9th Workshop on Speech and Language Technology in Education (SLaTE)
18-20 August 2023, Dublin, Ireland

Comparing phoneme recognition systems on the detection and diagnosis of reading mistakes for young children's oral reading evaluation

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

[1]Lalilo by Renaissance Learning, France
[2]IRIT, Paul Sabatier University, CNRS, Toulouse, France
{lucile.gelin, thomas.pellegrini, julien.pinquier}@irit.fr,
morgane.daniel@renaissance.com

Abstract

In the scope of our oral reading exercise for 5-8-year-old children, models need to be able to precisely detect and diagnose reading mistakes, which remains a considerable challenge even for state-of-the-art ASR systems. In this paper, we compare hybrid and end-to-end acoustic models trained for phoneme recognition on young learners' speech. We evaluate them not only with phoneme error rates but through detailed phoneme-level misread detection and diagnostic metrics. We show that a traditional TDNNF-HMM model, despite a high PER, is the best at detecting reading mistakes (F1-score 72.6%), but at the cost of low precision (73.8%) and specificity (74.7%), which is pedagogically critical. A recent Transformer+CTC model, to which we applied our synthetic reading mistakes augmentation method, obtains the highest precision (81.8%) and specificity (86.3%), as well as the highest correct diagnosis rate (70.7%), showing it is the best fit for our application.
Index Terms: child speech, misread detection and diagnosis, end-to-end, synthetic reading mistakes augmentation

1. Introduction

The speech of children aged 5 to 7 is subject to peculiarities linked to the growth of their speech production apparatus and to their poor body control: unstable articulatory mechanisms, intra- and inter-speaker spectral variability [1], higher fundamental and formant frequencies [2], phonological errors [3], etc. These morphological and phonological differences are the main reasons for the poor performance of automatic speech recognition (ASR) systems on children's voices.

Prior studies on ASR for child speech have indeed demonstrated that the performance is lower than for adult speech [4, 5, 6]. Due to limited available children data in most languages, deep neural network (DNN)-based systems only recently started to be exploited, with hybrid DNN-HMM (Hidden Markov Model) acoustic modelling approaches. For a children language learner application, [7] presents a DNN-HMM that, even trained on less data, surpasses Gaussian Mixture Models (GMM)-HMM systems. A factorised time-delay neural network (TDNNF-HMM) also shows to outperform GMM-HMM models for child speech recognition in [8]. Valuable insights on acoustic modelling for child speech recognition with DNN-HMM are given in [5]. End-to-end acoustic modelling for child speech recognition is not yet common due to limited available child speech data. In [9], 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 sequence-to-sequence (Seq2Seq) architectures for child speech recognition, is a new research subject, as shows their absence on a systematic review on child ASR published in 2022 [10], and the very recent communication of studies [11, 12] on this matter.

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 [13, 14, 15, 16]. 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: non-proficient readers' speech contains many disfluencies and reading mistakes that can be laborious to detect automatically [17, 18]. Nonetheless, our aim is specifically to be able to detect these reading mistakes, as well as to diagnose their nature: if the child inserts a phoneme, does the insertion constitute a familiar word? Do the most often deleted phonemes correspond to particularly hard-to-read graphemes? If two phonemes were confused, which ones? And most importantly, is the child making the same mistakes recurrently? The ASR system must be very precise on what it transcribes to identify the student's difficulties, provide reliable and relevant feedback, and efficiently help them in their learning.

In our previous studies, we were focused on improving our phoneme recognition system's ability to transcribe young readers' speech as precisely as possible. For this we compared several phoneme recognition systems based on their error rates. In this work, we push further the comparison done in [12] between several hybrid and end-to-end models, by evaluating how well they perform from a misread detection and diagnosis (MDD) point of view. We take inspiration from language learners mispronunciation detection and diagnosis studies [19, 20, 21]. We also apply our synthetic reading mistakes data augmentation method [22] to the best model, and examine the impact on its MDD performance. Our analysis shows that the different models display different strengths and weaknesses and brings valuable insights to choose the right system for a given application.

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 themself 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 misread words.

[1]https://www.lalilo.com/
[2]Corpus available at: https://voice.mozilla.org/fr

10.21437/SLaTE.2023-2

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. Young readers' speech is very difficult and costly to annotate due to the presence of reading mistakes. Our manual annotation process has two levels: 1) labeling the words as correct/incorrect 2) transcribing at the phoneme level if the word is incorrect. The correct words are automatically phonetized with a pronunciation dictionary. Annotations are done by two human judges, and recordings have been discarded in case of disagreement. The level 2 is the most difficult to attain, thus we have very few data containing reading mistakes annotated at this level. We prioritized including this data in the test set to evaluate our systems in real-life conditions.The training and validation sets contain respectively 13 and 0.41 hours of data that does not contain reading mistakes. The test set contains 0.48 hours of utterances, only sentences to ease the analysis, that may or may not contain reading mistakes.

3. Systems description

In this work we aim at comparing several systems: one hybrid and three end-to-end. Our systems are trained for phoneme recognition rather than word recognition through characters, which is mostly seen in end-to-end systems. For a reading assistant application, we prefer phonemes over words, the latter being less precise and making it difficult to handle phoneme-level reading mistakes that constitute non-existing words.

All models are trained with the same procedure: we first train a source model on the Common Voice adult speech dataset, then adapt it to child speech characteristics through transfer learning (TL). The transfer is done by retraining all layers with the Lalilo child speech dataset, which is advised in [5] for a dataset of 10+ hours of speech from very young children (5-8). All structures and hyperparameters are fully detailed in [12]. Models are trained on a single GTX 2080 Ti GPU.

3.1. TDNNF-HMM: the baseline

Hybrid approaches for automatic speech recognition consists in linking an HMM to a DNN. We use as a baseline a TDNNF-HMM [23] that was shown to be more efficient than a TDNN-HMM on a limited quantity of child speech data [8]. Training the source model on the Common Voice dataset, then the TL model on the Lalilo dataset, takes in total 48 hours. A single training of this system consumes approximately 5.2 kg CO2eq[3], which equals to the consumption of a 1-person car over 20 km.

3.2. RNN-CTC

The CTC paradigm, introduced by [24], discards the obligation of having an HMM by learning automatically alignments between the input and output sequences. Our first end-to-end model is named RNN-CTC and is composed of a simple encoder with Bidirectional Gated Recurrent Unit layers (BiGRU) and a CTC function. This model takes in total (source + TL) 28 hours to train, which consumes approximately 3 kg CO2eq per training[3].

3.3. LAS+CTC

Our second end-to-end model is a Seq2Seq model that follows the Listen, Attend and Spell (LAS) architecture [25]. It contains an encoder and a decoder that are based on Bidirectional Long-Short Term Memory (BiLSTM) layers and linked by an attention mecanism. Our LAS+CTC system combines a Cross-Entropy (CE) with a CTC loss function and uses a joint CTC/attention decoding, as proposed in [26]. Training the LAS+CTC takes 54 hours, consuming approximately 5.8 kg CO2eq[3].

3.4. Transformer+CTC

Presented by [27] and adapted to speech recognition by [28], the Transformer model follows a Seq2Seq encoder-decoder architecture, but relies solely on attention mechanisms, instead of recurrent neural networks in classical Seq2Seq systems. It is composed of self-attention-based encoder and decoder, which are linked by an attention module. In the same way as the LAS+CTC, it is trained with a multi-objective CE+CTC learning method and uses a joint CTC/attention decoding (proposed for Transformer architectures in [29]). Discarding the need for recurrent neural networks enables to compute dependencies between each pair of positions at once, instead of one by one. It allows for faster training in comparison with LAS systems. This model thus takes less time to train: 33 hours in total, which consumes approximately 3.6 kg CO2eq[3].

3.5. Synthetic reading mistake data augmentation

We observed in [12] that the TDNNF-HMM and Transformer+CTC models had difficulties accurately transcribing words that contain reading mistakes. It can be explained by the acoustic and linguistic characteristics being modified when a child misreads a word (slow speech rate, modified pronunciation, hesitations, etc). Another explanation is linked to the training set that contains only correctly read words: the model does not learn uncommon phoneme sequences that can be generated by reading mistakes. In the case of the Transformer+CTC, the decoder module in particular acts like a language model that favours known sequence of phonemes and thus tend to cover mistakes. We do not use a language model with the TDNNF-HMM to avoid this undesirable effect.

To counter it, we proposed in [22] a data augmentation method that consists in creating synthetic reading mistakes to train the model on more diverse content, including non-existing words. The method reduces the Phoneme Error Rate (PER) both on the whole test set and on a reduced set containing only words with reading mistakes, showing its efficiency on mitigating the implicit language model effect. In this work, we apply this technique to the model named Transformer+CTC+aug. As expected, applying it to the TDNNF-HMM did not bring significant improvement, and results will not be displayed here.

4. Metrics

Our models' objective is to detect phoneme-level reading mistakes that a child makes: we thus stay at the phoneme-level to evaluate their performance. Three types of phoneme sequences will be used:

- Prompted sequence: what the child was supposed to read;
- Uttered sequence: what the child actually read;
- Predicted sequence: what the ASR model transcribed.

The prompted sequence is phonetized automatically from words with a word-to-phoneme dictionary. It includes the liaisons between words that are featured in the French language. The uttered sequences are manually annotated at the phoneme-level.

[3]Calculator: `https://mlco2.github.io/impact`

4.1. Phoneme error rate

The traditional metric to assess an ASR model's performance is the WER. Here we compute PER values, by aligning uttered and predicted sequences and counting correct predictions, insertions, deletions and substitutions.

4.2. Misread detection rates

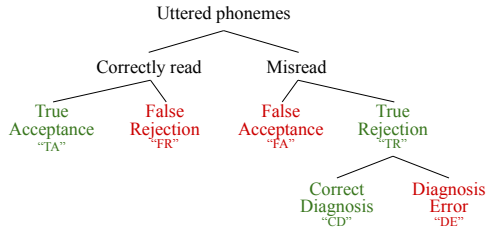

Figure 1: *Representation of MDD classification: in green (respectively, red), numbers to maximize (resp. minimize)*

To evaluate the capacity of a model to detect misread phonemes, we adapt word-level mispronunciation detection and diagnosis from second language automatic assessments [19, 20, 21] to phonemes. We compare the three phoneme sequences to obtain the number of true acceptances (TA), false rejections (FR), false acceptances (FA) and true rejections (TR) as shown in Fig. 1. The computation is done in several steps. An example is provided in Fig. 2 for better understanding.

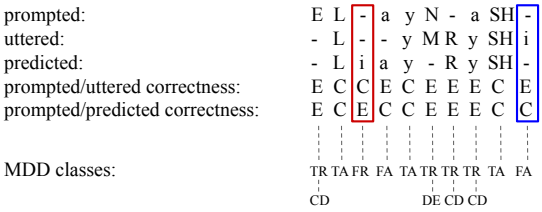

Figure 2: *Example of prompted, uttered and predicted phoneme sequences, correctness vectors and MDD classes for each phoneme, for the French sentence "elle a une hache"*

1. Alignment of the uttered and predicted sequences using the Levenshtein distance [30], adding of a blank phoneme "-" to model insertions and deletions;
2. Computation of a correctness vector between prompted and uttered sequences (prompted/uttered correctness in Fig. 2), that will describe whether each uttered phoneme is correctly read or not. We go through every operation of the alignment between the two sequences:
   - If the operation is "=", we label it as correct "C";
   - If the operation is a deletion or substitution, we label it as erroneous "E";
   - If the operation is an insertion, and the uttered phoneme is the blank phoneme, we label it as "C" (see red box in Fig. 2). It the uttered phoneme is not the blank phoneme, then it is labeled as "E" (blue box).

   Because prompted-uttered and prompted-predicted alignments sometimes do not match, we perform an additional automatic correction step that handles exceptional cases.
3. Computation of the correctness vector between prompted and predicted sequences, in the same way as the previous step;
4. Computation of the TA, FR, FA, TR values by comparing the two correctness vectors.

From these values, we compute i) the precision, that represents the proportion of phonemes detected as misreads that really were misread by the child ii) the recall, that corresponds to the proportion of misread phonemes that are detected as such iii) the specificity, that measures the proportion of phonemes detected as correct that really were correctly read by the child.

4.3. Diagnosis rates

Among the true rejections, we also want to differentiate the correct diagnosis (CD) from the diagnosis errors (DE), as shown in Fig.1. These values will measure our models' ability to diagnose the correct mistake. Starting again after step 3, we now compute the CD and DE numbers by taking only the true rejections and checking whether the ASR predicted the correct phoneme:

- If the child deletes a phoneme, the correct diagnosis corresponds to the ASR deleting the phoneme as well;
- If the child substitutes (respectively, inserts) a phoneme, the correct diagnosis corresponds to the ASR substituting (resp. inserting) the same phoneme as the child at the same location in the sequence.

From these numbers, we compute the CD and DE rates by dividing by the number of true rejections.

5. Evaluation

When computing MDD metrics, we encountered difficulties with utterances for which the predicted sequence was very different in length from the uttered sequence, with a lot of deletions and insertions. Our alignment algorithm was not capable of aligning the sequences properly and we were confronted to correctness vectors with different lengths, thus not comparable. We chose to manually remove any utterance for which the prediction of at least one model could not be aligned properly, discarding 69 utterances out of 353. We conducted a qualitative analysis on the corresponding recordings, and observed that the vast majority contain classroom noise or saturation or the child is whispering or speaking very loudly.

We display in Table 1 both the PER computed on the whole test dataset (353 utterances, results published in [6, 31]) and on the reduced dataset (284 utterances). The value between parentheses corresponds to the number of utterances for which the model's prediction could not be aligned and consequently were added to the list of utterances to discard for all models. We can see that the PER values on the reduced dataset are lower due to the discarding of hard to transcribe utterances. However, the ranking of models remains the same. The Transformer+CTC+aug model obtains the lowest PER value, with a 10.2% (9.5% on the reduced set) absolute improvement over the baseline hybrid TDNNF-HMM. We observe that the lower the PER is, the lower is the number of utterances for which the prediction was not good enough to be aligned. It suggests that we need to reduce further our model's PER to be able to compute MDD metrics on all utterances.

Table 2 displays the misread detection rates: precision (Prec), recall (Rec), specificity (Spec) and F1-score. The last column shows the number (#) and percentage (%) of correct diagnosis among true rejections. Results show that the TDNNF-

Table 1: *PER (%) obtained with different systems, tested on the test dataset and the reduced test dataset*

Model	Whole	Reduced (#utt)
TDNNF-HMM	28.9	26.3 (23)
RNN-CTC	31.0	28.5 (27)
LAS+CTC	23.2	20.5 (24)
Transf+CTC	19.6	17.9 (15)
Transf+CTC+aug	**18.7**	**16.8** (16)

HMM, despite its high PER, obtains the best F1-score. Among the end-to-end models, the LAS+CTC obtains the best F1-score, the RNN-CTC the best recall, and Transformer+CTC the best precision and specificity. By comparing the Transformer+CTC and Transformer+CTC+aug results, we can see that the synthetic reading mistakes data augmentation is highly effective for misread detection and diagnosis, since it improves each metric by 1.2 to 3.5%.

Applying the synthetic mistakes augmentation to the TDNNF-HMM did not bring any improvement on the PER nor on the MDD metrics, and is thus not relayed here. We explain this phenomenon by the fact that we do not use a language model for the TDNNF-HMM, while the Transformer+CTC model is subjected to an implicit language model effect, that tends to covering misread words by favouring existing words seen during training. The augmentation, that consists in showing reading mistakes during training, alleviates this undesirable effect, but is ineffective on the TDNNF-HMM.

Table 2: *Misread detection and diagnosis metrics (% or #) obtained with different systems, tested on the reduced dataset.*

Model	Prec	Rec	Spec	F1	CD (% / #)
TDNNF-HMM	73.8	**71.4**	74.7	**72.6**	68.4 / **357**
RNN-CTC	70.5	62.8	73.7	66.4	58.7 / 273
LAS+CTC	78.2	61.5	82.9	68.9	58.9 / 265
Transf+CTC	79.8	58.8	85.1	67.7	67.2 / 279
Transf+CTC+aug	**81.8**	61.5	**86.3**	70.2	**70.7** / 307

6. Discussion

We had shown in a previous study [6] that although the PER of the TDNNF-HMM is higher than the PER of the Transformer+CTC model, it was better at correctly transcribing words containing reading mistakes. This is due to the implicit language model effect conveyed by the encoder-decoder structure of the Transformer, that tends to cover reading mistakes. We see in this work that the TDNNF-HMM obtains a significantly better recall than the Transformer+CTC, which means that it is indeed better at detecting reading mistakes as such, and a better correct diagnosis rate, which confirms that it transcribes with better accuracy the erroneous phonemes.

Applying the synthetic mistake augmentation greatly improves the Transformer+CTC model's ability to correctly detect and diagnose the reading mistakes. It improves its F1-score by 2.5% absolute and its CD rate by 3.5%. These results corroborate our previous work [22, 6], where we had shown that the augmentation improves the Transformer+CTC model's PER on word containing reading mistakes. In terms of CD rate, the Transformer+CTC+aug outperforms the TDNNF-HMM by 2.3% absolute. However, since the latter has a better recall, it diagnoses correctly more phoneme-level mistakes (357 vs 307). This analysis brings depth to previous findings: the Transformer+CTC+aug obtains a better PER than the TDNNF-HMM on words containing reading mistakes, but still makes more diagnosis errors.

In the MDD point of view, we want to maximize the F1-score, that is based on the precision and recall. Our best model from this aspect is the hybrid TDNNF-HMM, which also obtains the best recall by far. However, this high detection of reading mistakes is made at the cost of having one of the lowest specificity and precision, which implies that a high proportion of correctly read phonemes are wrongly detected as misreads. In the pedagogical point of view, it is certainly important to detect as many mistakes as possible, but it is even more important to avoid giving children negative feedback when they have read correctly. It generates a lot of frustration, especially in a human-machine interaction where the child is autonomous, and can severely disrupt a child's learning. Our two main objectives are thus, in this order, to 1) minimize false rejections and 2) maximize true rejections. It implies that we want firstly to maximize precision, then specificity, and finally recall. The best model from this aspect is undoubtedly the Transformer+CTC+aug. It indeed displays the best precision and specificity by far, while ranking second in F1-score and third in recall.

Additionally, its good CD rate would enable us to bring valuable insights to teachers on the nature of reading mistakes their students need. For example, knowing that a given student has confused several times the sounds /b/, /p/, /d/, /k/ –which is common in French since the letters b, p, d, and q are mirrored letters– can encourage the teacher to offer remediation on this particular skill to the student. This knowledge can also be used in our platform's adaptive learning algorithm, that chooses the best reading exercise for each child, to prioritize exercises on that subject. We will therefore use the Transformer+CTC+aug model in our application to maximize its efficiency in helping children learning to read.

Finally, training a Transformer+CTC model consumes 1.6 kg CO2eq less than training a TDNNF-HMM. It might seem negligible, but is not since we usually train dozens of models to find the right hyper-parameters. It is thus an important parameter to take into account when doing ASR research.

7. Conclusions

Automatic speech recognition systems are usually evaluated through error rate measures, that assess the accuracy of the generated transcription. In the scope of our oral reading exercise for 5-8 year-old children, we aim at detecting and diagnosing the reading mistakes they make, to help them in their learning. Choosing the right model architecture for our application thus necessitates to evaluate models from a misread detection and diagnosis (MDD) point of view.

In this work, we compare several hybrid and end-to-end phoneme recognition system with traditional PER and MDD metrics and show that the different models display different strengths and weaknesses. The hybrid TDNNF-HMM obtains the worse PER but the best F1-score and recall, at the cost of a high proportion of false rejections, which is pedagogically critical. An end-to-end Transformer+CTC, to which we applied our innovative synthetic reading mistakes data augmentation technique, obtains the best precision and specificity, which are the metrics to maximize in the pedagogical point of view. It also displays the highest correct diagnosis rate, which will be useful to provide insightful remediation reports to teachers.

8. 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] 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.

[5] 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.

[6] L. Gelin, "Reconnaissance automatique de la parole d'enfants apprenant·e·s lecteur·ice·s en salle de classe : modélisation acoustique de phonèmes," Theses, Université Paul Sabatier - Toulouse III, Feb. 2022. [Online]. Available: https://theses.hal.science/tel-03715653

[7] 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.

[8] 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.

[9] 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.

[10] V. Bhardwaj, M. B. Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. Goud, A. Rehman, M. Shafiq, and H. Hamam, "Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review," *Applied Sciences*, vol. 12, no. 9, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/9/4419

[11] P. Gurunath Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech Language*, vol. 72, p. 101289, 2022.

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

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2017.

[20] W.-K. Leung, X. Liu, and H. Meng, "Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8132–8136.

[21] Y. Feng, G. Fu, Q. Chen, and K. Chen, "Sed-mdd: Towards sentence dependent end-to-end mispronunciation detection and diagnosis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3492–3496.

[22] L. Gelin, T. Pellegrini, J. Pinquier, and M. Daniel, "Simulating Reading Mistakes for Child Speech Transformer-Based Phone Recognition," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno*, 2021, pp. 3860–3864.

[23] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyberabad*, 2018, pp. 3743–3747.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals." *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966, doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

[31] L. Gelin, M. Daniel, T. Pellegrini, and J. Pinquier, "Reconnaissance de phones fondée sur du Transfer Learning pour des enfants apprenants lecteurs en environnement de classe," in *Conférence conjointe Journées d'Études sur la Parole (JEP), Traitement Automatique des Langues Naturelles (TALN), et Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)*, vol. 1. Nancy, France: ATALA, 2020, pp. 253–261. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02798545