INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Simulating reading mistakes for child speech Transformer-based phone recognition

Lucile Gelin[1,2] *Thomas Pellegrini*[1], *Julien Pinquier*[1], *Morgane Daniel*[2]

[1]IRIT, Paul Sabatier University, CNRS, Toulouse, France
[2]Lalilo, France

{lucile.gelin, thomas.pellegrini, julien.pinquier}@irit.fr, morgane@lalilo.com

Abstract

Current performance of automatic speech recognition (ASR) for children is below that of the latest systems dedicated to adult speech. Child speech is particularly difficult to recognise, and substantial corpora are missing to train acoustic models. Furthermore, in the scope of our reading assistant for 5-8-year-old children learning to read, models need to cope with disfluencies and reading mistakes, which remain considerable challenges even for state-of-the-art ASR systems. In this paper, we adapt an end-to-end Transformer acoustic model to speech from children learning to read. Transfer learning (TL) with a small amount of child speech improves the phone error rate (PER) by 48.7% relative over an adult model and outperforms a TL-adapted DNN-HMM model by 21.0% relative PER. Multi-objective training with a Connectionist Temporal Classification (CTC) function further reduces the PER by 4.8% relative. We propose a method of reading mistakes data augmentation, where we simulate word-level repetitions and substitutions with phonetically or graphically close words. Combining these two types of reading mistakes reaches a 19.9% PER, with a 13.1% relative improvement over the baseline. A detailed analysis shows that both the CTC multi-objective training and the augmentation with synthetic repetitions help the attention mechanisms better detect children's disfluencies.

Index Terms: child speech, transformer, connectionist temporal classification, data augmentation, synthetic reading mistakes

1. Introduction

The human body changes continuously during the early years of life, and in particular the speech production apparatus takes time to fully develop. At ages 5-7, children's articulatory mechanisms are not stable, which implies intra- and inter-speaker spectral variability. Developmental stabilisation of pitch control only occurs around age 8 [1]. Due to slow growth of the vocal tract, their fundamental and formant frequencies do not reach a mature adult level before age 15 [2]. Furthermore, phonological errors, like weak-syllable deletion, or phoneme substitution due to bad positioning of the tongue and lips, are customary among young children's speech, and show to disappear with age [3].

Reading tutors have a strong pedagogical impact for reading learners, and several projects, applied to different languages, age groups and reading tasks, have been implemented over the years [4, 5, 6, 7]. Lalilo provides an online reading assistant[1] for 5-8 year-old children, featuring a reading aloud exercise where children record themselves reading, and get feedback on their reading. Speech recognition for children learning to read is an arduous task since it combines the difficulties of child speech acoustic and prosodic characteristics with more difficulties brought by non-proficient readers' speech. Young readers indeed demonstrate very specific prosodic events such as hesitations, false starts, repetitions and various mispronunciations that may be laborious to detect automatically [8, 9]. Detecting reading mistakes necessitates specific adaptation of the acoustic model. ASR studies for second language learning, which aim at detecting pronunciation mistakes, use manual annotations of non-native speakers, which are time-consuming to obtain, to incorporate pronunciation mistakes in their training dataset [10, 11]. In this work, we present a method to simulate two types of reading mistakes, with the objective of training our model to better recognise children read speech.

Prior studies on ASR for child speech demonstrated that the performance is lower than for adult speech [12, 13]. A study on grade-specific ASR performances showed significant gaps of word error rate (WER) between children of age 5, 6 and 7+, confirming that the younger the children, the lower the accuracy [9]. [14] used a hybrid deep neural network - hidden Markov model (DNN-HMM) system, jointly trained on adult and children data and adapted to speech from an age-specific group. A factorised time-delay neural network (TDNNF-HMM) shows to outperform Gaussian mixture models - hidden Markov model (GMM-HMM) models for child speech recognition in [15]. For a children language learner application, [10] presents a DNN-HMM that, even trained on less data, surpasses GMM-HMM systems. Valuable insights on acoustic modelling for child speech recognition with DNN-HMM are given in [13].

Recent end-to-end architectures have the advantage to directly predict the final label sequence without the need of preprocessed alignments nor of an HMM to yield the hypothesis. They have shown to reach the performance of hybrid architectures on a broad range of ASR datasets [16]. A first step towards end-to-end ASR was made with the Connectionist Temporal Classification (CTC) objective function [17]. Sequence-to-sequence (seq2seq) encoder-decoder structures were then proposed [18], coupled with attention mechanisms to link the encoder and decoder of the models [19, 20]. The Transformer architecture [21, 22] replaced the customary recurrent neural networks (RNN) by self-attention modules, and showed to outperform RNN-based seq2seq models [16].

End-to-end acoustic modelling for child speech recognition is not yet common due to limited available child speech data. In [23], the authors show improvement with a CTC-based end-to-end system trained on very large quantities of mixed adult and child speech data. Usage of seq2seq models for child speech recognition is a new research subject, as show the extremely recent communication of technical reports on this matter [24, 25, 26]. Our work aims at bringing valuable insights to this new research domain, while finding solutions to improve

[1]https://www.lalilo.com/

http://dx.doi.org/10.21437/Interspeech.2021-2202

the performance on young readers' speech. We show that:

- A phone-level Transformer model trained with transfer learning (TL) and a CTC multi-objective outperforms a TL-trained TDNNF-HMM model. We observe that the CTC function keeps the attention mechanisms from missing disfluencies such as repetitions.
- Data augmentation with synthetic reading mistakes, an original method that simulates word-level repetitions and substitutions, reveals effective in improving the recognition performance on a wider proportion of disfluencies and reading mistakes.

In section 2, we present the speech material and a detailed presentation of typical children reading mistakes. Section 3 presents the methods used to improve the robustness on young readers' speech, including our novel data augmentation method with synthetic reading mistakes. Section 4 describes our Transformer model, and section 5 displays the results of the various methods used. Finally, section 6 provides deeper studies on the robustness improvements towards reading mistakes.

2. Speech material

We use two sets of French speech: the *Common Voice* adult corpus, and an in-house children corpus, hereinafter called *Lalilo*.

2.1. Adult dataset: Common Voice

The Commonvoice corpus[2] is created through a participatory online platform, where everyone can record himself reading sentences. In French, the training set we used for these experiments contains approximately 150 hours of speech. Each recording is validated by two annotators, thereby the corpus contains few miscues.

2.2. Child dataset: Lalilo

The Lalilo corpus contains recordings of Kindergarten-to-2nd-Grade children, aged 5-8, reading aloud isolated words, sentences and short stories. The recordings are mainly gathered through the read aloud exercise of the Lalilo platform, which is most often used in classrooms under reduced supervision: they contain variable levels of babble noise.

The training and validation sets contain respectively 13 and 0.41 hours of data. They are only composed of correctly pronounced utterances (sentences and isolated words) that have been manually labelled as correct. The text prompted to the student has been converted to phones with a pronunciation dictionary. The test set contains 0.48 hours of utterances that may or may not contain reading mistakes. The test is only composed of sentences in this work, as they contain a wider variety of reading mistakes in comparison to isolated words. Phones read by children have been manually transcribed by two human judges, and recordings have been discarded in case of disagreement.

2.3. Reading learners mistakes

Reading mistakes done by beginner children are diverse and sometimes unique, which makes the manual annotation of data quite difficult for human experts, and can cause all the more confusion for a phone recognition system. Understanding the different types of reading mistakes can help analyse the system's behaviour when encountering these. The percentages displayed represent the proportion of each category of mistakes into our test set, which makes a total of 13.1% of words that are not correctly read.

- **Mispronunciation** (5.1%): a word contains one or several phone substitutions, insertions or deletions. The resulting word can exist, in which case it is called a *word substitution*.
- **Repetition** (4.5%): a word is repeated one or several times. Each repeated word may contain a mistake.
- **Deletion** (2.9%): a word is skipped by the student.
- **Hesitation** (0.6%): a word contains one or several silences due to the child hesitating.

Among mispronunciations, 50% can be considered as word substitutions, meaning that the resulting word exists either in the French language, or in the Lalilo vocabulary that contains pseudo-words as well. The word substitutions represent in this way 2.5% of the dataset. As for repetitions, 77% of the repeated words do not contain a reading mistake, which makes 3.5% over the whole set.

3. Methods

3.1. Transfer learning

Transfer learning (TL) consists in adapting a model trained on a large out-of-domain dataset with a small amount of in-domain data. It is used in many ASR applications, such as adaptation between languages [27, 28], between different speech types, such as broadcast news and conversational speech in [27], or between native and non-native speech [29]. In our case, adult-to-child adaptation, the source model is a model trained on adult speech, while the adaptation data is child speech data. This particular adaptation is very sensitive to the amount of target data and the target children's age, since their vocal apparatus and speech quality are so different to adults, and vary greatly during children's growth [13]. In [13], the authors observe that young children's acoustic and prosodic characteristics are highly complex and variable, and suggest to adapt the whole network with a minimum amount of 10 hours of adaptation data. Our training child corpus containing approximately 13 hours of speech from very young children (5-8 years old), we follow their findings and apply transfer learning on the whole source model.

3.2. CTC multi-objective training

The CTC paradigm, introduced by [17], discards the obligation of having an HMM by learning automatically alignments between the input and output sequences. In phone recognition applications, it aligns in a monotonic way the speech frames input and their phone sequence output, using the conditional independence assumption and a dynamic forward-backward algorithm to find the most probable alignment.

Attention mechanisms can sometimes be too flexible for ASR, since they allow non-monotonic alignments while speech recognition inputs and outputs are intrinsically sequential. Multi-objective training can be done with the cross-entropy and CTC functions for attention-based encoder-decoder systems such as the Transformer [30, 31]. Using the CTC objective aims at constraining the attention mechanism to find monotonic alignments. This constraint is particularly interesting for recognising repetitions of words, as attention mechanisms will tend to merge several occurrences of a word into a single one, thus missing the repeated words. The CTC loss is combined with the cross-entropy loss with a certain weight (0.3 in this work).

[2]Corpus available at: `https://voice.mozilla.org/fr`

3.3. Synthetic reading mistakes data augmentation

As seen in section 2.3, 13.1% of the words in our test set contain a reading mistake. Since our training data contains only correctly read utterances, the model is not prepared to deal with children's reading mistakes, and would most certainly benefit from encountering some during training. However, the phone-level annotation process being very time-consuming, we aim at creating synthetic reading mistakes for data augmentation. Since mispronunciations and repetitions are the most common mistakes (respectively 5.1% and 4.5%), we propose to simulate these two types of mistakes. Furthermore, since most of these mistakes result in existing words that are present in the training set, we can easily extract whole words to generate mistakes, avoiding imprecise phone-level operations.

All possible synthetic reading mistakes are detailed in Table 1. Each original training utterance is aligned with a word-level GMM-HMM, to obtain each word's time boundaries. For word substitutions, we create for each existing word a list of possible substitutions, inspired by our analysis of children mistakes: substitutions of two vowels (1) or two consonants (2), inversion of phones (3, only for two-phones words), and phonetically- or graphically-based false starts (4). Among this list, we filter out the words that do not exist in the French language. For each possible substitution word, we extract all occurrences in our training audio dataset, obtaining real audio segments. Finally, to create the word-substitution-augmented utterances, we randomly: select a word to be substituted, get a possible substitution word, and select a corresponding audio segment (possibly from a different speaker) that we insert in place of the original word. We limited the number of substituted words to one per utterance to avoid severe mismatches between words, and normalised the energy of the recordings to smooth out the transitions. For repetitions, we randomly select the word(s) to be repeated in an utterance, then extract the corresponding audio segments and repeatedly insert them in the original recording. Words can be repeated in a pattern (5), or individually (6).

Table 1: *Description of possible synthetic reading mistakes on the sentence: [il ʁul a velo]*

Mistake type	Description
(1) Sub vowel	A vowel is substituted by another *Example:* [il ʁ**ɑl** a velo]
(2) Sub consonant	A consonant is substituted by another *Example:* [il **b**ul a velo]
(3) Sub inversion	Two-phones words are inverted *Example:* [**li** ʁul a velo]
(4) Sub false start	Only the beginning of the word is read *Example:* [il **ʁu** a velo]
(5) Rep pattern	A pattern of words is repeated *Example:* [**il ʁul a** il ʁul a velo]
(6) Rep individual	Word(s) are repeated individually *Example:* [**il** il ʁul a **velo** velo]

Both methods are applied on a subset of training sentences. For each utterance, we create one substitution-augmented version, and one repetition-augmented version. Based on observed proportions in our data, we searched for the optimal proportions of substituted and repeated words to insert in our training data, in ranges of 0.5-3.5% for word substitutions and 2.5-5.5% for word repetitions. The proportions that yielded the best validation loss when training with both augmentations were selected: 1.4% substituted words and 3.8% repeated words.

4. System description

Presented in [21] and adapted to speech recognition in [22], the Transformer model follows a seq2seq encoder-decoder architecture, but relies solely on attention mechanisms, instead of recurrent neural networks. The recurrence, essential to extract position information in speech frames, is replaced by positional encodings concatenated to the input encodings, as well as multi-head self-attention mechanisms and position-wise feed-forward neural networks in the encoder and decoder blocks. Discarding the need of recurrent neural networks allows to compute dependencies between each pair of positions at once, instead of one by one. It enables faster training and more parallelisation in comparison with RNN-based encoder-decoder systems.

Our Transformer models follow the architecture of the original paper [21]. The hyper-parameters are the same as the models trained in a previous study [32]. The audio information, 80-dimensional filter banks following [16, 31], is given to the encoder, and the reference phone labels to the decoder. The model dimension is 256, which is used for embeddings, sinusoidal positional encoding, layers in both encoder and decoder, and for the classification network used to process the output. The encoder contains six layers, and the decoder four. The models are trained with the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon =$ 1e-9. We use the same specific learning rate scheduler as [21] for training, with 4000 warm-up steps. Models are trained on 100 epochs on a single GTX 2080 Ti GPU, which take an average of 30 and 3 hours for adult and TL models, respectively. The final model contains a total of 14.3M parameters.

5. Evaluation

In this work we aim at transcribing accurately what the child has read, including potential phone-level reading mistakes. Therefore, we measure performance with a phone error rate (PER), that is computed as the number of phone recognition errors over the number of reference phones.

Table 2: *PER (%) obtained with TL-adapted acoustic models and our augmentation methods, tested on child speech*

Model	PER
Transformer *(baseline)*	22.9
Transformer +CTC	21.8
+Sub	21.7
+Rep	20.4
Transformer +CTC +Sub +Rep	**19.9**

The source Transformer model, trained over approximately 150 hours of adult speech, scored 7.5% PER on matched adult speech, but scored very high on child speech (44.6%). Transfer learning reduces the PER by 48.7% relative, reaching a score of 22.9%, as can be seen in Table 2. This model was shown in a previous work to outperform an hybrid TDNNF-HMM trained with TL by 21.0% relative PER [32]. Multi-objective training with CTC enables to lower further the PER, reaching 21.8%, which corresponds to a relative improvement of 4.8%.

Data augmentation with synthetic substitutions, designated by *+Sub* in Table 2, slightly improves the Transformer+CTC model's PER. Data augmentation with synthetic repetitions (*+Rep*) is more efficient, yielding a 6.4% relative PER reduction. Combining both augmentations shows that their effects are complementary since it leads to the lowest PER, 19.9%, outperforming by 13.1% relative the Transformer baseline.

6. Discussion

As mentioned in section 3, the attention mechanisms are very flexible, which causes them to miss out on some typical beginning readers' disfluencies, such as repetitions. Additionally, the decoder can act like a language model, covering students' mistakes by recognising what should have been read. We offer a detailed study on the effectiveness of the various methods to improve the model's robustness on reading mistakes.

On Figure 1 are displayed the PER score for Transformer, Transformer+CTC and Transformer+CTC+Sub+Rep models, computed on words that are repeated, substituted or correctly read. We can first note that the scores for incorrect words are drastically higher than for correct words. This striking phenomenon is partly due to the presence of repetitions, substitutions, but also of other potential reading mistakes and slow reading rates since all are interlinked with a low reading level. CTC multi-objective training brings a significant improvement on repeated words (10.2% relative), and to a lesser extent on substituted words (1.9% relative), without degrading the performance on correct utterances. Data augmentation further reduces the PER both on repeated and substituted words, with 8.1% and 5.7% relative improvement over the Transformer+CTC. The improvement brought by data augmentation on correctly read words may be linked to a greater diversity in the training content, slightly breaking the decoder's language model effect.

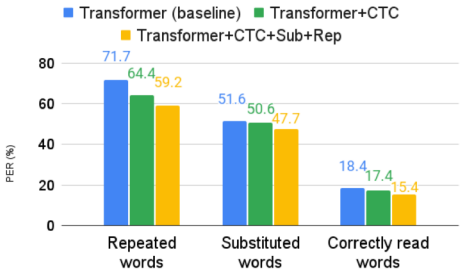

Figure 1: *PER (%) of models on repeated, substituted and correctly read words*

Figure 2 displays the attention weights and transcriptions obtained with the Transformer and Transformer+CTC+Rep models for an utterance where the child repeated a 4-phones word ([**papa** papa fɛ ʁəliʁ lili]). We can see that the Transformer+CTC+Rep model detects the repetition, while the Transformer does not. The parallel diagonals on the Transformer's weights, highlighted by two white lines, show that the model merges the two occurrences of the repeated word into one. On the contrary, the Transformer+CTC+Rep model's weights display a single diagonal, correctly detecting the two occurrences as separate. This success is most probably due to the CTC monotonic alignment constraint, supplemented by the synthetic repetitions, which the model learnt to apprehend.

Since the correct simulation of reading mistakes depends greatly on the quality of alignments, our augmented recordings could be refined by improving the performance of the model that generates the alignment. We could also upgrade the methods by simulating non-existing words that children could utter, through phone-level substitutions, insertions or deletions. Gathering more children training data would enable us to choose substituted words from the same speaker, which would reduce the risk of harming mismatches. Finally, we could tailor more

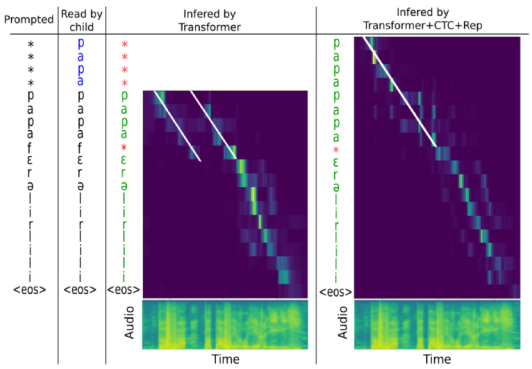

Figure 2: *Phone-level prompted text, actual text read by the child and transcriptions obtained with models Transformer and Transformer+CTC+Rep of the utterance: [papa fɛ ʁəliʁ lili]. The child's repetitions are in blue, and correct detections and errors in transcriptions in green and red, respectively. Attention weights captured between the models' encoders and decoders are displayed, as well as the spectrograms for time reference.*

the mistakes to our application by replacing the random choices by adapted-distribution-driven choices.

7. Conclusion

Read speech from young readers is very challenging to accurately recognise, as, in addition to child speech acoustic and prosodic characteristics, the model has to handle the presence of disfluencies and reading mistakes. This work offers various methods to adapt a Transformer acoustic model to speech from children learning to read. We use transfer learning (TL) for a baseline PER of 22.9%, which outperforms a TL-adapted DNN-HMM model by 21.0% relative. Multi-objective with CTC training improves the performance by 4.8% relative. Two original methods of data augmentation, consisting in simulating reading mistakes, show significant improvements over the baseline, reaching a PER of 19.9%. We analyse the model's behaviour on utterances that contain repeated words, a common mistake for beginner readers, and show that both the CTC multi-objective and the synthetic repetition augmentation help the model better recognise this type of disfluencies.

Future work will seek a better robustness to classroom babble noise, present in our datasets, which more than doubles the PER between clean and noisy recordings. Preliminary experiments were led and will be pursued: we obtained performance improvements (up to 11.7% relative) for very noisy recordings, but at the cost of significant deterioration (7.8% relative) on clean recordings. Finally, hybrid CTC/attention decoding [33] could bring improvement by taking advantage of the CTC function, which has proven useful through multi-objective training for better handling children reading mistakes.

8. Acknowledgements

We used the OSIRIM GPU platform, administered by IRIT and supported by CNRS, the Region Midi-Pyrénées, the French Government, ERDF (https://osirim.irit.fr/site/en).

9. References

[1] S. Lee, A. Potamianos, and S. S. Y. 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.

[2] R. Mugitani and S. Hiroya, "Development of vocal tract and acoustic features in children," *The Journal of the Acoustical Society of Japan*, vol. 68, no. 5, pp. 234–240, 2012.

[3] E. Fringi, J. F. Lehman, and M. J. Russell, "Evidence of phonological processes in automatic recognition of children's speech," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Dresden*, 2015, pp. 1621–1624.

[4] J. Mostow and G. Aist, "Evaluating tutors that listen: An overview of Project LISTEN," in *Smart machines in education: The coming revolution in educational technology.* The MIT Press, 2001, pp. 169–234.

[5] D. Bolaños, R. Cole, W. Ward, E. Borts, and E. Svirsky, "FLORA: Fluent oral reading assessment of children's speech," *ACM Trans. Speech Lang. Process.*, vol. 7, no. 4, p. 16, 2011.

[6] J. D. L. Proença, "Automatic assessment of reading ability of children," Ph.D. dissertation, University of Coimbra, 2018.

[7] E. Godde, G. Bailly, D. Escudero, M.-L. Bosse, and G. Estelle, "Evaluation of reading performance of primary school children: Objective measurements vs. subjective ratings," in *Proc. of the International Workshop on Child Computer Interaction (WOCCI)*, 2017, pp. 23–27.

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

[9] G. Yeung and A. Alwan, "On the difficulties of automatic speech recognition for kindergarten-aged children," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad*, 2018, pp. 1661–1665.

[10] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children English language learners," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Singapore*, 2014, pp. 1468–1472.

[11] A. Renduchintala, S. Ding, M. Wiesner, and S. Watanabe, "Multimodal data augmentation for end-to-end ASR," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad*, 2018, pp. 2394–2398.

[12] A. Potamianos and S. Narayanan, "Robust Recognition of Children's Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. November 2003, pp. 603–616, 2003.

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

[14] R. Serizel and D. Giuliani, "Deep neural network adaptation for children's and adults' speech recognition," in *Proc. of the Italian Computational Linguistics Conference (CLiC-it)*, 2014, pp. 137–140.

[15] F. Wu, P. Garcia, D. Povey, and S. Khudanpur, "Advances in automatic speech recognition for child speech using factored time delay neural network," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz*, 2019, pp. 1–5.

[16] S. Karita, X. Wang, S. Watanabe, T. Yoshimura, W. Zhang, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, and R. Yamamoto, "A Comparative Study on Transformer vs RNN in Speech Applications," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, no. April 2020, pp. 449–456, 2019.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. of the International Conference on Machine learning (ICML)*, 2006, pp. 369–376.

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. of the International Conference on Neural Information Processing Systems (NIPS)*, Cambridge, MA, USA, 2014, p. 3104–3112.

[19] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," in *Proc. of the International Conference on Neural Information Processing Systems (NIPS): Workshop on Deep Learning*, 2014, pp. 1–10.

[20] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, Attend and Spell: A neural network for large vocabulary conversational speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of the International Conference on Neural Information Processing Systems (NIPS)*. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[22] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888.

[23] Andrew, H. Sak, F. de Chaumont Quitry, T. Sainath, and K. Rao, "Acoustic modelling with CD-CTC-SMBR LSTM RNNS," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 604–609.

[24] S.-I. Ng, W. Liu, Z. Peng, S. Feng, H.-P. Huang, O. Scharenborg, and T. Lee, "The CUHK-TUDELFT system for the SLT 2021 Children Speech Recognition Challenge," *ArXiv preprint:2011.06239*, 2020.

[25] G. Chen, X. Na, Y. Wang, Z. Yan, J. Zhang, S. Ma, and Y. Wang, "Data augmentation for children's speech recognition – the "Ethiopian" system for the SLT 2021 Children Speech Recognition Challenge," *ArXiv preprint:2011.04547*, 2020.

[26] P. G. Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *ArXiv preprint:2102.09918*, 2021.

[27] A. Abad, P. Bell, A. Carmantini, and S. Renals, "Cross lingual transfer learning for zero-resource domain adaptation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6909–6913, 2020.

[28] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiát, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 521–527.

[29] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo, "Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 391–401, 2020.

[30] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[31] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz*, 2019, pp. 1408–1412.

[32] L. Gelin, M. Daniel, J. Pinquier, and T. Pellegrini, "End-to-end acoustic modelling for phone recognition of young readers," *ArXiv preprint:2103.02899*, 2021.

[33] T. Hori, S. Watanabe, and J. Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proc. of the Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 518–529.