EXPLORING ADAPTERS WITH CONFORMERS FOR CHILDREN’S AUTOMATIC SPEECH RECOGNITION

Thomas Rolland, Alberto Abad

INESC-ID, Lisbon, Portugal
Instituto Superior Técnico, Universidade de Lisboa, Portugal

ABSTRACT

The high variability in acoustic, pronunciation, and linguistic characteristics of children’s speech makes it a complex task. Training a dedicated ASR model from scratch for children remains challenging, mainly due to the limited availability of children’s data. To tackle this limitation, a common strategy involves fine-tuning a pre-trained ASR model. However, this approach faces challenges due to the diversity of speakers and data scarcity, especially when dealing with large ASR models like the Conformer. In this study, we explore an alternative approach known as Adapter transfer. Adapter transfer requires training fewer parameters and can be more effective in adapting large ASR models for children’s speech. In this paper, we assess various Adapter configurations in the literature and introduce a novel configuration called Two Serial Adapter (TSA). The experimental results indicate that Adapter transfer consistently outperforms traditional fine-tuning across various configurations for the Conformer model.

Index Terms— Children’s speech, Model Adaptation, Adapters, Conformer

1. INTRODUCTION

Automatic speech recognition (ASR) has made considerable progress thanks to advances in deep learning methods that exploit large amounts of training data. Nevertheless, automatic speech recognition for children poses a major challenge. This difficulty is mainly attributed to the significant intra- and inter-speaker acoustic variability arising from developmental changes in their speech production mechanism [1]. It manifests as shifted fundamental and formant frequencies, modified temporal and spectral characteristics [2], an increased amount of disfluencies [3] and a limited linguistic knowledge [4]. Additionally, the lack of available children’s speech data limits the capability of the acoustic model to be more robust to all those variabilities.

To tackle the challenges presented by children’s speech characteristics during model training, some efforts have focused on the feature level. These efforts encompass techniques like Vocal Tract Length Normalization (VTLN) [5] and pitch and formants modifications[6]. Furthermore, recent studies have also suggested the use of adversarial multi-task learning to generate age-invariant features in a data-driven manner [7]. Additionally, at the acoustic model level, several studies applied transfer learning. This involves fine-tuning an adult model to adapt it to children’s speech characteristics [8, 9]. More recently, advances on self-supervised learning (SSL) focused on training models to extract representations from large volumes of unsupervised data. Subsequently, the entire network undergoes fine-tuning to use the extracted representations effectively in supervised tasks. Notably, these advances have led to significant improvements in child speech recognition [10, 11]. However, fine-tuning the entire network in this way can lead to significant computational costs, including long training times and substantial storage requirements. Moreover, SSL is vulnerable to domain shift, when the data domain used for fine-tuning differs from the initial pre-training domain. Although SSL yields improved performances with extensive unlabeled data during pre-training, recent research suggests that even greater improvements can be achieved by including target domain data during the pre-training phase [12]. However, larger models require more data for effective fine-tuning to prevent over-fitting.

Residual adapter modules offer an alternative approach to address traditional transfer learning limitations in the context of automatic children’s speech recognition of large ASR models. Adapters, specifically designed for Transformer-based systems integrate a compact set of additional layers into a pre-trained source model [13, 14]. These adapters typically provide enhanced computational efficiency, resulting in faster training and mitigating the problem of catastrophic forgetting. In contrast to conventional transfer learning, where the source model’s weights are completely replaced, adapter transfer preserves the backbone model, leaving it unchanged even when the adapter layers are removed. Additionally, due to their limited number of trainable parameters, adapters tend to be less prone to over-fitting.

In this study, we extend research presented in [15] and [16] by conducting a comprehensive evaluation of various adapter configurations specifically designed for Conformer models in the domain of children’s speech recognition. Furthermore, we introduce a novel adapter configuration, the Two Serial Adapter (TSA). We also perform a comparative analysis between Conformer-based models and traditional Transformer-based models, a comparison that has not been explored in the context of children’s speech recognition to the best of our knowledge. Then, considering the potential benefits of age-dependent acoustic models, and given the strong correlation between children’s age and acoustic variability [17, 18], we propose a novel cluster-based strategy for training adapters.
2. PREVIOUS WORK

2.1. Transformer and Conformer for ASR

Transformer models, which are built upon an encoder-decoder architecture and attention mechanisms, have earned considerable recognition for their effectiveness in capturing global interactions within data, especially in sequential tasks such as natural language and speech processing [19]. In parallel, Convolutional Neural Networks (CNNs) excel at capturing local features within data. Consequently, the Conformer architecture was introduced to effectively leverage the strengths of both Transformers and CNNs [20]. This is achieved by integrating CNNs into the conventional Transformer architecture. More specifically, a Conformer block consists of four modules arranged sequentially: a feed-forward module, a self-attention module, a convolution module, and a second feed-forward module. This architectural configuration empowers the Conformer to effectively capture both local and global dependencies within speech sequences. In the context of children’s speech recognition, recent research has primarily revolved around the full fine-tuning of Transformer-based models such as Whisper or SSL models [21, 10]. In contrast, Conformer models have received relatively less attention in the domain of children's ASR. However, a recent study [10] presented compelling evidence that Conformer-based models outperformed other state-of-the-art ASR models, including Wav2Vec and Hubert. This outcome underscores the potential of Conformer models to improve children’s speech recognition systems.

2.2. Adapters

Adapters were first introduced in the natural language processing field to efficiently adapt large models like Transformers for text classification [13]. They are a simple alternative to full model fine-tuning, adding a small number of parameters at each transformer layer, generally after the feed-forward layer. Adapters use a bottleneck architecture (projection-down followed by projection-up) and have benefits such as parameter efficiency, faster training, and modularity compared to full fine-tuning. Adapter can be expressed as follows:

\[ adapter(x) = x + (W_{up}(f(W_{down}g(x) + b_{down}))) + b_{up} \]  

where \( f(\cdot) \) is the non-linear activation function and \( g(\cdot) \) a layer normalization or identity function. In terms of computation, Adapters offer faster training as they update fewer parameters. However, they might introduce a slight processing delay compared to fully fine-tuned models at inference, but this difference is typically minimal and can be well-managed [22].

Adapters in the context of children’s ASR have received limited attention, with only one notable study [11]. That work proposes integrating adapters into self-supervised models and fine-tuning the entire model, including the adapter weights, to better model children’s speech. In contrast, our primary objective is to update only the adapter weights during supervised training, maintaining the parameter efficiency and modularity of adapters, without relying on semi-supervised pre-training.

3. INVESTIGATING ADAPTERS FOR CHILDREN ASR

In this paper, we investigate the application of Adapters in both Transformer and Conformer architectures. For the Transformer, we examine two integration methods: parallel and serial placement with the Feed-Forward Network (FFN) component [23]. Within the Conformer architecture, we explore six Adapter configurations, as illustrated in Figure 1. The first two configurations mirror our Transformer investigation, involving both parallel and serial placements, either after or in parallel with the second FFN layer [15]. Furthermore, we assess the configuration that incorporates an Adapter following the convolution module, referred to as the “serial-conv” setup used in [16]. Furthermore, [15] introduces two variants of the parallel setup: “parallel-conv,” where the Adapter operates in parallel with the convolution module, and “TPA,” which deploys two adapters in parallel with both FFN modules in the Conformer layer. In the case of serial configurations, we integrate the adapter information with the preceding component denoted as \( P \). The specific component \( P \) varies depending on the configuration and can be either the FFN or the convolution modules. This integration is accomplished through the following process:

\[ output = Adapter(P(x)) \]  

In the parallel configurations, the integration process differs slightly. Here, we combine the adapter’s output with the output of component \( P \) as follows:

\[ output = x + 0.5 \cdot P(x) + (Adapter(x) - x) \]

where \( x \) is the input of the component \( P \). In order to comprehensively explore all feasible configurations, we introduce a novel one called “TPA,” which stands for Two Serial Adapters. In this configuration, one Adapter is positioned sequentially after each FFN component.

Additionally, for a comprehensive assessment of Adapter behaviour, we consider three distinct configurations where Adapters are placed in the Decoder. Note that in the Conformer architecture, the decoder is a regular Transformer. Therefore, we initially evaluate the “Serial” and “Parallel” setups. Subsequently, we investigate the combination of the most effective encoder Adapter configuration with both Decoder configurations. To the best of our knowledge, there is no prior research that formally investigates the influence of Adapters within an ASR decoder.

Finally, motivated by the strong correlation between children’s speech variability and age, we investigate the possibility of training specialized Adapters for groups of speakers exhibiting similar acoustic characteristics based on unsupervised clustering. In practice, we apply a k-means clustering algorithm on the x-vector representation [24] of each training utterance. Then, a different adapter model is trained for each speaker cluster. During test, the closest speaker cluster is found for each test utterance and the corresponding Adapter weights are used for decoding. The primary objective of these experiments is to investigate whether Adapters trained on comparable speech characteristics yield improvements over a general Adapter on the entire training set.

3.1. Relation to prior work

Children’s speech is inherently atypical and displays a significant degree of variability, making it imperative to assess the efficacy of existing methods in modelling children’s speech. This paper introduces several contributions to the field of automatic children’s speech recognition. Firstly, the paper conducts a comparative analysis between traditional Transformer and Conformer-based models. Secondly, we propose the use of residual adapters to tackle the challenges associated with children’s ASR in large models. This study conducts a comprehensive and systematic evaluation of various
adapter configurations designed explicitly for Conformer models extending prior work [15, 16], pinpointing the most effective setups for children's ASR tasks. Indeed, in prior work different configurations were employed resulting in a lack of standardised evaluation. Additionally, a new adapter configuration called the Two Serial Adapter (TSA) is proposed to cover all conceivable configurations. Furthermore, to the best of our knowledge, there has been no work exploring Adapters in the decoder of an ASR model. Lastly, a cluster-based training strategy is proposed as an innovative approach to further enhance children's ASR by leveraging speaker-dependent acoustic characteristics. Altogether, these contributions advance the state of children's ASR and provide valuable insights for future research in this domain.

4. EXPERIMENTAL SETTINGS

4.1. Children speech corpus

The My Science Tutor (MyST) Children Speech Corpus [25] offers around 400 hours of English speech data. It includes conversations between students and a virtual tutor spanning 8 scientific domains and involves 1,372 students in third, fourth, and fifth grades. The corpus has been carefully partitioned to improve its usability, ensuring fair representation across domains and allocating each student’s data to a single partition. However, it is important to note that only 45% of the utterances have been transcribed at the word level. For our experiments, we applied specific filtering criteria, excluding utterances shorter than one second and longer than 20 seconds. You can find detailed information about the filtered corpora in Table 1.

<table>
<thead>
<tr>
<th>Table 1. My Science Tutor Children Speech Corpus statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
</tr>
<tr>
<td># of utterances</td>
</tr>
<tr>
<td># of speakers</td>
</tr>
<tr>
<td># of hours</td>
</tr>
</tbody>
</table>

4.2. Implementation details

All experiments were performed using the SpeechBrain toolkit [26]. We used 12 Transformer or Conformer layers for the encoder, for the Transformer and Conformer model respectively, and 6 Transformer layers for the decoder, all with dimensions 512. These models have been pre-trained using the LibriSpeech dataset [27] and are publicly available¹. Furthermore, for all of our experiments, we used the same Transformer language model, trained on 10 million words on the LibriSpeech transcriptions. The adapter architecture consists of a first linear layer projection to dimension 512 with a ReLu activation, followed by another linear layer projection to dimension 512 with a residual connection of the adapter input. The use of a hidden-dimension size equal to the model size (instead of a bottleneck) was motivated by previous research exploring hidden-dimension size, that consistently demonstrated that larger dimensions tend to yield improved performance scores [15]. All models were trained for 30 epochs, with a learning rate of 8 · 10⁻⁴ for Adapters experiments and of 8 · 10⁻⁵ for fine-tuning the entire model. The training was performed using a combination of Connectionist Temporal Classification and Sequence-to-Sequence losses, with weights of 0.3 and 0.7 respectively.

For the clustering experiments, we use the k-means clustering algorithm on the speaker-embedding of each utterance. The speaker embeddings were extracted using a publicly pre-trained ECAPA-TDNN model, trained on adult speech².

5. RESULTS

5.1. Configurations

In this section, we present a comprehensive evaluation of Adapter configurations applied to both Transformer and Conformer models, assessing their performance based on Word Error Rate (WER), as presented in Table 2. First, we assess the Transformer model when no fine-tuning was applied (Frozen), resulting in a WER of 25.04%. Conversely, Full Fine-Tuning involved complete fine-tuning of the

¹https://huggingface.co/speechbrain/asr-transformer-transformerlm-librispeech
²https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
Table 2. Results of the different Adapters configurations in both Transformer and Conformer.

<table>
<thead>
<tr>
<th>Method</th>
<th>WER</th>
<th>Trained params</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transformer</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frozen</td>
<td>25.04%</td>
<td>-</td>
</tr>
<tr>
<td>Full fine-tuning</td>
<td>12.99%</td>
<td>71.5M</td>
</tr>
<tr>
<td>Serial</td>
<td>12.78%</td>
<td>6.3M</td>
</tr>
<tr>
<td>Parallel</td>
<td>12.62%</td>
<td>6.3M</td>
</tr>
<tr>
<td><strong>Conformer</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frozen</td>
<td>21.75%</td>
<td></td>
</tr>
<tr>
<td>Full fine-tuning</td>
<td>12.28%</td>
<td>109.1M</td>
</tr>
<tr>
<td>Serial</td>
<td>11.76%</td>
<td>6.3M</td>
</tr>
<tr>
<td>Serial-Conv</td>
<td>11.78%</td>
<td>6.3M</td>
</tr>
<tr>
<td>Parallel</td>
<td>11.72%</td>
<td>6.3M</td>
</tr>
<tr>
<td>Parallel-conv</td>
<td>11.79%</td>
<td>6.3M</td>
</tr>
<tr>
<td>TPA</td>
<td><strong>11.58%</strong></td>
<td>12.6M</td>
</tr>
<tr>
<td>TSA</td>
<td>11.75%</td>
<td>12.6M</td>
</tr>
<tr>
<td>Serial (Decoder)</td>
<td>18.09%</td>
<td>3.2M</td>
</tr>
<tr>
<td>Parallel (Decoder)</td>
<td>17.76%</td>
<td>3.2M</td>
</tr>
<tr>
<td>TPA + Parallel (Decoder)</td>
<td><strong>11.47%</strong></td>
<td>15.8M</td>
</tr>
</tbody>
</table>

In addition, we evaluated the use of Adapters in the decoder. As the decoder of the Conformer architecture is a regular Transformer, we only evaluate the “Serial” and “Parallel” setup, which respectively reached 18.09% and 17.76% WER with 3.2 million parameters. Results showed that Adapters are more relevant when plugged into the encoder. It confirms that acoustic variability plays a critical role in the degradation of children’s ASR performance. Finally, combining “TPA” in the encoder layers with “Parallel” Adapters in the decoder significantly improves performance, reducing the WER to 12.99%, at the expense of 71.5 million trainable parameters. Turning to the Adapter setups, we investigate the “Serial” and “Parallel” configurations, both equipped with 6.3 million trainable parameters. The “Parallel” emerged as the best configuration, achieving the lowest WER of 12.62% compared to 12.78% for the “Serial”. These results underscore the effectiveness of Adapter configurations within the Transformer architecture, as they both perform slightly better than the full-finetuning.

Next, we investigated the Conformer model, we once again explored Frozen and Full Fine-Tuning. In Frozen the pre-trained model remained untouched, yielding a WER of 21.75%. The Full finetuning, in a similar way as the Transformer, led to enhanced performance, reducing the WER to 12.28% with a total of 109.1 million trainable parameters. We can observe that given the same pretraining dataset, the Conformer architecture outperforms the regular Transformers. Within the set of adapter configurations, “Serial” achieved a WER of 11.76%, while “Parallel” demonstrated slightly better performance with a WER of 11.72%. These results indicate that “Parallel” Adapters were more effective in improving WER in the Conformer model. When Adapters are placed after the convolution layer, the “Serial-conv” and “Parallel-conv” configuration, both slightly under-perform compared to Adapters placed after the second FFN component with respective scores of 11.78% and 11.79%. Finally, we evaluated the “TPA” and “TSA” configurations. The “TPA” configuration emerged as the most promising, with a very remarkable WER of 11.58% using 12.6 million trainable parameters, while “TSA” achieved a WER of 11.75%, which is slightly underperforming compared to the “TPA” configuration.

In addition, we evaluated the use of Adapters in the decoder. As the decoder of the Conformer architecture is a regular Transformer, we only evaluate the “Serial” and “Parallel” setup, which respectively reached 18.09% and 17.76% WER with 3.2 million parameters. Results showed that Adapters are more relevant when plugged into the encoder. It confirms that acoustic variability plays a critical role in the degradation of children’s ASR performance. Finally, combining “TPA” in the encoder layers with “Parallel” Adapters in the decoder significantly improves performance, reducing the WER to 12.99%, at the expense of 71.5 million trainable parameters. We can observe that given the same pretraining dataset, the Conformer architecture outperforms the regular Transformers. Within the set of adapter configurations, “Serial” achieved a WER of 11.76%, while “Parallel” demonstrated slightly better performance with a WER of 11.72%. These results indicate that “Parallel” Adapters were more effective in improving WER in the Conformer model. When Adapters are placed after the convolution layer, the “Serial-conv” and “Parallel-conv” configuration, both slightly under-perform compared to Adapters placed after the second FFN component with respective scores of 11.78% and 11.79%. Finally, we evaluated the “TPA” and “TSA” configurations. The “TPA” configuration emerged as the most promising, with a very remarkable WER of 11.58% using 12.6 million trainable parameters, while “TSA” achieved a WER of 11.75%, which is slightly under-performing compared to the “TPA” configuration.

In addition, we evaluated the use of Adapters in the decoder. As the decoder of the Conformer architecture is a regular Transformer, we only evaluate the “Serial” and “Parallel” setup, which respectively reached 18.09% and 17.76% WER with 3.2 million parameters. Results showed that Adapters are more relevant when plugged into the encoder. It confirms that acoustic variability plays a critical role in the degradation of children’s ASR performance. Finally, combining “TPA” in the encoder layers with “Parallel” Adapters in the decoder significantly improves performance, reducing the WER to 12.99%, at the expense of 71.5 million trainable parameters. We can observe that given the same pretraining dataset, the Conformer architecture outperforms the regular Transformers. Within the set of adapter configurations, “Serial” achieved a WER of 11.76%, while “Parallel” demonstrated slightly better performance with a WER of 11.72%. These results indicate that “Parallel” Adapters were more effective in improving WER in the Conformer model. When Adapters are placed after the convolution layer, the “Serial-conv” and “Parallel-conv” configuration, both slightly under-perform compared to Adapters placed after the second FFN component with respective scores of 11.78% and 11.79%. Finally, we evaluated the “TPA” and “TSA” configurations. The “TPA” configuration emerged as the most promising, with a very remarkable WER of 11.58% using 12.6 million trainable parameters, while “TSA” achieved a WER of 11.75%, which is slightly under-performing compared to the “TPA” configuration.

In addition, we evaluated the use of Adapters in the decoder. As the decoder of the Conformer architecture is a regular Transformer, we only evaluate the “Serial” and “Parallel” setup, which respectively reached 18.09% and 17.76% WER with 3.2 million parameters. Results showed that Adapters are more relevant when plugged into the encoder. It confirms that acoustic variability plays a critical role in the degradation of children’s ASR performance. Finally, combining “TPA” in the encoder layers with “Parallel” Adapters in the decoder significantly improves performance, reducing the WER to 12.99%, at the expense of 71.5 million trainable parameters. We can observe that given the same pretraining dataset, the Conformer architecture outperforms the regular Transformers. Within the set of adapter configurations, “Serial” achieved a WER of 11.76%, while “Parallel” demonstrated slightly better performance with a WER of 11.72%. These results indicate that “Parallel” Adapters were more effective in improving WER in the Conformer model. When Adapters are placed after the convolution layer, the “Serial-conv” and “Parallel-conv” configuration, both slightly under-perform compared to Adapters placed after the second FFN component with respective scores of 11.78% and 11.79%. Finally, we evaluated the “TPA” and “TSA” configurations. The “TPA” configuration emerged as the most promising, with a very remarkable WER of 11.58% using 12.6 million trainable parameters, while “TSA” achieved a WER of 11.75%, which is slightly under-performing compared to the “TPA” configuration.

5.2. Unsupervised Clustering of utterances

In this section, we present the outcomes of our clustering approach, as summarised in Table 3. We investigated the influence of varying cluster numbers on ASR performance, ranging from 1 to 4 clusters using the “TPA” in the encoder-only configuration. Notably, when the data remained unclustered (one cluster), the ASR system exhibited a WER of 11.58%. However, the two-cluster configuration surpassed the others, achieving a superior performance of 11.50%. This result suggests that partitioning the data into two distinct clusters enables the Adapters to more effectively capture underlying patterns intricately linked to their respective clusters, consequently enhancing the recognition scores. Furthermore, we explored the impact of increasing the number of clusters to three and four, revealing only marginal differences in performance, with WERs of 11.57% and 11.51%, respectively. These findings underscore the role of data clustering in children’s ASR systems.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we conducted a comprehensive evaluation of Adapter performance within the context of children’s ASR, considering both Transformer and Conformer architectures. Our findings consistently favoured the “Parallel” Adapter configuration over the “Serial” Adapter configuration, whether in the encoder or decoder. The most promising configuration entailed employing two parallel adapters in the encoder, in conjunction with parallel adapters in the decoder, yielding the best results. In future works, we aim to extend our evaluation of Adapters to include other SSL or large ASR models like HuBERT and Whisper. Additionally, we intend to explore novel Adapter architectures tailored to specifically fit children’s speech characteristics.

3https://github.com/usnistgov/SCTK/tree/master
7. REFERENCES


