splits in MyST corpus are as follows: train (133h), dev (21h), and test (25h).

The CSLU OGI Kids corpus contains around 50 hours of speech by 1100 children (from kindergarten to grade 10) read-

[2] https://huggingface.co/spaces/hf-audio/open_asr_leaderboard

Table 2: *Zero-shot performance of the supervised speech foundation models in terms of WER. Bold numbers are the best performance among the supervised SFMs.*

Model	Model Size	MyST		OGI	
		dev	test	dev	test
Whisper-tiny	39M	18.5	20.6	40.1	53.8
Whisper-base	74M	15.6	16.8	36.8	38.0
Whisper-small	242M	14.4	13.4	21.2	25.4
Whisper-medium	769M	13.3	13.1	18.8	20.8
Whisper-large	1.55B	14.4	12.5	21.2	22.9
Whisper-largeV3	1.55B	12.3	12.6	14.9	19.9
Canary	1.0B	**9.3**	**9.5**	14.8	18.2
Parakeet-rnnt	1.1B	10.7	11.1	**14.3**	**16.7**

ing from a list that contains either simple words, sentences, or digit strings. The utterances are randomly split into train (70%), development (15%) and test (15%) sets without speaker overlap, which is the same as [23, 33].

The utterance ID list for the two corpora are released as standard evaluations with the training code.

3.2. Finetuning and Evaluation Setup

When finetuning the supervised SFMs, we use the same vocabulary and objective function as those used in the pretraining stage. When finetuning self-supervised SFMs, we use all characters in the transcriptions to create a vocabulary and apply a CTC loss to perform ASR. All results are reported by greedy search decoding without any external language model. The PEFT and DA experiments are not investigated for the OGI corpus because of the page limitation.

3.3. Zero-shot Performance for the Supervised SFMs

Since supervised speech foundation models are trained with ASR loss, we first compare their zero-shot abilities on child speech. Top performing models from the OpenASR benchmark for adult speech are selected for comparisons. The results are presented in Table 2. The Canary and Parakeet models have been shown to perform better than Whisper, on average, on the adult speech benchmark [34]. The same conclusions can be drawn here for child speech, which is surprising because the Whisper models are trained with more data than Canary and Parakeet (training data sizes are shown in Table 1). Considering that many of the data for Whisper training is weakly-supervised, we conclude that data quality is sometimes more important than the size of data for obtaining a robust supervised speech foundation model, which has also been observed for large language models [35].

3.4. Foundation Models with Finetuning

In addition to the supervised foundation models, self-supervised foundation models are also widely used for child ASR. We compare the full-finetuning performance between the two types of foundation models and present the results in Table 3. The results show that supervised SFMs can achieve better performance than the self-supervised SFMs after finetuning with similar model parameters (e.g. Whisper-small with 242M and WavLM with 311M parameters). Among the most widely used SSL models, WavLM achieves the best performance because it used more data and included a masked reconstruction from noisy and

Table 3: *WER comparisons of finetuning supervised and self-supervised speech foundation models. Note finetuning on each corpus separately. Supervised and self-supervised SFMs are finetuned with 2GPUs for 4k and 12k steps, respectively.*

Model	MyST			OGI	
	Dev	Test	Time	Dev	Test
Supervised SFM					
Whisper-tiny	11.6	11.6	2.0h	2.7	3.0
Whisper-base	9.1	10.4	2.5h	2.0	2.3
Whisper-small	8.4	9.3	6.0h	5.0	1.8
Whisper-medium	8.4	**8.9**	8.0h	**1.6**	1.5
Whisper-large	**8.2**	13.0	9.2h	1.8	1.7
Whisper-largeV3	8.5	9.1	13.0h	**1.6**	**1.4**
Self-supervised SFM					
Wav2vec2.0	10.6	11.1	10.5h	2.1	2.5
HuBERT	10.5	11.3	10.5h	2.2	2.5
WavLM	9.6	10.4	13.5h	1.7	1.8

multi-talker speech data during pretraining. Note that an advantage of the SSL models are that they might also be robust to other speech tasks because the SSL loss is not specifically designed for ASR. We don't compare this ability of SSL models since we are mainly focusing on the ASR task. The full finetuning results for the Canary and Parakeet models are as follows: Canary (MyST dev: 8.6, MyST test: 9.2, OGI dev: **1.4**, OGI test: 1.5); Parakeet (MyST dev: **7.9**, MyST test: **8.5**, OGI dev: 1.8, OGI test: 1.8). Due to the recent release of these models, PEFT and DA experiments were not explored in detail.

3.5. Comparisons of Data Augmentation Methods

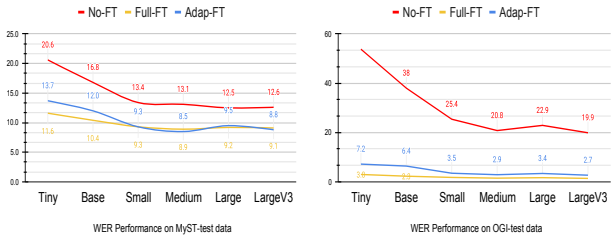
Table 4: *WER comparisons of different data augmentation methods on MyST dataset using the Whisper-small model. PIF stands for perturbation invariant training. x3 indicates three copies of the original data for augmentation.*

Whisper-small	Augmentation	MyST-dev	MyST-test
Baseline	no	14.4	13.4
Finetuning	no	8.4	9.3
	PP (x3)	8.6	**8.8**
	VTLP (x3)	8.6	9.0
	SP (x3)	**8.1**	8.9
	SA	8.2	9.0
	SA + PP	8.2	8.9
	SA + VTLP	**8.1**	9.0
	SA + SP	8.3	8.9
PIF	VTLP (x3)	8.3	9.0
	PP (x3)	8.3	8.9

Data augmentation (DA) is an important technique to deal with low-resource tasks, such as child ASR. However, previous works either used private data or conducted experiments with their own settings, making the comparisons of different methods difficult. In addition, previous DA methods are proposed based on training from scratch. It is unknown whether these methods improve the performance when using SFMs. To address this issue, we made a comparison of different DA methods and explored their role in finetuning SFMs. The experimental results on MyST dataset using Whisper-small model are shown in Table 4. The reason we use the Whisper-small model is that it is computationally efficient given our limited number

Table 5: *WER comparisons of different parameter efficient finetuning (PEFT) methods on MyST dataset using the Whisper-small model. Params indicates the number of updated parameters during finetuning. Enc. and Dec. represents finetuning encoder and decoder only, respectively.*

Model	PEFT	MyST-dev	MyST-test	Params
Baseline	no	14.4	13.4	0
Full-FT	no	**8.4**	9.3	242M
Whisper -small	Enc.	9.0	**9.2**	88M
	Dec.	8.9	9.5	154M
	Prompt [2]	10.4	10.4	92k
	Prefix [32]	8.9	10.2	541k
	LoRA [29]	9.1	9.6	917k
	Adapter [23]	**8.4**	9.3	1.29M



(a) *MyST-test* (b) *OGI-test*

Figure 1: *The impact of the Whisper model size for full and adapter finetuning (Adap-FT in the figure) on WER. The model size of each whisper model can be found in Table 1.*

of GPUs, and achieves a reasonable WER on child speech. We can observe from the table that different augmentation methods achieve similar WER improvements compared to the finetuning baseline. Interestingly, the combination of two DA methods does not provide further gains compared to using only one method. This is slightly different from the conclusion in [36] when the model is trained from scratch. This might be because the SFM itself is already robust to some variations created by the DA methods. Note that F0-based data augmentation in [36] achieves similar performance to pitch perturbation.

The improvements of PP and VTLP are not stable, and we propose a perturbation invariant finetuning (PIF) technique to stabilize the VTLP and PP. Specifically, an additional distance loss between the encoder outputs of original and perturbed utterance is added as a regularization for finetuning. The results in Table 4 show that PIF can lead to more consistent improvements of perturbation methods on the MyST-dev and MyST-test sets. PIF is only used for VTLP and PP as they are not stable when FT on kids speech while other DAs are stable.

3.6. Comparisons of Parameter Efficient Finetuning

When speech foundation models are large, full finetuning with the entire model parameters would be difficult because of the high GPU memory costs. Parameter efficient finetuning (PEFT) can retain the performance of full tuning but update less parameters during the finetuning stage. We compare several widely used PEFT methods in the NLP area on Whisper-small model on the MyST dataset and present the results in Table 5. It can be seen from the table that adapter tuning achieves similar performance compared to the full finetuning while having only 1.29M parameters for updates. Note that the initialization of the adapters are important for good performance of adapter tuning. For example, the inserted adapter module should be equivalent to the identity function at the start of the finetuning. However, LoRA, the most popular PEFT method in the area of NLP, achieves worse performance than the full finetuning. Prompt and prefix tuning behave not as good as LoRA and adapter FT might be because they alter the positional information of the speech sequence and restrict the model capacity for learning from finetuning data.

3.7. Impact of Model size on PEFT performance

As shown in Table 3, the WER of the Whisper model decreases when the model size increases. We further explore whether the model size would affect the performance of PEFT, specifically adapter tuning because it behaves better than other PEFTs as shown in Table 5. The results of both full finetuning and adapter finetuning on MyST and OGI test data are plotted in Figure 1. We can observe from the figure that the adapter tuning does not work as well as the full finetuning for small models. However, when the model size increases, the gap between adapter tuning and full finetuning decreases. For example, the adapter tuning matches the performance of the full finetuning for the Whisper-largeV3 on the MyST-test data. This interesting behavior provides us with guidance on how to select the appropriate finetuning strategy. That is, performing full finetuning for small models and PEFT for large models. It would also be interesting to investigate the impact of the model size for data augmentation methods, which will be included in future work.

4. Conclusions and Future Work

In this paper, we presented the first benchmark for child ASR with a comparison of various speech foundation models, such as Whisper, Canary, Parakeet, Wav2vec2.0, HuBERT, and WavLM. We found that the Canary and Parakeet models are better than Whisper models on child speech with much less training data, indicating the data quality is sometimes more important than the data quantity. As expected, supervised SFMs performed better than the self-supervised SFMs after finetuning. Moreover, we investigated finetuning strategies by comparing various data augmentation (pitch perturbation, speed perturbation, VTLP and SpecAugment) and parameter-efficient finetuning (PEFT) methods (prompt tuning, prefix tuning, adapter tuning, and LoRA). To stabilize the finetuning using the augmented data, we propose a perturbation invariant finetuning (PIF) loss as a regularization. Various parameter-efficient finetuning (PEFT) strategies were compared, and we observed that the behaviors of PEFT are different when the model size increases. For example, PEFT performed better than full finetuning for large models but worse for small models. This study may offer guidance in selecting appropriate models, data augmentation and PEFT strategies to develop robust child ASR systems.

Future work will include: 1) Evaluations on other child speech datasets; 2) Comparisons with new data augmentation methods; 3) Evaluations of other open-sourced speech foundation models, such as SeamlessM4T [37], OWSM [3] and W2VBERT2.0 [8]; 4) Migration of models not supported in Huggingface, e.g. the Canary and Parakeet models developed using the NeMo [38] framework, since our finetuning code is implemented based on Huggingface.

5. References

[1] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, "Emergent abilities of large language models," *Transactions on Machine Learning Research*, 2022.

[2] Y. Li, Y. Wu, J. Li, and S. Liu, "Prompting large language models for zero-shot domain adaptation in speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[3] Y. Peng, J. Tian *et al.*, "Reproducing whisper-style training using an open-source toolkit and publicly available data," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[4] 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, ICML 2023*, vol. 202. PMLR, 2023, pp. 28 492–28 518.

[5] A. Baevski *et al.*, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[6] W.-N. Hsu, B. Bolte *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[7] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[8] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 244–250.

[9] Y. Zhang, D. S. Park *et al.*, "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1519–1532, 2022.

[10] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," *Sixteenth annual conference of the international speech communication association*, 2015.

[11] G. Yeung, R. Fan, and A. Alwan, "Fundamental frequency feature warping for frequency normalization and data augmentation in child automatic speech recognition," *Speech Communication*, vol. 135, pp. 1–10, 2021.

[12] H. K. Kathania, M. Singh *et al.*, "Data augmentation using prosody and false starts to recognize non-native children's speech," in *Interspeech 2020*. ISCA, 2020, pp. 260–264.

[13] M. Le, A. Vyas *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," *Advances in neural information processing systems*, vol. 36, 2024.

[14] S. Shahnawazuddin *et al.*, "Voice Conversion Based Data Augmentation to Improve Children's Speech Recognition in Limited Data Scenario," in *Proc. Interspeech 2020*, 2020, pp. 4382–4386.

[15] A. Mohamed, H.-y. Lee *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[16] R. Fan, Y. Zhu, J. Wang, and A. Alwan, "Towards better domain adaptation for self-supervised models: A case study of child asr," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1242–1252, 2022.

[17] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer speech & language*, vol. 63, p. 101077, 2020.

[18] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of Whisper models to child speech recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 5242–5246.

[19] A. A. Attia *et al.*, "Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults," *arXiv preprint arXiv:2309.07927*, 2023.

[20] 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.

[21] K. Shobaki, J.-P. Hosom, and R. Cole, "The ogi kids' speech corpus and recognizers," *Proc. of ICSLP*, pp. 564–567, 2000.

[22] W. Ward, R. Cole *et al.*, "My science tutor: A conversational multimedia virtual tutor for elementary school science," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 4, pp. 1–29, 2011.

[23] R. Fan and A. Alwan, "DRAFT: A novel framework to reduce domain shifting in self-supervised learning and its application to children's ASR," in *Interspeech 2022*, 2022, pp. 4900–4904.

[24] D. Rekesh, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, K. Puvvada, A. Kumar, J. Balam *et al.*, "Fast conformer with linearly scalable attention for efficient speech recognition," *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2023.

[25] R. Patel, C. Niziolek, K. Reilly, and F. H. Guenther, "Prosodic adaptations to pitch perturbation in running speech," *Journal of speech, language, and hearing research : JSLHR*, 2011.

[26] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, p. 21, 2013.

[27] D. S. Park, W. Chan *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*. ISCA, 2019, pp. 2613–2617.

[28] W. Liu, Y. Qin, Z. Peng, and T. Lee, "Sparsely shared lora on whisper for child speech recognition," *IEEE ICASSP 2024*, 2024.

[29] E. J. Hu, P. Wallis *et al.*, "Lora: Low-rank adaptation of large language models," *International Conference on Learning Representations*, 2021.

[30] N. Houlsby, A. Giurgiu *et al.*, "Parameter-efficient transfer learning for nlp," *International Conference on Machine Learning*, pp. 2790–2799, 2019.

[31] X. Liu, K. Ji *et al.*, "P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 61–68, 2022.

[32] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 4582–4597, 2021.

[33] R. Fan, A. Afshan, and A. Alwan, "Bi-apc: Bidirectional autoregressive predictive coding for unsupervised pre-training and its application to children's asr," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7023–7027.

[34] V. Srivastav, S. Majumdar, N. Koluguri, A. Moumen, S. Gandhi, Hugging Face Team, Nvidia NeMo Team, and SpeechBrain Team, "Open automatic speech recognition leaderboard," https://huggingface.co/spaces/hf-audio/open_asr_leaderboard, 2023.

[35] C. Zhou *et al.*, "Lima: Less is more for alignment," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[36] G. Yeung, R. Fan, and A. Alwan, "Fundamental frequency feature normalization and data augmentation for child speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6993–6997.

[37] L. Barrault, Y.-A. Chung, M. C. Meglioli *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[38] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "Nemo: a toolkit for building ai applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.