Introduction To Partial Fine-tuning: A Comprehensive Evaluation Of End-to-end Children's Automatic Speech Recognition Adaptation

Thomas Rolland[1,2], *Alberto Abad*[1,2]

[1]INESC-ID, Lisbon, Portugal
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal
thomas.rolland@inesc-id.pt, alberto.abad@inesc-id.pt

Abstract

Automatic Speech Recognition (ASR) encounters unique challenges when dealing with children's speech, mainly due to the scarcity of available data. Training large ASR models with constrained data presents a significant challenge. To address this, fine-tuning strategy is frequently employed. However, fine-tuning an entire large pre-trained model with limited children's speech data may overfit leading to decreased performance. This study offers a granular evaluation of children's ASR fine-tuning, departing from conventional whole-network tunning. We present a partial fine-tuning approach spotlighting the importance of the Encoder and Feedforward Neural Network modules in Transformer-based models. Remarkably, this method surpasses the efficacy of whole-model fine-tuning, with a relative word error rate improvement of 9% when dealing with limited data. Our findings highlight the critical role of partial fine-tuning in advancing children's ASR model development.

Index Terms: speech recognition, children speech, transfer learning, over-parameterisation

1. Introduction

The success of end-to-end Automatic Speech Recognition (ASR), particularly Transformer-based architectures, has led to state-of-the-art performances across various adult speech datasets [1, 2]. This paradigm integrates all components of the traditional Hidden Markov Model (HMM)-based ASR pipeline into a unified neural network, encompassing language, pronunciation, and acoustic models. However, this unification of components results in a substantial increase in the number of parameters, posing significant challenges when training end-to-end ASR models from scratch with small-sized speech datasets like children's speech [3, 4]. Indeed, children's ASR typically underperforms adult ASR due to the high intra- and inter-speaker acoustic variability present in children's speech. This variability is primarily caused by the developmental changes in their speech production apparatus [5, 6]. These changes encompass shifts in fundamental and formant frequencies, as well as alterations in temporal and spectral characteristics [5]. Additionally, the limited linguistic and phonetic knowledge of children presents another challenge for ASR systems [7]. In response to these challenges, researchers have explored various techniques, including pitch-normalised features [8], Vocal Tract Length Normalisation (VTLN) [9], multi-task learning [10], and synthetic data augmentation [11].

In this context, Transfer Learning (TL) has emerged as a promising strategy for addressing the challenges of children's ASR. Typically, TL involves adapting a pre-trained adult ASR model to children's speech through a re-training phase, leveraging knowledge acquired during pre-training. This fine-tuning approach eliminates the need to train an ASR model from scratch, making it more manageable for limited-size children's speech datasets. Notably, TL has shown promising performances across both traditional HMM-based [12] and end-to-end paradigms [13, 4, 3].

However, with the recent increase in model size, exemplified by models like Whisper [14] and HuBERT [15], which can encompass billions of parameters, it is crucial to assess TL performance when a large-size pre-trained model is used for fine-tuning with limited data. Indeed, research has highlighted the interconnected nature of the training dataset size and the number of tunable parameters, identifying them as mutual bottlenecks that influence model performances [16]. Therefore, using TL on such large models on relatively small datasets may lead to overfitting and recognition performance drop. Additionally, the issue of over-parameterisation in large Transformer-based models further complicates the fine-tuning of large ASR pre-trained models. Over-parameterisation was initially recognised in the context of Natural Language Processing (NLP) models, where certain components or layers of the architecture can be removed without compromising performance or even resulting in slight gains [17, 18]. Therefore, it is essential to investigate which components are important to fine-tune, avoiding training unimportant parameters, which can be problematic, especially for small dataset tasks such as children's speech. While these ablation studies have been explored in other domains [19, 20], their application in speech tasks remains relatively unexplored [21]. However, recent success of distillation and pruning techniques for speech models also suggest the presence of over-parameterisation in ASR models [22, 23].

While previous work using TL solely focused on fine-tuning the entire model [13, 4, 3], our work investigates a granular fine-tuning process, which we refer to as partial fine-tunning. With this partial fine-tuning, we aim to pinpoint the specific components of Transformer-based end-to-end models that yield the most significant contribution for adapting to children's speech during TL. Through our experimentation, we provide valuable insights and recommendations that can guide future children's ASR developments. While such investigation has been addressed for children's ASR in HMM-based models in the past [12], to the best of our knowledge, it has never been explored for end-to-end ASR.

2. Partial Fine-tuning

In this work, our objective is to conduct a comprehensive exploration of TL, specifically on end-to-end children's ASR. Previous research in this field has solely focused on HMM-DNN models [12], providing valuable recommendations for improving children's ASR systems. Nevertheless, such a study is cur-

Figure 1: *Overview of a Transformer and Conformer module.*

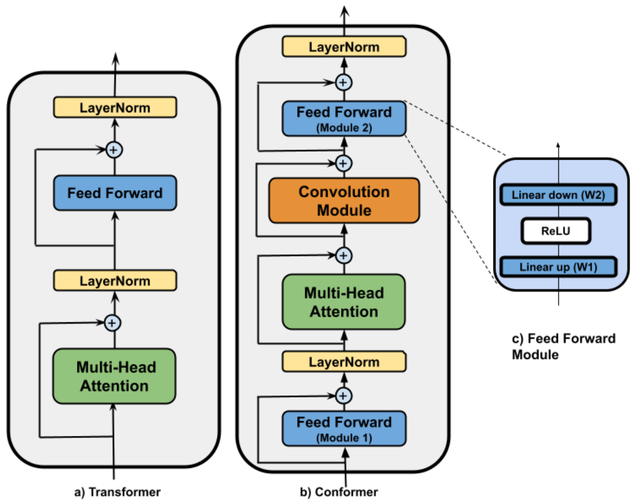


Table 1: *My Science Tutor Children Speech Corpus statistics.*

Training	Validation	Test
60897 utterances 566 speakers 113 hours	10044 utterances 79 speakers 18 hours	4079 utterances 91 speakers 13 hours

Table 2: *Fine-tuning the Encoder and Decoder separately for the Transformer and Conformer architectures.*

Transformer	WER ↓	# Params
Full model	12.99%	71.5M
Encoder only	**12.55%**	37.8M
Decoder only	15.95%	25.2M
Conformer		
Full model	12.28%	109M
Encoder only	**11.24%**	75.9M
Decoder only	16.94%	25.2M

rently missing for the end-to-end paradigm, motivating our investigation. Notably, existing works on TL in the end-to-end paradigm have solely focused on fine-tuning the entire model, leaving a notable gap in understanding the impact of fine-tuning individual components. To address this gap, we evaluate fine-tuning of both Transformer [24] and Conformer [2] architectures, known for their remarkable results in children's speech recognition [13, 4, 3]. Figure 1 presents an overview of a single layer of both the Transformer and Conformer architecture.

We start by undertaking a meticulous examination of the TL process, focusing specifically on isolating the effects of fine-tuning the Encoder and Decoder individually, as opposed to fine-tuning the entire model in Transformer-based architectures. Furthermore, a prevailing hypothesis posits that the Encoder predominantly captures acoustic information, whereas the Decoder encodes more linguistic information. Given the significant presence of acoustic variability in children's speech [6, 12], our investigation extends to discerning which layers of the Encoder are most pertinent for achieving effective TL and determining the optimal number of layers that need to be fine-tuned.

Subsequently, we introduce the *Partial fine-tuning*, where our focus shifts to delineating the distinctive contributions of modules within both the Transformer and Conformer architectures during the fine-tuning process. In contrast to previous research, which typically involves fine-tuning the entire model or entire layers, our partial fine-tuning takes a new approach by evaluating the individual contributions of Transformer-based components independently, regardless of the layers in which they are embedded. Therefore, our objective with this approach is to require fewer trainable parameters while maintaining the model's effectiveness. Indeed, by departing from the conventional whole-model or whole-layer fine-tuning, our approach aims to optimise the utilisation of limited data resources, contributing to the development of more efficient and tailored end-to-end models for children's ASR.

3. Experimental setup

3.1. Corpus

In our experiments, we used the MyST Children Speech Corpus (MyST) dataset. MyST is one of the largest publicly accessible collections of English children's speech, comprising approximately 400 hours. It encompasses dialogues between children and a virtual tutor across eight scientific domains, involving 1,372 students in grades three to five. The corpus is pre-partitioned, ensuring equitable representation of scientific domains and unique student occurrences within each partition. However, only 45% of utterances are transcribed at the word level. In our experiments, utterances shorter than one second, mainly containing silence, and those longer than 20 seconds, due to GPU constraints, were excluded. Following this filtering, 75,020 utterances from 736 speakers, totalling 144 hours, remain. More detailed statistics are provided in Table 1.

3.2. Implementation details

All experiments were conducted using the SpeechBrain toolkit [25]. The Transformer model encompasses 12 Transformer layers in the Encoder and 6 Transformer layers in the Decoder. Similarly, the Conformer architecture featured 12 Conformer layers in the Encoder and 6 Transformer layers in the Decoder. Both configurations used a hidden dimension of 512, 8 heads for all Multi-Head Attention (MHA), a Feed-Forward Network (FFN) hidden dimension of 2048, and dropout rate of 0.1. These models were pre-trained on the LibriSpeech dataset [26] a large English adult speech corpus comprising 1,000 hours of data. For reproducibility, these pre-trained models are publicly available[1]. Furthermore, for all experiments, the same Transformer language model was employed, trained on 10 million words from LibriSpeech transcriptions. Our training involved 30 epochs with a learning rate of 8×10^{-5}. Furthermore, a combination of CTC and Sequence-to-Sequence losses was used, with respective weights of 0.3 and 0.7. Finally, all statistical tests performed in this work use SCTK, the NIST Scoring Toolkit [27], specifically Matched Pairs Sentence-Segment Word Error.

[1] https://huggingface.co/speechbrain/ASR-Transformer-Transformerlm-librispeech
https://huggingface.co/speechbrain/ASR-Conformer-Transformerlm-librispeech

4. Results

4.1. Encoder-Decoder Transfer learning

Table 2 summarises the results of the impact of isolating fine-tuning of the Encoder and Decoder components only. For the Transformer model, fine-tuning the entire model exhibits a WER of 12.99% using 71.5 million parameters. Isolating the Encoder component leads to an improved WER, with 12.55% WER with a reduced parameter count of 37.8 million. In parallel, fine-tuning only the Decoder underperformed compared to the full model and Encoder-only fine-tuning strategies by achieving a WER score of 15.95% with 25.2 million parameters.

For the Conformer model, the full model TL achieves a WER of 12.28% with 109 million parameters updated. Isolating the fine-tuning on the Encoder yields a remarkable improvement, resulting in a WER of 11.24% with 75.9 million parameters. Conversely, fine-tuning only the Decoder leads to degraded performance, with a higher WER of 16.94% using 25.2 million parameters. Notably, the Conformer architecture consistently outperforms the Transformer across all configurations, emphasising its effectiveness for speech-related tasks, specifically for children's ASR. Additionally, the results underscore the important role of the Encoder in both the Transformer and Conformer ASR models compared to the Decoder, particularly in capturing the inherent variabilities of children's acoustics.

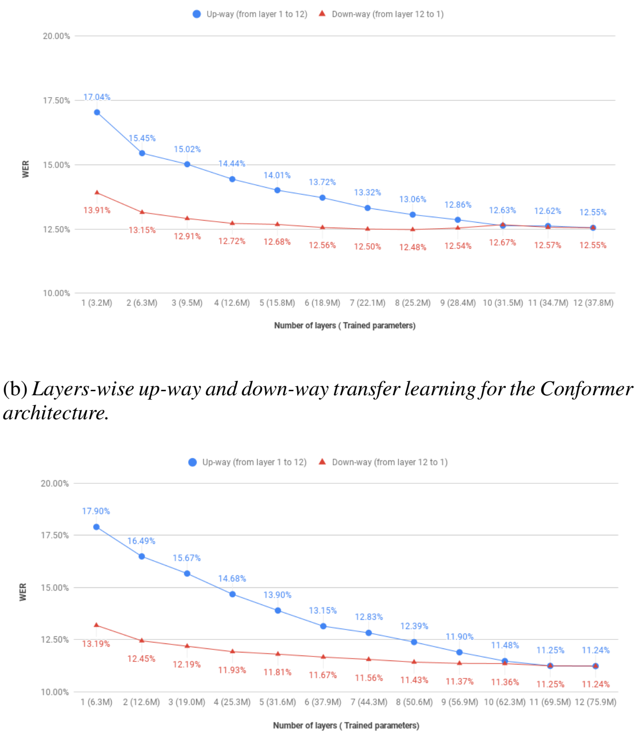
For both architectures, we performed statistical tests which showed that the *p*-value for fine-tuning the Encoder compared to the full model fine-tuning is less than or equal to 0.001. This result indicates the statistical significance and provides evidence against the null hypothesis.

4.2. Layer-wise fine-tuning

Recognising the pivotal role of the Encoder in the success of the fine-tuning process for children's ASR, our investigation delves deeper into determining the specific layers that are more important during this TL process. To achieve this, we adopted a meticulous approach where we incrementally fine-tuned the Encoder by adding one layer at a time for each fine-tuning experiment. This layer-wise fine-tuning procedure is executed bidirectionally, encompassing both an up-way trajectory, commencing from the input layer, and a down-way trajectory, starting from the output layer of the Encoder. In the up-way trajectory, fine-tuning progresses by adding one layer at a time, starting from the input layer and systematically integrating subsequent layers towards the output layer of the Encoder. Conversely, the down-way trajectory initiates fine-tuning from the output layer, systematically incorporating preceding layers towards the input layer at each experiment. The results of this layer-wise fine-tuning procedure for both the Transformer and Conformer architectures are presented in Figure 2.

In both the Transformer and Conformer, a consistent pattern emerged, adding more layers was found to be consistently beneficial, with optimal performance stabilisation occurring when 10 out of the 12 layers are employed. Interestingly, we observed that it is more advantageous to use TL from the top layers in the down-way direction (i.e., those close to the output of the Encoder) compared to the bottom layers in the up-way direction (i.e., those close to the input of the Encoder). The success of the down-way trajectory indicates that the top layers of the Encoder are the most relevant to fine-tuning for children's ASR.

Figure 2: *Transfer learning experiments increasing the number of adapted layers for Transformer and Conformer architectures.*



(a) *Layers-wise up-way and down-way transfer learning for the Transformer architecture.*

(b) *Layers-wise up-way and down-way transfer learning for the Conformer architecture.*

4.3. Modules Transfer learning

The results of our *Partial fine-tuning* experiments, focusing on fine-tuning specific components of Transformer and Conformer ASR models for children's speech, are presented in Table 3. We decomposed our experiments into three parts. Firstly, we fine-tuned components in both the Encoder and Decoder. Secondly, we partially fine-tuned only the components present in the Encoder, motivated by the results of the Encoder fine-tuning. Finally, we investigated different combinations of fine-tuning components in both the Encoder and Decoder.

4.3.1. Fine-tuning components in both Encoder-Decoder

The baseline performance of the Transformer pre-trained model without any fine-tuning yields a WER of 25.04%. In contrast, the fine-tuning of the entire Transformer model exhibits a noteworthy improvement, achieving a WER of 12.99%. The fine-tuning of the normalisation layers alone results in a WER of 17.00% by using 57.9 thousand parameters. The fine-tuning of MHA modules outperforms normalisation and full fine-tuning, achieving a WER of 12.19% with a parameter count of 25.2 million. The most important improvement was observed with the FFN modules, which attained a remarkable WER of 11.84% using 37.8 million parameters. Remarkably, both MHA and the FFN modules, when fine-tuned individually, already outperform the full model performance. This implies that the decrease in the number of parameters, coupled with the significance of

Table 3: *Partial fine-tuning experiments.*

Transformer	WER $\downarrow$	# Params
Frozen pretrained	25.04%	-
Full model	12.99%	71.5M
Normalisation	17.00%	57.9K
Attention	12.19%	25.2M
FFN	**11.84%**	37.8M
Normalisation$_{Encoder}$	17.75%	25.6K
Attention$_{Encoder}$	12.83%	12.6M
FFN$_{Encoder}$	12.20%	25.2M
Attention + FFN	12.39%	63.0M
Normalisation + FFN	12.19%	37.9M
Normalisation + Attention	12.29%	25.3M
Conformer		
Frozen pretrained	21.75%	-
Full model	12.28%	109M
Normalisation	15.61%	63.7K
Attention	11.74%	28.4M
Convolution Module	11.67%	9.7M
FFN	**11.10%**	63M
$\hookrightarrow$ Module 1	11.44%	25.2M
$\hookrightarrow$ Module 2	11.48%	25.2M
$\hookrightarrow$ Up-linear (W_1)	11.47%	31.5M
$\hookrightarrow$ Down-linear (W_2)	11.40%	31.5M
Normalisation$_{Encoder}$	15.88%	37.9K
Attention$_{Encoder}$	11.91%	15.7M
FFN$_{Encoder}$	11.17%	50.4M
FFN + Attention	11.20%	91.4M
FFN + Convolution Module	11.11%	72.7M
FFN + Normalisation	11.15%	63.1M
Attention + Normalisation	11.67%	28.4M
Attention + Convolution Module	11.44%	38.0M
Convolution Module + Normalisation	11.62%	9.7M

these modules, may play a substantial role in the enhanced performances of the fine-tuning process with a limited dataset.

Turning to the Conformer model, the baseline WER without fine-tuning results in 21.75% WER, while full fine-tuning yielded an improved WER of 12.28% with 109 million parameters. Fine-tuning specific Conformer modules offers further granularity. First, the normalisation layers fine-tuning, in a similar way as observed in the Transformer configuration, yields a score of 15.61% WER, with 63.7 thousand parameters. Then, MHA modules proved to be effective by already providing better results than the full full-tuning with a WER of 11.74%, by training 28.4 million parameters. The convolution modules outperformed the MHA with a WER of 11.67% with fewer parameters used, 9.7 million. Moreover, as in the Transformer model, FFN modules stand out significantly, demonstrating a WER of 11.10% with a parameter count of 63 million. Notably, all the MHA, convolution modules and FFN modules, when fine-tuned in isolation, surpass the performance of the full Conformer model. For both Transformer and Conformer, we performed statistical tests which showed that the p-value for fine-tuning the FFN modules compared to the full model fine-tuning is less than or equal to 0.001. This result indicates statistical significance and provides evidence against the null hypothesis.

Given the consistent importance of the FFN components in fine-tuning children's ASR, we delve deeper into the fine-tuning of the FFN modules. To this end, we identify two ways to subdivide the FFN components of a Conformer model. The first split consists of fine-tuning only Module 1 or Module 2, from Figure 1b). The second subdivision involves focusing solely on the up-linear and down-linear layers of each FFN module, denoted as W_1 and W_2. Fine-tuning Module 1 and Module 2 achieved WERs of 11.44% and 11.48%, respectively, while W_1 and W_2 yield WERs of 11.47% and 11.40%, respectively. This fine-grained analysis did not improve the fine-tuning of the full FFN modules, emphasising the importance of fine-tuning the entire FFN modules in Transformer-based end-to-end models.

4.3.2. Encoder Only

Based on the findings outlined in Section 4.1, where the significance of fine-tuning the Encoder module for children's ASR was highlighted, we investigated whether fine-tuning various components solely located in the Encoder of Transformer and Conformer layers would yield better results. It is important to note that for this experiment, we excluded the fine-tuning of Convolution modules, as they are already exclusively present in the Encoder within the Conformer architecture, in other words, $Convolutionmodules_{Encoder} = Convolutionmodules$. Our findings revealed that isolating the fine-tuning of Normalisation$_{Encoder}$, Attention$_{Encoder}$, and FFN$_{Encoder}$ modules solely within the Encoder resulted in minimal performance degradation compared to fine-tuning these modules in both the Encoder and Decoder. Nevertheless, this approach offers the advantage of reducing the number of fine-tuned parameters, while maintaining consistent performance levels.

4.3.3. Combination of different components

Furthermore, we investigate the impact of combining different components within both the Transformer and Conformer architectures. In the case of the Transformer model, the results consistently showed that, while all combinations improved upon the full mode fine-tuning, they consistently fell short of the score achieved by fine-tuning the FFN components alone. The most promising combination attained a WER score of 12.19%, by adapting normalisation layers and FFN modules in combination. Similarly, within the Conformer architecture, a comparable trend emerged. While all combinations exhibited improvements compared to the full model, they still lagged behind the performance achieved with the FFN-only scenario. A noteworthy result was the observation that the combination of the FFN and convolution modules proved to be as effective as the FFN in isolation, yielding a WER score of 11.11%. This particular experiment accentuates the notion that employing a single component is more advantageous than utilising combinations, thereby suggesting the benefits of a more parsimonious use of parameters while fine-tuning on a small dataset.

5. Conclusion

In this work, we evaluated different fine-tuning levels for children's ASR, departing from previous whole-network end-to-end fine-tuning. We found that fine-tuning the Encoder is crucial for capturing children's speech variabilities. Furthermore, when using a small dataset for TL, our partial fine-tuning identifies the adaption of the FFN modules as outperforming the entire model fine-tuning. Overall, this work showed the importance of partial fine-tuning when a small dataset is available, opening the way for better children's ASR model development. For future research directions, we aim at extending the scope of this study by evaluating the effectiveness of our proposed partial fine-tuning approach on larger models, such as Semi-Supervised Learning models [15] and Whisper [14]. Moreover, we would like to enhance the adaptability of the fine-tuning process by exploring dynamic component selection mechanisms.

6. Acknowledgements

This work was partially supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project DOI:10.54499/UIDB/50021/2020, as well as by the Portuguese Recovery and Resilience Plan (RRP) through project C644865762-00000008 (Accelerat.AI), and European Union funds through Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 766287.

7. References

[1] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, "A comparison of transformer and lstm encoder decoder models for asr," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 8–15.

[2] M. Zeineldeen, J. Xu, C. Lüscher, W. Michel, A. Gerstenberger, R. Schlüter, and H. Ney, "Conformer-based hybrid asr system for switchboard dataset," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7437–7441.

[3] P. Gurunath 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.

[4] L. Gelin, M. Daniel, J. Pinquier, and T. Pellegrini, "End-to-end acoustic modelling for phone recognition of young readers," *Speech Communication*, vol. 134, pp. 71–84, 2021.

[5] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999. [Online]. Available: https://doi.org/10.1121/1.426686

[6] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, ser. WOCCI '09. New York, NY, USA: Association for Computing Machinery, 2009. [Online]. Available: https://doi.org/10.1145/1640377.1640384

[7] A. Potamianos and S. Narayanan, "Spoken dialog systems for children," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 1, 06 1998, pp. 197 – 200 vol.1.

[8] S. Shahnawazuddin, R. Sinha, and G. Pradhan, "Pitch-normalized acoustic features for robust children's speech recognition," *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1128–1132, 2017.

[9] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition," in *SLT Workshop*, 2014, pp. 135–140.

[10] T. Rolland, A. Abad, C. Cucchiarini, and H. Strik, "Multilingual transfer learning for children automatic speech recognition," in *LREC 2022*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 7314–7320.

[11] W. Wang, Z. Zhou, Y. Lu, H. Wang, C. Du, and Y. Qian, "Towards data selection on TTS data for children's speech recognition," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6888–6892.

[12] P. Gurunath 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230820300103

[13] 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.

[14] 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*. PMLR, 2023, pp. 28 492–28 518.

[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "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.

[16] J. Kaplan, S. McCandlish, T. J. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *ArXiv*, vol. abs/2001.08361, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:210861095

[17] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, "Revealing the dark secrets of BERT," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4365–4374. [Online]. Available: https://aclanthology.org/D19-1445

[18] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" *Advances in neural information processing systems*, vol. 32, 2019.

[19] Z. Shen, Z. Liu, J. Qin, M. Savvides, and K.-T. Cheng, "Partial is better than all: revisiting fine-tuning strategy for few-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, 2021, pp. 9594–9602.

[20] P. Ye, Y. Huang, C. Tu, M. Li, T. Chen, T. He, and W. Ouyang, "Partial fine-tuning: A successor to full fine-tuning for vision transformers," *arXiv preprint arXiv:2312.15681*, 2023.

[21] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[22] S. Gandhi, P. von Platen, and A. M. Rush, "Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling," *arXiv preprint arXiv:2311.00430*, 2023.

[23] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7087–7091.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[25] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[27] National Institute of Standards and technology, "SCTK, the NIST Scoring Toolkit," 2021, accessed Febuary 11th, 2024. https://www.yale.edu/about-yale/yale-facts.