End-to-end Deep Learning For Child Speech Recognition

Anthony Li
Stanford University
antli@stanford.edu

Anthony Galczak
Stanford University
agalczak@stanford.edu

Abstract

Child Speech Recognition (CSR) is a less explored and more challenging task than typical Automatic Speech Recognition (ASR). This task has significant applications in the classroom and is especially important in a remote learning environment. We present findings from training deep-learning based speech recognition models on the MyST corpus, the largest publicly-available English-language child speech corpus. We obtained 27.26% word error rate (WER) on the MyST test set with a DeepSpeech2 baseline. Our best model, a Conformer model pre-trained on LibriSpeech and fine-tuned using the MyST corpus, achieved a test WER of 23.45%. Our results show that pre-training on adult speech is essential for model performance. We also provide additional error analysis on our best model and discussion of the results.

1 Introduction

Automatic speech recognition (ASR) systems have become ubiquitous in recent years, powering an abundance of consumer products such as smartphone digital assistants, smart speakers, and more. It is no surprise, therefore, that there has been substantial investment in developing robust and highly accurate ASR systems. Commercial ASR systems are able to achieve strong performance on adult speech, often with WER of under 5% (Booth et al., 2020).

However, there has been much less focus on developing similarly accurate systems for child speech. Child speech recognition (CSR) is challenging because speech characteristics are very different between adults and children. Acoustic properties have higher variation in child speech than in adult speech because childrens' vocal tracts change rapidly as they mature (Booth et al., 2020). Further compounding the problem, there has historically been a lack of large, freely available child speech datasets for the research community, which may in part be due to privacy laws around data collection from minors. As a result, most earlier CSR systems have been trained on fairly small datasets, making it difficult for them to achieve high performance.

Recently however, the My Science Tutor (MyST) Children's speech corpus from Boulder Learning has been made available for research and commercial use. The corpus contains nearly 500 hours of speech from approximately 1300 students in grades 3-5, who engaged in dialogs with a virtual science tutor. About 45% of the utterances in the corpus have word-level transcriptions (bou). The availability of this large corpus presents an opportunity to build more robust English-language CSR systems.

Recent model architectures such as Conformer (Gulati et al., 2020) and ContextNet (Han et al., 2020) have shown strong performance on LibriSpeech, but there has not been substantial work evaluating these models on child speech data. We present results from training and evaluating these models on the MyST corpus.

2 Related Works

Wu (Wu, Fei, 2020) studied child speech recognition (CSR) as a low-resource automatic speech recognition (ASR) task, training on the CMU Kids and CSLU Kids datasets which contain about 9.1 hours and 69.3 hours of speech respectively. The paper explored hybrid DNN-HMM systems as well as end-to-end deep learning approaches. The primary hybrid system studied was a TDNN-F model, a variation on time-delay neural networks (TDNN) for phoneme recognition which was hypothesized to improve ASR performance on low-resource tasks. The primary end-to-end DNN system was a seq2seq with attention model fine-tuned on the LibriSpeech and WSJ corpora, with different components frozen during fine-tuning.

The hybrid systems trained on the CMU Kids corpus had very high test set WERs (>70%), but they were able to achieve 22.3% test set WER when trained on CSLU Kids corpus. End-to-end deep learning models pre-trained on LibriSpeech and fine-tuned were able to achieve test WER of 24.2%.

Booth et al. (Booth et al., 2020) used DeepSpeech2 to train and test against the CMU Kids dataset. This study used a model trained on LibriSpeech as a baseline model and then fine-tuned using the CMU Kids training data. Additionally, a custom-built child speech data collection tool was used to collect an additional 454 utterances. Their final transfer learning model obtained a WER of 29%, which is comparable to our best results from DeepSpeech2/LibriSpeech transfer learning on MyST. Additional analysis included grouping by grade level and experimenting with models trained on data from only one grade—for example, testing on the lowest performing grade (1st) with just 1st grade training data. This actually resulted in a worse WER (42%) than using the all age groups model (39.4%).

Shivakumar et al. (Shivakumar and Narayanan, 2021) is the only work we are aware of that has directly investigated the performance of CSR systems trained on the MyST corpus. It studied a variety of recent neural network architectures pre-trained on LibriSpeech and LibriVox and then fine-tuned on child speech corpora. On the MyST test set, the best result recorded was 16.01% WER for a Transformer+CTC model fine-tuned on MyST. Interestingly, this result was achieved using greedy decoding rather than beam search decoding, though beam search decoding still performed better in out-of-domain evaluation.

3 Approach

After initial dataset exploration, we began by training baseline models using DeepSpeech2 (Amodei et al., 2015). Released in 2015, DeepSpeech2 is able to achieve solid performance of approximately 5.33% WER on LibriSpeech test-clean. However, it is surpassed by more recent models such as ContextNet and Conformer, which are able to achieve approximately 2.3% and 2% WER respectively.

Because previous work (Booth et al., 2020) has evaluated DeepSpeech2 using the CMU Kids dataset, we felt it is a natural choice of baseline model. We evaluated a DeepSpeech2 model pre-trained on LibriSpeech but with no additional training using the MyST corpus. We then continued training the pre-trained model using the MyST training set, performing several training runs for hyperparameter tuning. Evaluation was performed using both greedy decoding and beam search decoding with a LibriSpeech 3-gram language model. We trained DeepSpeech2 using an open source framework, *SeanNaren/deepspeech.pytorch* (Naran, 2021).

Next, we looked to improve upon the best results from DeepSpeech2 by training additional models. We trained several Conformer models using different hyperparameter settings and text featurization methods, one of which was trained from scratch without LibriSpeech pre-training. We also trained a ContextNet model from scratch. These models were trained using the open source *TensorFlowASR* framework (Nguyen, 2021).

We did not add special handling for non-word tokens like <laugh>, <breath>, <noise>, etc. Instead, we allowed models to treat them as regular words and learn to predict them organically. We do believe that removing these tokens from both the training and evaluation data could result in further WER improvements.

We aimed to train each model for at least 10 epochs, which was challenging due to limited compute resources. This generally took several days for each model. Models were trained using a mixture of compute resources including Google Cloud, Azure, Google Colab, and local GPUs.

4 Experiments

4.1 Exploration of MyST Corpus

Table 1 summarizes our initial exploration of the MyST corpus. Word counts were obtained using nltk's word_tokenize. We discovered some need for data cleanup—for example, some transcript files were empty or contained only <NO_SIGNAL>. Additionally, very short (<0.5s) and very long (>60s) utterance recordings were present which caused issues in model training. Before training our initial models, we removed training examples with utterance recordings shorter than 0.5s or longer than 60s, as well as examples with empty or degenerate transcripts. In total this amounted to about 3660 training examples removed. Figure 1 shows the distribution of recording lengths across the MyST corpus and Figure 2 shows the distribution of transcript lengths.

	Train	Dev	Test	All
Total files	181323	23652	22592	227567
Duration(hrs)	379	48	47	474
avg words	16.875	15.917	N/A	16.692
unique words	10191	4570	N/A	10873
transcripts	76992	12261	13180	102433
no transcripts	104331	11391	9412	125134

Table 1: MyST corpus statistics.

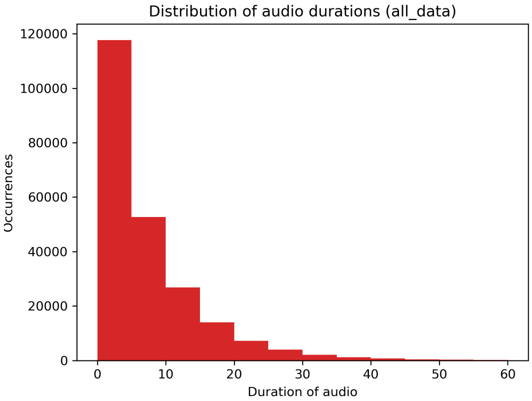

Figure 1: Distribution of utterance recording durations (seconds). A small number of very long recordings are excluded from this histogram.

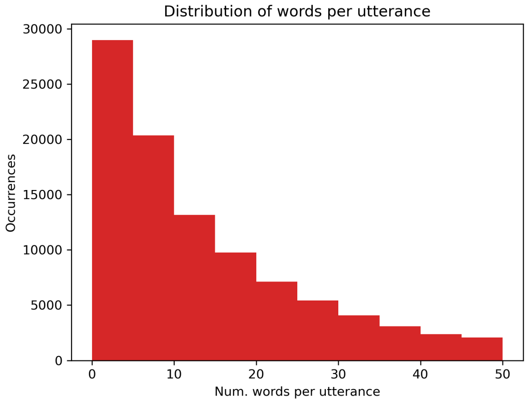

Figure 2: Distribution of transcript word counts. Some longer transcripts are excluded from this histogram.

4.2 Results

Our results are summarized in Table 4. We used word error rate (WER) as our primary evaluation metric, which is standard in the Spoken Language community. The MyST corpus comes with an official train/dev/test split. All models were trained using the MyST corpus training set, intermediate evaluations were performed using the dev set and final results were obtained using the test set. Note that the dev set was left out of training and testing for our final results.

The DeepSpeech2 model pre-trained on LibriSpeech but not further trained on MyST (*ds2_untrained*) saw poor performance on the MyST test set, achieving 58.07% and 48.61% WER from greedy and beam search decoding respectively. This was not unexpected, due to the aforementioned differences between child and adult speech. From here, we looked to improve performance of DeepSpeech2 by continuing training using the MyST training set. This resulted in a substantial reduction of WER, reaching 27.67% WER after 10 epochs of training (*ds2_ft_10*).

We additionally tested the effect of enabling SpecAugment when continuing training. While we saw mild improvements, reaching 27.26% test WER after 5 epochs (*ds2_ft_5_aug*), the model began to diverge after further training for unclear reasons (*ds2_ft_10_aug*) and produced blank predictions. Based on the performance of *ds2_ft_10* and *ds2_ft_5_aug*, it does not appear that beam search decoding improved performance for fine tuned models. This is consistent with (Shivakumar and Narayanan, 2021), which found that the model achieving best performance on the MyST test set used greedy decoding. However, beam search did improve WER by over 10 percentage points for the DeepSpeech2 model trained from scratch on MyST (*ds2_scratch_14*).

Since it did not appear that further tuning of

	Audio Duration			Number of Words		
	0.5s-3s	3s-8s	8s+	1-5	6-15	16+
WER	34.01	25.95	22.40	29.28	25.21	22.48
# samples	4073	4254	4542	4223	4143	4631

Table 2: WER of *conf_ft_subword_10* model evaluated against utterances of various lengths.

ID	Ground Truth	Prediction
1	humus gravel clay and silt	humus gravel clay and silt
2	sixty seven grams of um sixty seven grams is level with the solution	(()) seven grams of (()) seven grams is level with the solution
3	good <breath>	good <laugh>
4	<silence>	<silence>
5	ricky john jones ricky john jones ricky john john john john john	break down the break down spot and jump
6	i'm doing good how about you	i'm doing good how about you
7	you could um feel which one was heavier than the other like um the metal was isn't hollow like the wood the plastic was not hollow but it was lighter than the wood by its um by how small it was and um th s cause it was about um oh it was really little	you could um see which one was heavier than the other like um the metal was isn't ho like the wood the plastic was not but it was lighter than the wood by um by how small it was and um cause it was about um well it was really little

Table 3: A sampling of test-time utterance predictions from the *conf_ft_subword_10* model.

DeepSpeech2 would significantly improve performance, we opted to shift to newer model architectures. These include Conformer and ContextNet, which have open source implementations in the *TensorFlowASR* library. Similar to our DeepSpeech2 approach, we initially evaluated a Conformer model pretrained on LibriSpeech but not further trained on MyST (*conf_untrained*). This still performed poorly overall, but achieved a significantly better result than the equivalent DeepSpeech2 model (41.90% WER for *conf_untrained* vs. 58.07% WER for *ds2_untrained*).

Next, we continued training the Conformer model pre-trained on LibriSpeech using the MyST training data. This took considerably more time than DeepSpeech2, requiring approximately six days to reach 10 epochs training on a single K80. This model (*conf_ft_subword_10*) achieved a WER of 23.59%. We continued training this model to 14 epochs (*conf_ft_subword_14*) but we did not see improved performance. However it is difficult to determine if this model has fully converged as the evaluation loss curve may still be slowly decreasing, as shown in Figure 3. The evaluation loss curve also has a peculiar spike at the beginning, which could indicate non-optimal hyperparameter settings. We also tested training Conformer with a higher number of warmup steps (80000) and gradient accumulation steps (8). This produced a slightly better result of 23.45% test WER (*conf_ft_subword_10_v2*).

The Conformer model trained from scratch without LibriSpeech pre-training (*conf_char_10*) did not converge to a usable degree. When evaluated, it merely produced a transcription of "the" for every utterance, resulting in a very high WER. This model likely needs different hyperparameters and/or more training time.

We also trained ContextNet and Jasper using the *TensorFlowASR* framework, but we did not have pre-trained LibriSpeech checkpoints for these models. As such, we trained these models from scratch. ContextNet after 10 epochs produced intelligible transcriptions but obtained a poor evaluation result of 75.21% WER. From these results, it is clear that pre-training on a large adult speech corpus like LibriSpeech is still very useful, even if the model will be fine-tuned and evaluated on child speech. All of the models that were not pre-trained on LibriSpeech did not seem to converge to a useful degree and exhibited poor performance.

We also tried training a Listen-Attend-Spell

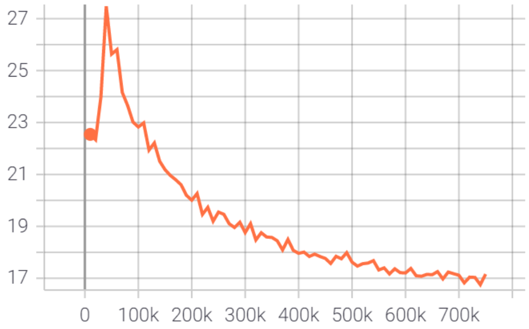

Figure 3: Evaluation loss curve (transducer loss) for *conf_ft_subword_10*.

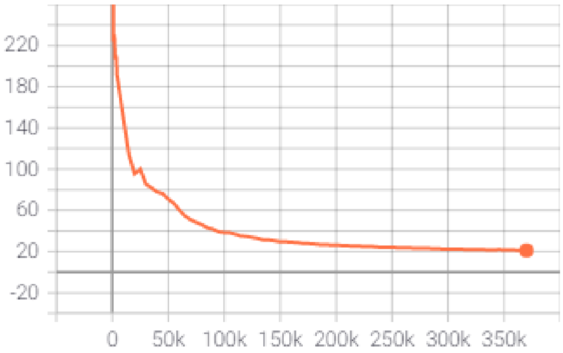

Figure 4: Evaluation loss curve (transducer loss) for *context_char_10*.

(LAS) model (Chan et al., 2015), but unfortunately after one epoch of training our loss curves were not decreasing and it didn't appear the model was learning. At evaluation time, it produced only transcriptions of "THE SAE" and "THE ST" for every utterance. We also trained the Jasper implementation in the *TensorFlowASR* framework for 10 epochs, but the model returned nonsensical transcriptions. These transcriptions consisted of random strings and characters and contained an excessive amount of spaces.

4.3 Error analysis and Discussion

In Table 3, we present some example ground truth utterances and predictions from our *conf_ft_subword_10* model. This model produced an overall test WER of 23%, but a qualitative analysis indicates that the transcriptions may be of better quality than the quantitative evaluation suggests because many errors come from ambiguity in the original audio. It performed especially well for in-domain utterances, like those dealing with geology and chemistry terms. As an example, utterance 1 shows that the model can reliably pick up words that are somewhat rare in everyday conversation.

Utterance 2 shows an example of a stray special token appearing in the prediction. Upon listening to this utterance, the pronunciation of "sixty" is fairly clear, but our model incorrectly identified it as "(())". This "(())" token in the ground truth training data seems to represent sounds that are difficult to transcribe. Anecdotally, we noticed that "(())" was often predicted for numbers in the audio, such as "fifty" and "sixty." This suggests that our model may not be confident predicting specific numbers. We hypothesize that this occurs because any specific number is unlikely to appear frequently in the training data, so additional training data may be needed specifically to improve number recognition.

Utterances 3 and 4 illustrate the model's ability to predict special tokens such as <laugh> or <noise>. Although we did not add any special handling for these tokens, the model was often able to insert them correctly into the predictions, albeit with lower accuracy than regular words. Special tokens may not be present consistently in the ground truth data, so it is not surprising that ASR models may also find it difficult to insert them correctly.

Utterance 5 shows an example of out-of-domain speech. The audio reveals that this utterance is spoken very quickly and in a manner that is quite different from that of regular speech. This shows that in some cases the model doesn't generalize very well outside of the fine-tuned domain. However, the model can still perform well on common phrases and similar utterances that do not strictly relate to science topics, such as Utterance 6.

Utterance 7 is an example of a long utterance the model performed fairly well on. The model missed some words such as "hollow," but upon listening to the transcription, the child's pronunciation of "hollow" was poor and it would be very difficult to predict the first "hollow" without context. Therefore the model's prediction of "ho" could be interpreted as an accurate transcription.

4.3.1 WER vs. Transcript and Audio Length

We performed an analysis of the relationship between model performance and audio/transcript length. Results are shown in Table 2. From this analysis, it is clear that shorter audio correlates to poor model transcriptions. We hypothesize there are two separate reasons for this correlation. First, shorter utterances have a disproportionate quantity of special tokens, like <laugh>, and tend to be noisier samples. In contrast, longer samples often

Model	Short name	Approx. Epochs Trained	Augmentation	Greedy decoder WER	Beam search WER
DeepSpeech2 (LibriSpeech)	ds2_untrained	0	–	58.07	48.61
DeepSpeech2 (LibriSpeech+fine tuning)	ds2_ft_10	10	None	27.67	30.63
DeepSpeech2 (LibriSpeech+fine tuning)	ds2_ft_5_aug	5	SpecAugment	27.26	31.24
DeepSpeech2 (LibriSpeech+fine tuning)	ds2_ft_10_aug	10*	SpecAugment	Diverged	Diverged
DeepSpeech2 (Trained from scratch)	ds2_scratch_14	14	None	37.41	28.77
Conformer (LibriSpeech)	conf_untrained	0	–	41.90	–
Conformer (Trained from scratch, character-level vocab)	conf_char_10	10	SpecAugment	96.93	–
Conformer (LibriSpeech+fine tuning, subword vocab)	conf_ft_subword_10	10	SpecAugment	23.59	–
Conformer (LibriSpeech+fine tuning, subword vocab)	conf_ft_subword_14	14*	SpecAugment	23.78	–
Conformer (LibriSpeech+fine tuning, subword vocab, increased warmup steps and grad accumulation)	conf_ft_subword_10_v2	10	SpecAugment	23.45	–
ContextNet (Trained from scratch, character-level vocab)	context_char_10	10	SpecAugment	75.21	–

Table 4: Summary of model performance for all models evauated. (*) indicates training was continued from model in previous row.

discuss a specific topic at length and use a rich in-domain vocabulary. Secondly, the Conformer model uses attention which can capture contextual information between sub-words. In very short utterances (0-5 words), there is likely less useful context available to the model.

5 Conclusion

Over the course of this project, we investigated several SOTA model architectures for ASR applied to the task of child speech recognition, many of which have not been previously evaluated on the MyST corpus. Our best model, *conf_ft_subword_10*, obtained a solid 23% WER and produced good-quality transcriptions of child speech. Based on qualitative analysis of these transcriptions, the vast majority of them convey the entire idea of the utterance and would be useful in real applications. We also demonstrated the importance of pre-training using LibriSpeech, which was essential to achieving strong performance with relatively few training epochs on MyST.

5.1 Challenges

Long training times made it difficult to experiment with hyperparameter tuning. Given more time, we would be able to test a greater variety of different model sizes, augmentation techniques, and many other hyperparameters for our Conformer model. We used thousands of hours of GPU time during this project across a combination of local GPU's, Google Colab Pro, and Google Cloud credits. Getting access to on-demand compute for long-running training was an issue for our project towards the end.

In our model evaluation, we also wanted to evaluate performance by age group, similar to the analysis performed in (Booth et al., 2020). However, we aren't aware of grade-level metadata in the MyST corpus and would likely need to utilize an external dataset. We would also like to evaluate how well our model generalizes to other child and adult speech datasets.

We intended to perform hyperparameter tuning using the dev set, but ultimately we did not significantly make use of dev evaluation results for hyperparameter tuning due to the long training times. We likely could have improved performance slightly by including the dev set in our final training runs.

We attempted to write a custom implementation of Conformer, but there were a lot of technical challenges, including the amount of time it would take to pretrain on LibriSpeech, that prevented us from continuing this effort. However, a custom implementation would provide more control over how we use the model for the project.

5.2 Future Research Directions

We performed a lengthy model architecture search to identify promising model architectures for our child speech recognition task. Continuing this search and experimenting with other models, such as QuartzNet and Wav2Vec, could identify other promising architectures that yield better empirical results.

The bulk of our non-baseline models didn't show convergence. Therefore, it would be reasonable to continue training all models tested until convergence. In the case of the larger models without pre-training, we may need to train for 2+ weeks. *conf_ft_subword_10*, one of our most performant, is one example of a model that may benefit from longer training.

An interesting research topic beyond the scope of this paper would be to design a model architecture that focused on CSR, rather than adult ASR.

5.3 Acknowledgements

We would like to thank Prof. Andrew Maas and the teaching staff of CS224S for their guidance and support throughout this project. We would also like to thank Dr. Ron Cole and Dr. Sameer Pradhan for providing us access to the MyST corpus.

5.4 Contributions

Anthony Li worked on infrastructure setup and training of Conformer, Jasper, and DeepSpeech2 without SpecAugment. Anthony Galczak worked on dataset statistics and training of ContextNet and DeepSpeech2 with SpecAugment. We shared responsibility for data preprocessing and model evaluation.

References

My science tutor (myst) children's speech corpus.

Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse H. Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Y. Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Y. Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. 2015. Deep speech 2: End-to-end speech recognition in english and mandarin. *CoRR*, abs/1512.02595.

Eric Booth, Jake Carns, Casey Kennington, and Nader Rafla. 2020. Evaluating and improving child-directed automatic speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6340–6345, Marseille, France. European Language Resources Association.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2015. Listen, attend and spell.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition.

Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. 2020. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context.

Sean Naran. 2021. Deepspeech2 for pytorch.

Huy Le Nguyen. 2021. Almost state-of-the-art automatic speech recognition in tensorflow 2.

Prashanth Gurunath Shivakumar and Shrikanth Narayanan. 2021. End-to-end neural systems for automatic children speech recognition: An empirical study.

Wu, Fei. 2020. Child speech recognition as low resource automatic speech recognition.