Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6340–6345
Marseille, 11–16 May 2020

Evaluating and Improving Child-Directed Automatic Speech Recognition

Eric Booth[‡], Jake Carns[*], Casey Kennington[*], Nader Rafla[‡],
[*]Department of Computer Science & [‡]Department of Electrical and Computer Engineering
Boise State University, USA
{ericbooth, jakecarns}@u.boisestate.edu
{caseykennington, naderrafla}@boisestate.edu

Abstract

Speech recognition has seen dramatic improvements in the last decade, though those improvements have focused primarily on adult speech. In this paper, we assess child-directed speech recognition and leverage a transfer learning approach to improve child-directed speech recognition by training the recent DeepSpeech2 model on adult data, then apply additional tuning to varied amounts of child speech data. We evaluate our model using the CMU Kids dataset as well as our own recordings of child-directed prompts. The results from our experiment show that even a small amount of child audio data improves significantly over a baseline of adult-only or child-only trained models. We report a final general Word-Error-Rate of 29% over a baseline of 62% that uses the adult-trained model. Our analyses show that our model adapts quickly using a small amount of data and that the general child model works better than school grade-specific models. We make available our trained model and our data collection tool.

Keywords: Speech Recognition, Children, transfer learning, Data Collection

1. Introduction

Adult automatic speech recognition (ASR) systems have rapidly improved over the last decade, with modern systems approaching recognition rates comparable to human-levels in controlled environments (Shu, 2017; Li, 2019), yet the accuracy of child speech recognition systems is lagging far behind (Kennedy et al., 2017). Some reasons may include the fact that children's vocal tracts are much smaller than those of adults, the vocal tracts of children change rapidly as they mature so the acoustic properties of speech vary between children much more than adults, and speech production is a complex motor activity that children are still learning to master, so the variation in speech production from the same speaker is much higher in children than in adults. For example, Figure 1 shows how the word error rate (WER) of a recent English ASR system, adapted specifically for children, declines significantly with the grade level (i.e., a proxy for age) of the speaker (Yeung and Alwan, 2018).

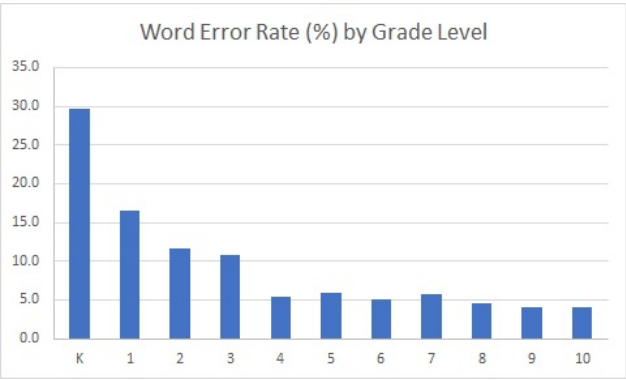

Figure 1: Word Error Rate (WER) by grade level.

Figure 1 shows that researchers cannot fully claim human-level recognition capabilities when the ASR system in question can only recognize adult speech, whereas humans are able to recognize adults and children. Moreover, speech disorders affect 8% of children (Black et al., 2015) in the United States (particularly for younger children ages 3-6, where the WER is the highest in Figure 1), yet only about half of those children receive timely intervention, and the frequency of that intervention is often not enough (Namasivayam et al., 2015). Computer based speech therapy is a promising approach to increase the effectiveness and accessibility of speech therapy for children, and research has shown that computer based speech therapy can be effective, especially when used in conjunction with traditional therapy (Chen et al., 2016; Furlong et al., 2017). However, this can only happen if child-directed ASR is reliable and useful to researchers and therapists.

There have been attempts to apply systematic pitch changes to adult data to improve child-directed ASR without marked improvements (Liao et al., 2015). As with adult ASR, more data will result in better models, yet one of the key challenges of developing ASR for children is collecting large data sets containing error free samples of child speech, in part because child speech data is protected by the Children's Online Privacy Protection Act (COPPA). For ongoing research, it is important that a child-directed ASR be usable offline in a COPPA-compliant environment.

In this paper, we offer a streamlined solution to collecting child speech samples to improve ASR, and show in an experiment how *transfer learning* can be used to train child directed ASR models using only a small amount of child speech. In our specific use of transfer learning, a deep neural network which has been previously trained on a large data set is adapted to work better for a different population on the same task; i.e., ASR. This is an appealing approach to create models directed towards child speech, including disordered speech, where large data sets are difficult to obtain. transfer learning has been used successfully in many deep learning applications, including image classification and natural language processing. For example, (Chen et al., 2018) showed how cross-language transfer learning (though in some cases arguably transfer learning) showed promising results for low resourced languages

(where the language is extinct or there are very few native speakers).

In the following section we describe our data – adult and child data – then explain our model, DeepSpeech2, and how we applied transfer learning to improve the WER in an evaluation set of child data. Our results show that the accuracy of models generated using transfer learning can approach the accuracy of state of the art cloud-based systems. We also show that when a small amount of child speech data is available for transfer learning, a general child-adapted model is more accurate than grade-specific models. Our additional analyses show that the model adapts quickly to a small amount of child speech data, though the adapted model, as one might expect, performs poorly on adult speech.

2. Data

In this section we describe the adult speech data (LibriSpeech) and child data (CMU Kids and our own novel recordings) that we used for training our initial model and for transfer learning.

2.1. Adult Data: LibriSpeech

The adult speech data used for this work is from the LibriSpeech corpus (Panayotov et al., 2015), which is commonly used for training and benchmarking ASR (Collobert et al., 2016) and downstream tasks that use ASR (Chung et al., 2016) applications. It consists of approximately 1000 hours of transcribed adult English speech. Most of the utterances are between 1 and 16 seconds long, with the average utterance about 12 seconds long. An example is shown in (1) below.

(1) AND I WAS BORN THERE YET I DO ASSURE YOU I OFTEN LOSE MY WAY AMONG THE VERY PILES OF WAREHOUSES THAT ARE BUILT UPON MY FATHER'S ORCHARD DO WE PART HERE

The data is derived from read audio books available in the public domain. The speech has been carefully segmented and separated into data sets as shown in Table 1. In this work, all three of the LibriSpeech training sets (about 960 hours) were used for training.

Table 1: LibriSpeech Data Set

subset	hours	min/speaker	female	male
dev-clean	5.4	8	20	20
test-clean	5.4	8	20	20
dev-other	5.3	10	16	17
test-other	5.1	10	17	16
train-clean-100	100.6	25	125	126
train-clean-360	363.6	25	439	482
train-other-500	496.7	30	564	602

2.2. Child Data: CMU Kids Dataset

The child speech data we used for our experiments is the CMU Kids Dataset (Eskenazi and Graff, 1997) which contains 5180 total North American English utterances. The corpus consists of 76 children ranging in age from six to eight (1st through 3rd grade at the time of recording) with the exception of one child who was 11 and in the 6th grade at the time of recording. There were 24 male and 52 female speakers. Each utterance is a short phrase, averaging 6 seconds in duration, resulting in a total of about 9 hours of child speech. We reserved 300 of those utterances for development and an additional 300 for testing (i.e., 4580 for training). An example utterance is shown in (2).

(2) A BLUE BUTTERFLY FLEW BY

Many of the utterances have small pronunciation errors which are typical of young children. For example, in one case when a child was prompted to give the example utterance shown in (2), the actual utterance was "A blue butterfly /s/flew by", where the child added the phoneme /s/ to the beginning of the word "flew". The transcripts used for training and testing were not modified to account for these small errors in speech production.

Besides differences in acoustic qualities, Examples (1) and (2) illustrate another important challenge: adult utterances tend to be longer and contain more complicated syntax and vocabulary than the child data.

2.3. Child Data: Novel Recordings

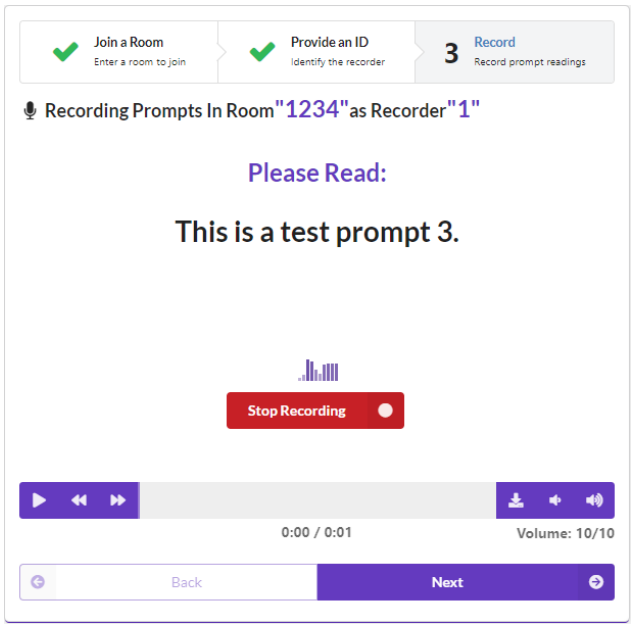

Figure 2: Interface used by children to recording prompt readings generated by the administration tool

Due to restrictions surrounding privacy and collection of child data, publicly available child speech data sets are scarce and difficult to obtain. The CMU Kids data set described above gives us a starting point. For a more rigorous evaluation, we collected new recordings from additional children while they read prompts selected from the CMU Kids data set. To streamline this process for this work and future work, we developed a web-based, audio recording tool. The tool is designed to easily administer prompt reading tasks to groups of speakers

via a child-friendly interface (as shown in Figure 2), and exports audio data packaged into user-defined groupings that can be downloaded in a zip file for external use.[1] The collected audio data is stored in 16k sample rate, uncompressed, wav files, which work for the DeepSpeech2 training requirements and match the sample rate of the previous data set. We recognize that text prompts bring an additional layer of complexity as it requires children to have learned how to read certain words. This makes our evaluation more robust as pronunciation errors due to misreading a word are recorded, giving the ASR model examples of those kinds of errors. We leave the ability to prompt children with pre-recorded audio (i.e., instead of text) for future work.

Using this tool, we collected novel data from 20 children, 9 male and 11 female, between grades 1 and 5. Each child read approximately 20 prompts, randomly selected from the CMU Kids Dataset, for a total of 454 recorded utterances. Most of the children were able to use the tool after a brief demonstration without any help. A touch screen laptop was provided for them as the user interface, and an external USB microphone was used for the recording. Per our observations, even the youngest of the children were able to navigate the user interface easily and appeared to find the tool and touch screen interface intuitive to use.

3. Model: DeepSpeech2

We used the DeepSpeech2 model described in Amodei et al. (2015), which has been shown to produce state-of-the-art results on adult speech when trained on sufficiently large data sets.[2] The model consists of two convolutional neural (CNN) network layers, followed by five bidirectional gated recurrent unit layers (GRU), and finally, a fully connected output layer. The total number of model parameters is 41.2 million. The input to the model is a spectrogram generated from the raw audio signal using a hamming window with a size of 20ms and a stride of 10ms. The number of frequency bands in the spectrogram is 160 and is computed as shown in equation 1, where the sample rate for all audio files in this experiment is 16k samples/second.

$$n\_spectrums = sample\_rate * window\_stride \quad (1)$$

lrec
Each CNN layer uses a two-dimensional filter. The first CNN layer uses a filter size of 41x11 (stride of 2x2) and the second CNN layer uses a filter size of 21x11 (stride of 2x1). Each of the bidirecitonal GRU layers have an input size of 800. The GRU is a type of recurrent neural network (RNN) with a gating function which aides in convergence during training (Cho et al., 2014). The output of the model is decoded using Connectionist Temporal Classification (CTC) (Graves et al., 2006) loss function during training. The model has 29 output classes {A, B, C, ..., Z, *apostrophe*, *space*, *blank*}. The *blank* character is a special character used for the CTC decoding. In this work, an additional language model is not used since the size of the adult training set (1000 hours) is sufficiently large to enable the GRU layers to learn a reasonably accurate language model directly. The goal of the model in this work is to output a transcription – a sequence of graphemes – directly. However, RNN-CTC based models have also been shown to work well for phoneme production (Miao et al., 2016), provided that phonemic transcriptions are available for training. Therefore, this model could potentially be used for future work related to child speech development such as pronunciation verification and language learning, where phoneme output is required. Importantly, this model can be used offline for child-directed ASR without worry that a cloud-based provider is collecting protected child data. Our trained models are available for download[3]

4. Baseline Evaluations on Child Speech Data

In this section, we evaluate our child test set against existing models to establish a baseline for the experiment in the following section.

4.1. Task, Metrics, & Procedure

To establish a baseline of results for this test set, we evaluated the 300 test utterances from the CMU Kids Dataset on three different models (one with two settings):[4]

- Sphinx4 with the most recent pre-trained English model (Walker et al., 2004)[5]
- DeepSpeech2 trained on the LibriSpeech adult speech training data set as described above
- DeepSpeech2 trained on the CMU Kids training data described above
- Google Speech API version 1[6]

Our target metric is WER; i.e., lower scores denote better results.

4.2. Baseline Results

Table 2: Baseline results on the child speech test set (WER)

Sphinx4 - Adult Model	100%
DeepSpeech2 - Adult Model	62%
DeepSpeech2 - Child Model	60%
Google Speech-to-Text Version1	24%

The results are summarized in Table 2. Though the Google Speech API results are encouraging, a WER of 24% still means that one in every four words is misrecognized;

[1] Our prompt recording tool is freely available: https://github.com/bsu-slim/prompt-recorder

[2] https://github.com/SeanNaren/deepspeech.pytorch

[3] Our trained models are available for download.https://bitbucket.org/bsu-slim/child-speech-models/

[4] We followed the approach taken by (Baumann et al., 2016) to systematically evaluate against Sphinx4 and the Google Speech API.

[5] https://github.com/cmusphinx/sphinx4

[6] https://cloud.google.com/speech-to-text/

clearly there is room for improvement – current adult ASR systems yield less than 5% WER. Moreover, the Google Speech API is cloud-based and protected child data may be compromised. The Sphinx4 system did not perform well with this particular dataset, we assume due to the modeling approach (i.e., Gaussian mixture model) and the domain-specific data that did not include child data. The DeepSpeech2 model performed better than Sphinx4, though with a correct recognition rate of one in three words using either the adult data or the child data for training. There is clearly room for improvement, which we explore in the experiment below. As DeepSpeech2 performed better than Sphinx4, we build on that model.

5. Experiment: Transfer Learning using DeepSpeech2

In this section, we explain the task, metrics, procedure, and results of our transfer learning experiment. We then offer some analyses of our model.

5.1. Task & Procedure

We task our model with transcribing from an audio input to a string output. For this experiment, we applied transfer learning to the DeepSpeech2 trained adult model using the CMU Kids training data. In our approach, we used the DeepSpeech2 model trained on the LibriSpeech adult data as a starting point without modification of the layers or re-initialization of the learned parameters. That model was trained for 20 epochs on the adult training data with a batch size of 16 and learning rate of 0.0003. The training regime included a learning rate annealing factor of 1.1 so that the learning rate decreased slightly with each epoch, a weight decay factor of 0.5 for regularization, and a momentum of 0.9 to aid in convergence. Following Amodei et al. (2015), we also applied Batch Normalization to all layers. This resulted in the DeepSpeech adult baseline model that yielded 62% WER, as reported in the baseline results.

After the model was initially trained on the adult data, the learning rate was reset to the original value, and we further trained an additional 20 epochs with the child speech data using the same hyper-parameters as in the initial training steps. We varied the amount of training data to show how much added child data is required for the model to improve, and we isolated grades 1, 2, and 3 to determine how well the model performs on specific age levels.

5.2. Metrics

As with the baselines above, we evaluate on the test set of 300 child utterances from the CMU Kids dataset to arrive at a WER score. In addition, we further evaluated our models against 454 utterances from our novel dataset.

5.3. Results

Table 3 shows the WER versus the amount of child speech data (random mix of all age groups) that was used for transfer learning. The results show that even with a very small amount of child speech data (25% of the available data, or just over 2 hours), transfer learning can be effective. The model which applied transfer learning using all of the data resulted in a WER of 29%, a significant improvement over either the adult-specific or child-specific models. Although the final results using all of the available data do not quite achieve the accuracy of a state-of-the art commercial system like Google's Speech API, the trend of continuously improving accuracy with more data shows that it may be possible to approach or exceed this accuracy by leveraging more data. As the CMU Kids data only contained roughly 9 hours of speech, we predict that an additional 4 hours of data would put this method on par with the Google Speech API with the added benefit of being usable offline.

Table 3: Transfer learning results on child speech test set (WER)

Training Data Used	WER
Adult data only	62%
Child data only	60%
Adult + 25% child data	47%
Adult + 50% child data	40%
Adult + 75% child data	34%
Adult + 100% child data	29%

5.4. Analyses

5.4.1. Checking for Generalizability

In Liao et al. (2015), Google researchers discuss techniques for training a general purpose model that could work well for both child and adult speech. In their approach, the primary focus was towards training on a massive data set containing millions of utterances of both child and adult speech. It should be noted that in our approach, a general purpose model is not the goal. When we tested our model on adult speech after transfer learning was applied, the accuracy on adult speech had degraded to 47.1%. The primary goal in our approach is to enable accurate child speech models to be created using as little child speech data as possible to serve specific use cases such as very young children, or children with speech, language or hearing disorders. We leave for future work developing a model that can be used generally for both child and adult speech.

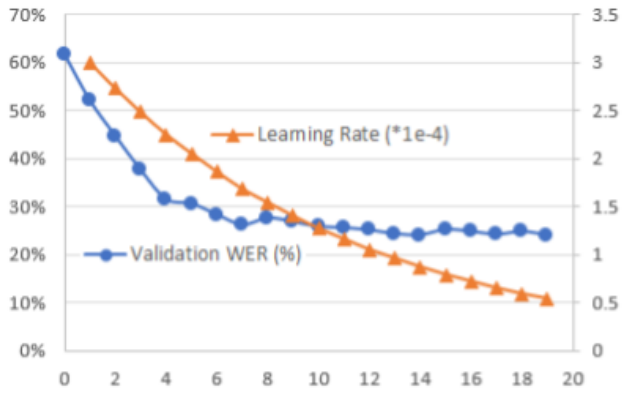

Figure 3: Validation results per epoch during the transfer learning training process. Epoch 0 corresponds to the adult model before transfer learning is applied. WER is reported for the development data set.

5.4.2. Model Learning

Here we explore and report what our model is learning. Figure 3 shows the WER of the validation data set and learning rate (which adaptively decreases) for each epoch of the transfer learning training process using all of the child data. This figure tells us that the model is indeed quickly learning to adapt the adult model to the child data; in fact, by as few as 6 epochs, the model has largely made the proper adjustments to improve results with the child data. We conjecture that the rapid adaptation–despite the small amount of child data–is due to the similarity of the task and domain; i.e., both are speech data and the model is learning a functional mapping from the pre-processed audio (as explained above) to the graphemes.

5.4.3. Transfer Learning by Grade Level

Table 4 shows the WER results on the test data set by grade level after transfer learning has been applied using all of the available training data. Google Speech API results are also provided as a reference. The WER of the 1st grade test set are significantly worse than the 2nd and 3rd grade sets. This is consistent with prior results from (Yeung and Alwan, 2018) and others which demonstrates that ASR systems typically perform worse for younger children. Although the accuracy of our transfer learning model on the 1st grade data set leaves significant room for improvement, it is on par with Google Speech to Text. This is encouraging, especially given that such a small amount of child speech data was required to train the model.

To address the high WER of the 1st grade test set, we attempted to generate a 1st grade specific model by performing transfer learning on the adult trained model using only the 1st grade training data (about 987 out of the total 4580 training utterances), but this yielded a WER of 42.0% on the 1st grade test set, even worse than the WER of 39.4% that was achieved when training data from all age groups was used for transfer learning. In this case, the general purpose child speech model trained on a larger data set outperformed the age specific model trained on less data.

Table 4: Transfer learning WER results on the child speech by age group. For these results, all age groups were used during transfer learning, but the test set was separated into age groups.

Test Set	Utterances	This Work	Google
1st Grade	61	**39.4%**	39.9%
2nd Grade	139	24.8%	**18.0%**
3rd Grade	90	26.9%	**20.9%**

5.4.4. Evaluation of Novel Data

Table 5 compares the WER performance on the novel data described above with the CMU Kids test set. The results show that the transfer learning model performs significantly better on the novel data than either the adult or child only models. This experiment also shows that the transfer learning model generalizes much better to a novel data set, as the WER using the transfer learning model increased by only 6% on the novel data when compared to the CMU Kids test set, while the child only model showed a 26% WER increase.

Table 5: Results on novel data set compared to CMU Kids test set (WER)

Training Data Used	CMU Kids	Novel data
Adult data only	62%	59%
Child data only	60%	86%
Adult + child data	29%	35%

5.4.5. Analysis of Mistakes

Using the best performing child model (i.e., trained on the adult data, then adapted using all of the child data), the three examples in (3)-(5) show some of the common mistakes (each has a *Ref* and a *Hyp*, the latter being what the model produced):

(3) *Ref:* **the** scientist was surprised
Hyp: **a** scientist was surprised

(4) *Ref:* some people recycle food garbage
Hyp: some people recycle**d** food garbage

(5) *Ref:* they jump from one tree to another looking for fruit
Hyp: they **jim** from one tree to **a nother lokin** for fruit

The mistakes in (3) and (4) show that some of the mistakes are common mistakes that, we assume, adults make when listening to children speech: **the** swapped with **a** in (3) or an added past tense in (4). For (5) on the other hand, **another** is split, and **lokin** is not even an English word in either of the datasets. These kinds of mistakes are a byproduct of the model which maps directly from audio to graphemes: it stays true to the actual utterance rather than attempt to match it to the closest word as done in other ASR models. This is desired behavior for future work as we want the model to produce transcriptions of what is actually uttered; i.e., not fitted to a pre-defined vocabulary.

5.5. Discussion

The results and analyses above show that further research into transfer learning for child directed speech recognition systems is promising and merits additional exploration. Specifically, our transfer learning approach explained above works well when transferring from adult to child speech with only 9 hours of child speech. This could be extended to improve ASR for children with speech disorders, non-native speech, or domains with very technical vocabularies (e.g., health care or law). Moreover, the model adapts quickly to the child data (i.e., with a small amount of data in only 8 epochs) as shown in Figure 3. Though when we evaluated the final model (i.e., trained on adult speech and adapted to child speech) and evaluated on the adult test set, the results were substantially worse, meaning our model is not completely generalizable, but as our results show in Table 4 our model does work as a generalized model for children roughly in grades 1-3.

6. Conclusion & Future Work

We conclude that transfer learning using an adult-trained deep learning model with CNN and GRU layers on a small

amount of child data works well and is a robust approach to the challenging problem of recognizing child speech, as explained in Section 2.

Next steps include collecting more child speech data to improve the model for effective use in real applications. We also plan to evaluate this approach on models trained for phoneme production. We will also explore more sophisticated approaches to transfer learning with DeepSpeech2 by freezing subsets of the CNN and/or GRU layers using an approach similar to Gale et al. (2019). Additionally, we would like to explore the impact of adding additional untrained layers.

7. References

Eskenazi, Maxine, J. M. and Graff, D. (1997). The cmu kids corpus ldc97s63.

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., and Woelfel, J. (2004). Sphinx-4 : A Flexible Open Source Framework for Speech Recognition. *Smli*, (TR-2004-139):1–9.

Graves, A., Fernandez, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks. *International conference on Machine Learning*, pages 369–376.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.

Black, L., Vahratian, A., and Hoffman, H., (2015). *Communication Disorders and Use of Intervention Services Among Children Aged 3 to 17 Years: United States, 2012*.

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Han, T., Hannun, A., Jun, B., Legresley, P., Lin, L., Narang, S., Ng, A., Ozair, S., Prenger, R., Raiman, J., Satheesh, S., Seetapun, D., Sengupta, S., Wang, Y., Wang, Z., Wang, C., Xiao, B., Yogatama, D., Zhan, J., and Zhu, Z. (2015). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *arXiv*.

Liao, H., Pundak, G., Siohan, O., Carroll, M. K., Coccaro, N., Jiang, Q.-m., Sainath, T. N., Senior, A., and Bacchiani, M. (2015). Large Vocabulary Automatic Speech Recognition for Children. In *The Proceedings of Interspeech*, pages 1611–1615. ISCA.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*.

Namasivayam, A. K., Pukonen, M., Goshulak, D., Hard, J., Rudzicz, F., Rietveld, T., Maassen, B., Kroll, R., and Van Lieshout, P. (2015). Treatment intensity and childhood apraxia of speech. *International Journal of Language and Communication Disorders*.

Chung, Y. A., Wu, C. C., Shen, C. H., Lee, H. Y., and Lee, L. S. (2016). Audio Word2Vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 08-12-September-2016:765–769.

Miao, Y., Gowayyed, M., and Metze, F. (2016). EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*, pages 167–174.

Baumann, T., Kennington, C., Hough, J., and Schlangen, D. (2016). Recognising Conversational Speech: What an Incremental ASR Should Do for a Dialogue System and How to Get There. In *Proceedings of the International Workshop Series on Spoken Dialogue Systems Technology (IWSDS) 2016*.

Chen, Y. P. P., Johnson, C., Lalbakhsh, P., Caelli, T., Deng, G., Tay, D., Erickson, S., Broadbridge, P., El Refaie, A., Doube, W., and Morris, M. E. (2016). Systematic review of virtual speech therapists for speech disorders. *Computer Speech and Language*.

Collobert, R., Puhrsch, C., and Synnaeve, G. (2016). Wav2Letter: an End-to-End ConvNet-based Speech Recognition System. pages 1–8.

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., and Belpaeme, T. (2017). Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations. In *Proceedings of HRI'17*.

Furlong, L., Erickson, S., and Morris, M. E. (2017). Computer-based speech therapy for childhood speech sound disorders. *Journal of Communication Disorders*, 68(February 2016):50–69.

Shu, C. (2017). Microsofts speech recognition system hits a new accuracy milestone. *TechCrunch*.

Yeung, G. and Alwan, A. (2018). On the difficulties of automatic speech recognition for kindergarten-aged children. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018-September(September):1661–1665.

Chen, W., Hasegawa-Johnson, M., and Chen, N. F. (2018). Topic and keyword identification for low-resourced speech using cross-language transfer learning. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018-September(September):2047–2051.

Li, A. (2019). Google speech recognition is now almost as accurate as humans. *9to5Google*.

Gale, R., Chen, L., Dolata, J., Van Santen, J., and Asgari, M. (2019). Improving ASR systems for children with autism and language impairment using domain-focused DNN transfer techniques. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2019-September, pages 11–15. International Speech Communication Association.