Investigating Further Fine-tuning Wav2vec2.0 in Low Resource Settings for Enhancing Children Speech Recognition and Word-level Reading Diagnosis

Lingyun Gao, Cristian Tejedor-Garcia, Catia Cucchiarini, Helmer Strik
Centre for Language Studies
Radboud University, Nijmegen, the Netherlands
{lingyun.gao, cristian.tejedorgarcia,
catia.cucchiarini, helmer.strik}@ru.nl

Abstract

Automatic reading diagnosis systems can substantially improves teachers' efficiency in scoring reading exercises and provide students with easier access to reading practice and feedback. However, few studies have focused on developing Automatic Speech Recognition (ASR)-based reading diagnosis systems due mainly to scarcity of data. This study explores the effectiveness and robustness of further fine-tuning the Wav2vec2.0 large model in low-resource settings, for recognizing children speech and detecting reading miscues using target domain and similar out-of-domain data. Our results show a word error rate (WER) of 10.9% and an F1 score of 0.49 for reading miscue detection achieved by our best fine-tuned model training with target domain data, while using similar out-of-domain non-native read speech can enhance the model performance for unseen speakers and noisy settings. The analyses provide insights into the robustness of further fine-tuning strategies on the Wav2vec2.0 model.

1 Introduction

Recent advances in Automatic Speech Recognition (ASR) have made many previously complex speech–based computer-assisted applications more feasible (Ivanko et al., 2023). One such application is the integration of ASR into primary school reading education (Shadiev and Liu, 2023). However, this initiative has encountered significant challenges in children speech recognition (Feng et al., 2024) and miscue detection (Shivakumar and Narayanan, 2022), largely due to the scarcity of speech data and annotations, especially for languages other than English.

Meanwhile, growing concerns over declining reading proficiency levels among low-resource language users (Swart et al., 2023) highlight the urgent need for innovative approaches to improve reading instruction. A common task in reading education is miscue detection (Limonard et al., 2020), which involves two steps: first, identifying general reading errors such as word substitution, insertion, and deletion, and second, classifying specific miscues, including various types of substitution and insertion errors (Shivakumar and Narayanan, 2022). This process requires the manual transcription of mispronunciations and the annotation of miscue categories, making data collection both time-consuming and costly.

State-of-the-art (SOTA) large pretrained ASR models have shown remarkable performance in adult speech recognition (Pratap et al., 2024) and offer potential for supporting practice and remedial teaching (Molenaar et al., 2023) in low-resource children's reading education. Wav2Vec 2.0-CTC and similar models are especially promising due to their ability to detect spoken errors more accurately (Gao et al., 2024). Research has also shown Wav2Vec 2.0's effectiveness in low-resource transfer learning tasks, improving children's speech recognition in English through fine-tuning pretrained models (Bartelds et al., 2023; Jain et al., 2023). For Dutch child speech, data augmentation with cross-lingual speaker diversity has proven effective, though it mainly benefits unseen speaker recognition (Zhang et al., 2024). However, these methods require significant computational resources and training data. Given the high cost of child speech data collection and ASR model training, further fine-tuning (Shen et al., 2021) trained ASR models offers a low-cost alternative by leveraging knowledge from adult speech, making it particularly suitable for low-resource languages. Moreover, previous research has shown that speech data from similar domains

Proceedings of AAAS Workshop 2025 – Automatic Assessment of Atypical Speech, pages 1–6
March 5, 2025 ©2025 University of Tartu Library

can be effectively used as augmentation for target domain speech recognition. In (San et al., 2024), training ASR with speech data from similar languages or accents has been found to improve target language speech recognition in low-resource settings. Nevertheless, further finetuning and augmentation with similar domain data has not been extensively explored in the context of Dutch children speech recognition and the impact of this method on downstream reading diagnosis.

In addition, most existing fine-tuning and augmentation studies have employed clean datasets, often collected in laboratories, while real-world child reading exercises typically take place in home and classrooms with diverse background noise and other environmental factors (Lavechin et al., 2020). The robustness of these strategies on ASR for real-world Dutch children's read speech remains unclear.

In this work, we make a novel contribution by filling the gap of investigating the effectiveness of further fine-tuning Dutch adult speech trained Wav2vec2.0, using target domain and similar out-of-domain data, for Dutch native child read speech recognition and reading miscue detection. Additionally, we address the research gap of exploring robustness of further fine-tuning in diverse real-world reading tasks and context where Dutch primary pupils read aloud. The research questions we address in our study are:

RQ1: *To what extent can low-resource further fine-tuning of the adult-speech trained Wav2vec2.0 model enhance the performance of Dutch native children's read speech recognition?*

RQ2: *To what extent can similar out-of-domain (native child dialogue and non-native child read speech) data used in further fine-tuning improve target-domain Dutch native children's read speech recognition?*

RQ3: *To what extent can the above mentioned further fine-tuning strategies enhance Dutch children's reading miscue detection?*

RQ4: *To what extent are the above-mentioned further fine-tuning strategies robust to real-world Dutch native children's read speech recognition?*

We address our research questions through a two-phase study. In the first phase, we explore the efficacy of different fine-tuning options on clean child read speech recognition. In the second phase, we select models representing effective training strategies for experiments on investigating the robustness of real-world child read speech and their ability to detect children's reading miscue.

Table 1: Reading error categories and reading miscue categories with their abbreviations in brackets

Error Type	**Reading Miscue Type**	**Other**
Substitution	Substitute a word in the prompt by another existing dutch word which was semantically identical (**SS**)	-
	Substitute a word in prompt by another existing dutch word which was orthographically similar (**OS**)	-
	Replace a word in prompt by another existing dutch word which was not orthographically or semantically similar (**O**)	-
Insertion		Restart
	Insertion of an extra word not in the prompt ($\boldsymbol{I_m}$)	-
Deletion	a word in the prompt is not read (**D**)	-

2 Methodology

2.1 Dataset and Preprocessing

This paper utilizes clean children speech from the Jasmin-CGN Corpus (Cucchiarini et al., 2008), and real-world Dutch child read speech from DART (Bai et al., 2021) and ST.CART (Wills et al., 2023). In the Jasmin-CGN Corpus, the target domain data, the native children read speech subset, includes recordings of 71 primary school children (ages 6-13, reading level 1-9) reading aloud at their mastery reading level, aligned with manual orthographic transcriptions. Children of the same reading level share the same reading prompt. Each prompt consists of three stories. The recordings of the first prompt story are aligned with the prompt text, reading miscue, and reading strategy annotations (data description available in (Limonard et al., 2020)). The native child dialogue speech and non-native child reading speech are used as similar out-of-domain data for augmentation. The native child dialogue speech consists of recordings of the same 71 speakers. The non-native child read speech consists of read speech from 53 non-native primary school children.

To investigate the impact of fine-tuning with different data on recognizing child speech, we split the Jasmin dataset, as shown in Table 2, into validation, training, and testing subsets. We created two child speech test sets: the full test set and the non-overlap test set. The full test set includes speakers overlapping with those in the training data, while the non-overlap test set consists of independent speakers. These test sets allow us to assess the ability of fine-tuning to handle unseen

Table 2: Data split details for Jasmin-CGN Corpus

Dataset Split	Content	Duration
Validation Set	Read Speech (story 2&3): First five samples from each of 65 native speakers.	31 minutes
Train:clean-FULL	Read Speech (story 2&3): Remaining sentences from 65 native speakers after validation samples are excluded.	4.4 hours
Train:clean-aug-nonna	Augmented set including native and non-native read speech.	8 hours
Train:clean-aug-dial	Augmented set including native read and child-only dialogue speech.	8 hours
Train:clean-SDS	Read Speech (story 2&3): Samples from the sentence order 8th to 20th	∼1 hour
Train:clean-SPDS	Read Speech (story 2&3): Random selection of sentence samples from 65 speakers emphasizing mispronunciations.	∼1 hour
Test:clean (Full)	Read Speech (story no.1): First prompt readings from all 65 speakers (71 recordings).	2.05 hours
Test:clean (Non-overlap)	Read Speech (story no.1): First prompt readings from six other speakers, avoiding overlap with the 65 speakers.	14 minutes

speakers.

Real-world speech recordings are more complex than speech recorded in a controlled lab setting, as it includes diverse speaking conditions and environmental noise. In this paper, we use the following two datasets to represent real-world speech. The DART test dataset consists of children reading speech recorded at home, primarily featuring environmental noise from different microphones and parents' voices. The ST.CART test dataset consists of children's reading speech recorded in a classroom, mainly including background noise from other children talking and reading. In both real-world datasets, usually the volume of speech is less well-controlled, and children are less attentive, leading to greater variation in speech speed compared to recordings made in a lab.

For evaluating fine-tuning robustness, we used three real-world testsets from DART and ST.CART. The DART test dataset, with 48 minutes of Dutch children reading sentences and stories at home, assesses robustness on real-world data seen during validation, but not included in training. The validation set includes 3 minutes of sentence recordings and 2.5 minutes of story recordings, while the DART testset includes 15 minutes of sentence recordings and 33 minutes of story recordings. The ST.CART testset, consisting of 36 minutes of Dutch children reading stories in classrooms, evaluates robustness on real-world data that was not seen during any training phase.

2.2 Reading Miscue Detection

In this paper, word-level reading errors include substitutions, insertions, and deletions. Word-level reading miscues, which are a subset of these errors, encompass specific substitution and insertion errors, as detailed in (Limonard et al., 2020) and shown in Table 1. Insertions in reading miscues are a subset of insertion errors, but correct readings after restarts or repetitions are not classified as insertion miscues, in line with Dutch reading test conventions (van Til et al., 2018).

For evaluating fine-tuned models, we focus on detecting word-level reading miscues, defined as errors where both the type and location match between prediction and ground truth. Analysis is based on detected general errors from manual transcriptions and ASR outputs, with miscue categorization outlined in Table 1. We follow the steps in section 2.3 of (Gao et al., 2024) to obtain and evaluate miscue labels.

2.3 ASR Models, Metrics and Tools

We evaluate the effectiveness of further finetuning ASR models in recognizing Dutch native children speech and detecting word-level reading miscues. The ASR foundation model Wav2vec2.0 we used in this paper is pretrained and finetuned with Dutch adult speech. We would like to further finetune the ASR model with Dutch child. ASR models are employed to predict word-level transcriptions, coupled with the Speech Recognition Toolkit SCTK (Lütkebohle, 2021). We employ Word Error Rate (WER) for evaluating children speech recognition at each testset and Precision, Recall, F1 for reading miscue detection evaluation, similarly in our previous work(Gao et al., 2024).

We (further) fine-tune the Wav2vec2.0 large model on different training dataset sourced from the Jasmin-CGN Dutch children speech. Our (further) fine-tuning experiments use hyperparameters similar to those reported by (Baevski et al., 2020) for comparable data sizes. In order to train models on a single A6000 GPU, following training settings in (Bartelds et al., 2023), We train the models with a batch size of 4 or 8 and apply gradient accumulation steps of 8 or 4, respectively, over 10k steps, using a learning rate of 1e-5 and a single seed.

3 Results

3.1 Performance on Different Fine-Tuning Options

We address RQ1 and RQ2 by evaluating further fine-tuning strategies on Wav2vec2.0 for children speech recognition performance, measured by WER, using different training datasets. The results are shown in Table 3. The baseline model is the Dutch adult pretrained Wav2vec2.0 fine-tuned on adult read speech, without further finetuning on child data.

Table 3: Evaluation of children speech recognition by WER of the baseline and fine-tuned models with different training sets.

Model	test-clean	
	full	non-overlap
RQ1		
pretrain-adult-ft-adult	13.2	13.9
pretrain-adult-ft-adult-clean-FULL	**10.9**	12.2
pretrain-adult-ft-adult-clean-SPDS	11.1	11.7
pretrain-adult-ft-adult-clean-SDS	11.3	11.7
RQ2		
pretrain-adult-ft-adult-clean-aug-dial	11.4	12.1
pretrain-adult-ft-adult-clean-aug-nonna	11.5	**11.2**

For RQ1, our results highlight that fine-tuning with a small dataset of Dutch native children's read speech (clean-FULL: 4.4 hours) can substantially enhance the model's accuracy for this demographic (WER=10.9%, absolute improvement=2.3%, Table 3). Meanwhile, using a speaker-prompt train subset SPDS can achieve competitive performance overall (WER=11.1% vs 10.9%) and improve results for unseen speakers (WER=11.7% vs 12.2%), underscoring the importance of data diversity in fine-tuning rather than sheer volume.

For RQ2, our findings show that further fine-tuning with non-native read speech augmentation improves recognition for unseen child speakers (best WER=11.2%, compared to 12.2% with clean-FULL fine-tuning on test-clean non-overlap), emphasizing the benefit of increased speaker diversity through out-of-domain data. However, on the test-clean full set, where speakers largely overlapped with the training data, neither fine-tuning with native dialogue augmentation (WER=11.4%) nor non-native speech data (WER=11.5%) improved performance over the clean-FULL model (WER=10.9%), as shown in the bottom part of Table 3, suggesting a significant domain transfer gap between dialogue and read speech for training ASR models, consistent with (Proença et al., 2018).

3.2 Detection of Reading Miscues

Then, to address RQ3, we compare precision, recall, and F1 scores of different models for miscue detection in Table 4. Our results confirm the effectiveness of low-resource fine-tuning with target-domain read speech and one out-of-domain (non-native) data in improving Dutch children's miscue detection. The best detection performance on the full testset was achieved by further fine-tuning with clean-full (F1=0.49), while the non-overlap testset was best handled by fine-tuning with clean-aug-nonna (F1=0.57), compared to 0.43 and 0.44 respectively for the baseline model. This trend mirrors WER improvements, indicating a strong correlation between speech recognition performance and miscue detection.

3.3 Robustness to Real-World Data and Reading Tasks

To address RQ4, we compare the WER performance of models trained with different strategies against a baseline model without further fine-tuning across three real-world testsets, as shown in Table 5. Our findings indicate that further fine-tuning strategies show limited robustness to unseen real-world data, as all fine-tuned models performed worse in these cases. In particular, the fine-tuning strategies, ft-adult-clean-FULL and ft-adult-clean-aug-dial, substantially improve WER on datasets similar to the training data (50.4% and 50.7%, respectively). On the unseen DART story testset, ft-adult-clean-aug-dial achieves the best performance with a WER of 39.0%. However, these models face challenges in generalizing to the ST.CART story test set. All fine-tuned models, including ft-adult-clean-FULL and ft-adult-clean-aug-nonna, underperform compared to the baseline (WER = 39.5%). This highlights a trade-off between in-domain optimization and broader generalization, as fine-tuning on small, clean datasets tends to reduce the model's ability to generalize effectively. Despite this, the results suggest that incorporating a limited amount of real-world data into the validation set can enhance the effectiveness of fine-tuning. Specifically, the strategy involving dialogue augmentation demonstrated the highest robustness among the various fine-tuning approaches.

Table 4: ASR model performance in reading miscue detection, evaluated by precision (P), recall (R), and F1 on the full testset and speaker-independent subset, with F1 for each miscue category in Table 1.

Model	**All miscues**				I_m	**D**	**OS**	**SS**	**O**
	P(full)	R(full)	F1(full)	F1(non-overlap)	F1	F1	F1	F1	F1
pretrain-adult-ft-adult	0.29	0.83	0.43	0.44	0.63	**0.59**	0.17	0.39	0.28
pretrain-adult-ft-adult-clean-FULL	**0.35**	**0.83**	**0.49**	0.54	0.71	**0.59**	**0.22**	0.4	**0.35**
pretrain-adult-ft-adult-clean-SPDS	0.34	0.83	0.48	0.55	**0.74**	0.57	0.20	**0.42**	0.33
pretrain-adult-ft-adult-clean-aug-dial	0.33	0.82	0.47	0.51	0.67	0.56	0.21	0.4	0.33
pretrain-adult-ft-adult-clean-aug-nonna	0.33	0.83	0.47	**0.57**	0.73	0.54	0.21	0.36	0.32

Table 5: ASR model evaluated by WER in three real-world test dataset

Model	DART		ST.CART
	sentence	story	story
ft-adult	62.4	53.2	**39.5**
ft-adult-clean-FULL	**50.4**	39.8	48.6
ft-adult-clean-SPDS	53.8	43.0	43.7
ft-adult-clean-aug-dial	50.7	**39.0**	44.5
ft-adult-clean-aug-nonna	55.2	40.1	51.4

4 Conclusion

This study demonstrates the potential of further fine-tuning the Wav2vec2.0 large model with domain-specific data to improve read speech recognition in Dutch native children. It highlights the effectiveness of augmenting training data with similar out-of-domain data, especially for unseen speakers in clean settings and real-world scenarios if a small amount of real-world audio can be utilized for validation.

Acknowledgments

This work was supported by the NWO research programme AiNed Fellowship Grants under the project Responsible AI for Voice Diagnostics (RAIVD) - NGF.1607.22.013.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Yu Bai, Ferdy Hubers, Catia Cucchiarini, and Helmer Strik. 2021. An asr-based reading tutor for practicing reading skills in the first grade: Improving performance through threshold adjustment. In *IberSPEECH 2021*, pages 11–15.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the ACL*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Catia Cucchiarini, Joris Driesen, Hugo Van hamme, and Eric Sanders. 2008. Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567.

Lingyun Gao, Cristian Tejedor-Garcia, Helmer Strik, and Catia Cucchiarini. 2024. Reading miscue detection in primary school through automatic speech recognition. In *Interspeech 2024*, pages 5153–5157.

Denis Ivanko, Dmitry Ryumin, and Alexey Karpov. 2023. A review of recent advances on deep learning methods for audio-visual speech recognition. *Mathematics*, 11(12):2665.

Rishabh Jain, Andrei Barcovschi, Mariam Yahayah Yiwere, Dan Bigioi, Peter Corcoran, and Horia Cucu. 2023. A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition. *IEEE Access*, 11:46938–46948.

Marvin Lavechin, Ruben Bousbib, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2020. An open-source voice type classifier for child-centered daylong recordings. In *Interspeech*.

S. Limonard, Catia Cucchiarini, R.W.N.M. van Hout, and Helmer Strik. 2020. Analyzing read aloud speech by primary school pupils: Insights for research and development. In *Interspeech 2020*, pages 3710–3714.

Ingo Lütkebohle. 2021. The nist speech recognition scoring toolkit (sctk) 2.4.12. https://github.com/usnistgov/SCTK. [Online; accessed 10-March-2024].

Bo Molenaar, Cristian Tejedor-Garcia, Catia Cucchiarini, and Helmer Strik. 2023. Automatic Assessment of Oral Reading Accuracy for Reading Diagnostics. In *Proc. INTERSPEECH 2023*, pages 5232–5236.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Jorge Proença, Carla Lopes, Michael Tjalve, Andreas Stolcke, Sara Candeias, and Fernando Perdigao. 2018. Mispronunciation detection in children's reading of sentences. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(7):1207–1219.

Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams, and Dan Jurafsky. 2024. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 100–112, St. Julian's, Malta. Association for Computational Linguistics.

Rustam Shadiev and Jiawen Liu. 2023. Review of research on applications of speech recognition technology to assist language learning. *ReCALL*, 35(1):74–88.

Zhiqiang Shen, Zechun Liu, Jie Qin, Marios Savvides, and Kwang-Ting Cheng. 2021. Partial is better than all: Revisiting fine-tuning strategy for few-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9594–9602.

Prashanth Gurunath Shivakumar and Shrikanth Narayanan. 2022. End-to-end neural systems for automatic children speech recognition: An empirical study. *Computer Speech & Language*, 72:101289.

Nicole Swart, Joyce. Gubbels, Melissa in 't Zandt, Maarten Wolbers, and Eliane Segers. 2023. PIRLS-2021: Trends in leesprestaties, leesattitude en leesgedrag van tienjarigen uit Nederland.

Alma van Til, Frans Kamphuis, Jos Keuning, Martine Gijsel, Judith Vloedgraven, and Anja de Wijs. 2018. Wetenschappelijke verantwoording lvs-toetsen dmt. *Arnhem: Cito*.

Simone Wills, Cristian Tejedor-Garcia, Catia Cucchiarini, and Helmer Strik. 2023. Enhancing asr-based educational applications: Peer evaluation of non-native child speech. In *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 16–20.

Yuanyuan Zhang, Zhengjun Yue, Tanvina Patel, and Odette Scharenborg. 2024. Improving child speech recognition with augmented child-like speech. In *Interspeech 2024*, pages 5183–5187.