ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889916

Speech Recognition for Automatically Assessing Afrikaans and isiXhosa Preschool Oral Narratives

Christiaan Jacobs*, Annelien Smith*, Daleen Klop*, Ondřej Klejch†, Febe de Wet*, Herman Kamper*
*Stellenbosch University, South Africa
†University of Edinburgh, United Kingdom

***Abstract*—We develop automatic speech recognition (ASR) systems for stories told by Afrikaans and isiXhosa preschool children. Oral narratives provide a way to assess children's language development before they learn to read. We consider a range of prior child-speech ASR strategies to determine which is best suited to this unique setting. Using Whisper and only 5 minutes of transcribed in-domain child speech, we find that additional in-domain adult data (adult speech matching the story domain) provides the biggest improvement, especially when coupled with voice conversion. Semi-supervised learning also helps for both languages, while parameter-efficient fine-tuning helps on Afrikaans but not on isiXhosa (which is under-represented in the Whisper model). Few child-speech studies look at non-English data, and even fewer at the preschool ages of 4 and 5. Our work therefore represents a unique validation of a wide range of previous child-speech ASR strategies in an under-explored setting.**

***Index Terms*—child speech recognition, low-resource languages, spoken language assessments, oral narratives**

I. Introduction

Less than 20% of South African 10-year olds can read for meaning [1]. But problems often start much earlier. Oral language skills provide the foundation for literacy and language development in preschool and predict later literacy and reading abilities [2]–[4]. While literacy and reading are essential, these skills cannot be assessed in preschool children aged 4 to 5. However, narrative and storytelling skills can be evaluated at this stage. Using narrative assessments to identify children at risk for literacy and reading difficulties early on could enable timely interventions, improving children's chances of later success [5]–[8]. To make such assessments possible in overcrowded preschool classrooms, this current work forms part of an initiative to develop automated systems for assessing oral narratives of preschool children in South Africa.

The first step of a spoken language assessment system [9]–[11] is to transcribe child speech using an automatic speech recognition (ASR) system. This paper describes the development of ASR systems on stories from isiXhosa and Afrikaans preschool children. ASR for child speech is difficult due to acoustic and linguistic variability [12], [13]. But, compared to previous work, our setting is even more challenging. First, most studies focus on older children (ages 7 and up) [14], [15], while we are targeting ages 4 and 5; it has been shown that ASR performance deteriorates dramatically for younger children [16]. Second, because many studies focus on early reading ability, few consider spontaneous speech. Our data consists of spontaneous speech, obtained by prompting a child with a picture sequence (Fig. 1). Third, almost all previous work is on English, with only a handful of publicly available non-English datasets [17]. We consider an extremely low-resource setting where only 5 minutes of transcribed in-domain speech data is available; developing ASR systems without relying on extensive labelled child speech data is crucial for making automated assessment possible in more languages.

We gratefully acknowledge the British Academy for their research grant (ECE 190079) and Fab Inc. for their research grant (W1/01B). We would like to also thank Julian Herreilers, Retief Louw, Emma Sharratt, Luke Crowley and Shelley O'Carroll for helpful suggestions and assistance with the data.

In this unique setting, the question is how we should develop our ASR systems. There is no shortage of strategies, but their benefits are often inconclusive. For instance, [18] reports voice conversion can be a useful data augmentation approach in some settings but not in others. To give support for or against a particular strategy, more data points are required. This paper describes a practical case study that provides two data points (Afrikaans and isiXhosa) for confirming or refuting strategies that have been proposed in the past for processing child speech. Our contribution is therefore not in promoting some new method, but in applying a range of strategies (wider than in most other work) to a particular real-world problem.

We use Whisper [19] as a starting point – a good base model for fine-tuning on transcribed child speech [20]. We then consider the following strategies: (1) parameter efficient fine-tuning [21]; (2) incorporating adult data from the same language to compensate for the absence of child data [22]; (3) using voice conversion as a data augmentation approach to generate more child-like speech [23]; and (4) semi-supervised learning for labelling untranscribed in-domain child speech that might be available [24]. Of these strategies, we find that using in-domain adult speech – where an adult produces stories matching our narrative domain – gives the biggest gain in both Afrikaans and isiXhosa. Semi-supervised learning also helps. Voice conversion and parameter-efficient fine-tuning are only beneficial in specific cases.

II. A Dataset for Oral Narrative Assessment

To develop our ASR systems, we use recordings of oral narrative assessments performed on Afrikaans- and isiXhosa-speaking preschool children in South Africa's Western Cape province. The assessment followed the Multilingual Assessment Instrument for Narratives

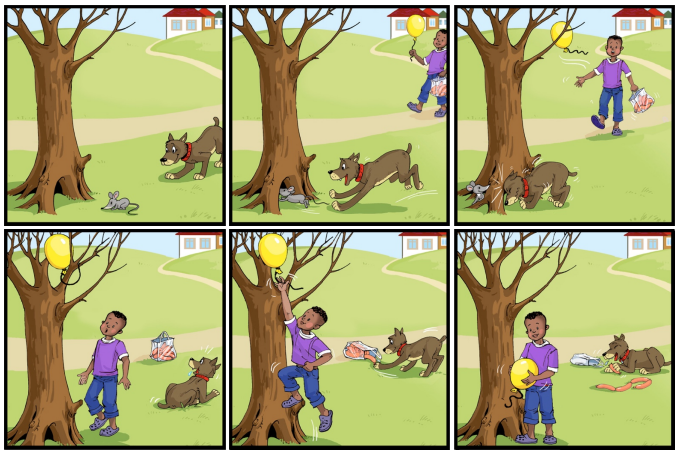

Fig. 1: A picture sequence used in the MAIN assessment protocol to elicit a narrative from a child.

TABLE I: Details of the oral narrative dataset, collected from preschool children aged 4 to 5.

	Afrikaans			isiXhosa		
Data	Mins.	Spks.	Utts.	Mins.	Spks.	Utts.
Train	256	120	5187	228	118	5289
Development	47	19	890	40	19	907
Test	30	14	680	35	14	824
Train 5m	5	5	91	5	5	82

TABLE II: WERs (%) on development data for Afrikaans and isiXhosa, trained on in-domain child speech.

	Afrikaans		isiXhosa	
Model	small	medium	small	medium
5m-child	54.4	47.4	**86.0**	**80.4**
5m-child (LoRA)	**52.1**	**41.1**	92.4	88.7
Baseline: Whisper w/o fine-tuning	90.3	88.1	107.7	105.6
Topline	21.3	18.3	54.4	50.9

(MAIN) protocol [25]: a facilitator presents a series of pictures to a child, who is then asked to verbalise a story based on the events illustrated in the pictures. The facilitator then asks the child questions to test their comprehension. MAIN was designed to be ecologically valid and culturally neutral, allowing it to assess children's narrative production and comprehension skills regardless of linguistic, socioeconomic or cultural backgrounds. Fig. 1 illustrates one of the picture sequences used. Based on a transcript of the session, a child is then scored using metrics that assess narrative comprehension and the structural complexity of the stories they produce.[1]

In total, 154 Afrikaans-speaking and 157 isiXhosa-speaking children, aged 4 to 5, were randomly selected from 55 classrooms to particpate in the study. Most recordings were sufficiently audible, but some contained noises such as vacuuming, background chatter and microphone fiddling, despite using a directional, noise-cancelling microphone. We applied minimal filtering since the recordings match the conditions under which our ASR models will operate. More details on the data are given in [26].

To prepare the data for ASR, we manually aligned the provided transcripts to the raw recordings using the Praat toolkit [27]. This resulted in roughly 5 hours of child-speech data for each language. We refer to this as our in-domain data. The data for each language was split into a train, development and test set, as shown in Table I. To develop ASR models for oral narrative assessments in more languages, collecting and annotating the amount of data presented in Table I is not feasible. In our experiments, we therefore explore an extremely resource-limited scenario by sampling a random 5-minute subset from the full training set (last row, Table I). We treat this 5-minute training set as the only labelled in-domain child speech for model development.

III. Base Speech Recognition Model: Whisper

We use Whisper [19] as our base ASR model since it gave good performance in previous studies on child ASR [20], [28], [29]. Whisper has five model sizes from which we only consider the small and medium variants due to hardware constrains. All model sizes were trained on 680k hours of weakly-supervised speech data from 97 languages (117k hours are non-English). The training data includes 4.1 hours of Afrikaans, but also thousands of hours from related Germanic languages such as Dutch and German. On the other hand, no isiXhosa or any closely related languages are included in Whisper. We select Swahili as the language token when fine-tuning on isiXhosa.

In our experiments, we fine-tune Whisper for up to 5k steps with a batch size of 64 and a learning rate of $1 \cdot 10^{-5}$. Without having additional text data to apply an external language model (LM), we benefit from Whisper's implicitly learned LM in final decoding with a beam size of 10. We apply a length penalty of -0.5 to penalise repetitive hallucinations caused by stuttering and false starts.[2]

IV. Experiments

We explore different strategies to build and improve our ASR models. While some of these have been effective in child-speech ASR in prior work, we now evaluate them in our own unique low-resource scenario. Before considering each strategy in turn, we look at the results of our base models, which are trained by fine-tuning Whisper (small and medium) using 5 minutes of transcribed in-domain child speech (last row, Table I).

The word error rates (WERs) for these models are presented in the first row of Table II. Despite using the same amount of training data, the WER for Afrikaans is significantly better than for isiXhosa: 47.4% compared to 80.4% for Whisper medium. This can partly be attributed to the mismatch between isiXhosa and the languages included in Whisper's pretraining data. Moreover, a previous study on adult ASR [31] found that it is more challenging for ASR models to transcribe agglutinative languages like isiXhosa than Germanic languages like Afrikaans. The character error rate (CER) for these 5-minute models are 27.9% for Afrikaans and 34.0% for isiXhosa (not shown in the table); this smaller difference in error rate shows that isiXhosa is more severely penalised for inaccurate word predictions.

To situate the results that follow, we consider lower and upper bounds. Representing the case where no child speech data is available, we apply Whisper without any fine-tuning. This baseline produces nonsensical results for isiXhosa with a WER of 105.6%, whereas for Afrikaans the results are poor but not random with a WER of 88.1% (third row, Table II, Whisper medium). Topline results (when fine-tuning on the full training sets) are shown in the last row of Table II.

We now consider strategies to improve our 5-minute models.

A. Does parameter-efficient fine-tuning help?

Related work: Low-rank adaptation (LoRA) is a method where a network is efficiently updated without having to retrain all the network parameters [32]. Apart from computational efficiency, LoRA has also resulted in better performance by preventing overfitting [21]. E.g. [28] achieved better child-speech ASR performance with LoRA fine-tuning than updating all weights. In contrast, in [20] the authors fine-tune Whisper for English child speech and find that LoRA is not always beneficial. Given these inconsistencies, we provide two additional data points to validate the effectiveness of LoRA for child-speech ASR.

[1]isiXhosa and Afrikaans are two of South Africa's 12 official languages, with respectively 8M and 7.2M native speakers. isiXhosa is a Southern Bantu language and has a very rich system of agglutinating morphology. Afrikaans is a West Germanic language that evolved from the 17th-century Dutch spoken by settlers and slaves in the Cape Colony. Both languages use the Latin alphabet.

[2]We also performed development experiments using XLSR [30] for ASR. Here we experimented with an external LM trained on the transcriptions of the full training set while using the 5-minute subset for the acoustic model. The LM provided benefits with XLSR, but the absolute performance was always worse than using Whisper with its implicit LM learned in its decoder.

Setup: LoRA inserts learnable decomposition matrices while freezing the original model parameters. We specifically add low-rank matrices with $r = 32$ to the attention layers $\{\mathbf{W}_q, \mathbf{W}_\text{k}, \mathbf{W}_\text{v}, \mathbf{W}_\text{out}\}$ with a learning rate of $1 \cdot 10^{-4}$. We fine-tune using the 5-minute in-domain sets.

Results: Comparing the LoRA fine-tuned variants (second row, Table II) to the full-parameter fine-tuned models (first row), we see that for Afrikaans LoRA helps by improving WER from 47.4% to 41.1% but it hurts on isiXhosa with WER going from 80.4% to 88.7%. This shows that LoRA is not always beneficial. It also highlights again that we might need particular strategies when dealing with underrepresented languages like isiXhosa – even though we are in the same domain, the absolute performance and the benefits of particular strategies differ dramatically between Afrikaans and isiXhosa.

Based on these results, we proceed with Whisper medium but without LoRA in the experiments below.

B. Does out-of-domain adult speech help?

Related work: Annotated child speech is very limited in most languages. To compensate for this, one approach is to use transcribed adult speech from the target language to combine with child speech during training. While some found this to be fruitful [22], others have not seen any benefit [33], [34]. We explore the effectiveness of this strategy using adult speech from the same language but a different domain than oral narratives.

Setup: We sample out-of-domain adult speech from the NCHLT speech corpus [31]. There are roughly 56 hours of speech from around 200 speakers for Afrikaans and isiXhosa each. The audio is mostly read text from government documents. We incorporate the adult data using a similar transfer learning approach to [22]: we first fine-tune Whisper using out-of-domain adult data, and then further fine-tune using the 5-minute, in-domain child data.

Results: WERs are presented in the upper section of Table III.[3] First, we look at the model fine-tuned only on the adult data (NCHLT-adult), achieving 94.5% and 94.3% WERs on Afrikaans and isiXhosa, respectively. These are poor scores and provide little or no benefit over the baseline models in Table II (88.1% and 105.6%). This aligns with the consensus that adding unmodified adult speech from a different domain does not help [33]–[35]. Where we do see a benefit is when the model is further fine-tuned on the in-domain child data (NCHLT-adult $\rightarrow$ 5m-child), with the WER improving by 4.8% absolute for Afrikaans and 6.5% for isiXhosa over the exclusive 5m-child model. The larger boost for isiXhosa could be ascribed to Whisper being exposed to isiXhosa for the first time. Since isiXhosa is underrepresented in Whisper, we also trained a multilingual Whisper model including languages related to isiXhosa, followed by 5-minute in-domain fine-tuning.[4] But this yielded only a 1.3% absolute improvement in WER (NCHLT-adult-multi. $\rightarrow$ 5m-child, Table III).

Our findings therefore indicate that using adult speech from a different domain can be beneficial when working with very limited amounts of target domain data (in the order of a few minutes). We found, however, that this gain did not realise when more in-domain training data is available: the topline results in Table II did not improve by including adult speech.

C. Does in-domain adult speech help?

Related work: In the previous experiment, we (and others) showed that incorporating adult speech from a different domain can be effective despite the linguistic mismatch to the target domain. However, these improvements are not as substantial as those achieved with even a small amount of in-domain child data, e.g., comparing 5m-child to NCHLT-adult $\rightarrow$ 5m-child in Table III. This prompted us to ask whether a model can benefit from recordings of an adult speaker that matches the target domain of child oral narratives. This would be much easier to obtain than capturing child speech data (which brings major practical and ethical challenges).

TABLE III: WERs (%) on development data when in- and out-of-domain transcribed adult data is used to supplement child speech.

Model	Afrikaans	isiXhosa
Out-of-domain adult speech:		
5m-child	47.4*	80.4*
NCHLT-adult	94.5	94.3
NCHLT-adult $\rightarrow$ 5m-child	**42.6**	73.9
NCHLT-adult-multi. $\rightarrow$ 5m-child	—	**72.6**
In-domain adult speech:		
5m-child	47.4*	80.4*
5m-child + 5m-adult	43.1	78.4
5m-child + 30m-adult	**33.7**	**70.2**

Setup: We asked an adult speaker from each target language to record themselves reading transcriptions from the in-domain training set which were not included in our 5-minute training set. For each language, we collected 30 minutes of adult speech. We also sampled a smaller set of 5 minutes of adult audio, for comparison. We pool the child (in-domain) and adult (in-domain) data for fine-tuning, rather than training sequentially as in Sec. IV-B.

Results: WERs are shown in the bottom section of Table III. For Afrikaans and isiXhosa, adding 5 minutes of in-domain adult data (5m-child + 5m-adult) improves WER by 4.3% and 2.0%, respectively. For Afrikaans, using only 5 minutes of in-domain adult data nearly matches the performance of using 52 hours of out-of-domain adult data (NCHLT-adult $\rightarrow$ 5m-child). When including up to 30 minutes of in-domain adult data, we achieve the best results that we have seen so far for both languages (33.7% and 70.2%).

Adult in-domain data is beneficial. Our analysis is idealised in that we assume we have text corresponding to child oral narratives – but these will still be easier to obtain than child recordings.

D. Does voice conversion help?

Related work: Secs. IV-B and IV-C showed that both in- and out-of-domain adult speech helps. However, the variability and pronunciation of child speech, especially at the age we consider, are vastly different from those of adults [12], [13]. To address this acoustic mismatch, voice conversion (VC) has proven effective by generating child-like speech from adult speakers [18], [23], [36], [37]. But improvements are not always consistent across domains and tasks [18].

Setup: We need target child speech to serve as reference for a VC system. To mimic the realistic scenario where we do not have appropriate child speech in the target language upfront, we opt to use clean child speech from a different language – a form of cross-lingual VC [38]. Specifically, we use 10 hours of read speech from 40 British English children. We apply VC to both our in- and out-of-domain adult speech using the recent kNN-VC system, a light-weight method giving state-of-the-art results [39].

Results: WERs are shown in Table IV. Training a model on the converted out-of-domain adult speech and then further fine-tuning on in-domain child speech (NCHLT-adult-vc $\rightarrow$ 5m-child) shows no

[3]Results repeating from a previous table are marked with an asterisk (∗).

[4]We use four Nguni languages (isiNdebele, Siswati, isiZulu and also isiXhosa) from the NCHLT corpus, amounting to 200 hours of training data.

TABLE IV: WERs (%) on development data when VC is applied to in- and out-of-domain adult speech to get more child-like speech.

Model	Afrikaans	isiXhosa
Out-of-domain:		
NCHLT-adult → 5m-child	42.6*	**73.9***
NCHLT-adult-vc → 5m-child	**42.3**	74.7
In-domain:		
5m-child + 5m-adult	43.1*	78.4*
5m-child + 5m-adult + 5m-adult-vc	41.1	75.7
5m-child + 30m-adult	33.7*	70.2*
5m-child + 30m-adult + 30m-adult-vc	**31.6**	**68.1**
30m-adult + 30m-adult-vc	53.9	76.1

TABLE V: WERs (%) on development data when semi-supervised learning is used to get additional training data from unlabelled child speech.

Model	Afrikaans	isiXhosa
5m-child	47.4*	80.4*
5m-child → 5m-child + 30m-child-ss	**39.6**	76.5
5m-child → 5m-child + 1h-child-ss	39.9	**74.5**
5m-child → 5m-child + 2h-child-ss	41.3	75.8

gain compared to using the unmodified adult speech (NCHLT-adult → 5m-child). However, when in-domain adult speech is converted, we see consistent improvements of between 2 and 3% in WER, with the second-to-last row in Table IV giving the best results achieved so far. Our findings differ from [23], [36], where converted adult speech from a different domain helped, but it matches [18], which showed that improvements from VC can be setting-specific. We show here specifically that it does not help to convert data with a large linguistic mismatch from the target domain (out-of-domain adult speech), but when the linguistic mismatch disappears (in-domain adult speech), then VC can be used to compensate for the acoustic mismatch.

The improvements we achieved here and in Sec. IV-C comes mostly from better linguistic rather than acoustic modelling. To show this, we also did a data augmentation experiment where we took the 5-minute child data and converted it to voices of other children. This increases acoustic diversity, but it gave no performance improvement. (This is different from [18], but [40] showed that the benefits of VC-based data augmentation is very dependent on the amount of training data.)

E. Does semi-supervised learning help?

Related work: An ASR model trained on limited amounts of data can be used to generate pseudo labels for unlabelled audio [41]–[43]. This strategy, a form of semi-supervised learning, can help improve a model by increasing the amount of training data without having perfect transcriptions. This proved helpful for child-speech ASR in [44]. It is also very relevant since getting more unlabelled audio is often much easier than getting transcriptions.

Setup: We apply the 5-minute fine-tuned ASR models (first row, Table II) to the remaining in-domain training data (first row, Table I), which we treat as unlabelled audio from the target domain. The resulting pseudo labels are combined with the transcribed 5-minute set to train a new model. Using all the predicted labels without filtering can hurt performance [24]. We therefore implement a data quality filtering strategy [44]: we only keep the pseudo labels with the highest log-likelihood scores according to the Whisper model. We consider the top 30-minute, 1-hour and 2-hour predictions.

TABLE VI: WERs (%) on development and test data when combining the best strategies.

	Afrikaans		isiXhosa	
Model	dev	test	dev	test
Base: 5m-child	47.4*	42.8	80.4*	78.7
Combined: NCHLT-adult → 5m-child + 30m-adult + 30m-adult-vc → [same as above] + 1h-child-ss	29.8	29.0	62.1	57.3
Topline	18.3*	17.5	50.9*	51.9

Results: WERs are shown in Table V. Both languages benefit from being exposed to more in-domain data even without perfect transcriptions. The 5.9% WER improvement for isiXhosa is noteworthy considering that the model used to generate the transcriptions has a WER of 80.4%. For both languages, the biggest improvement comes from using the the top 1 hour of predicted transcriptions, after which performance starts to deteriorate.

F. Combined systems

We now combine all the best strategies. We fine-tune Whisper on out-of-domain adult data (Sec. IV-B) and then further fine-tune on the pooled 5-minute in-domain child, unmodified in-domain adult (Sec. IV-C), and in-domain adult converted (Sec. IV-D) data. We then use this model to predict transcripts for the remaining unlabelled in-domain child speech (Sec. IV-E), add these predictions to the training pool, and fine-tune for a third time. The results for this model are presented in Table VI. It achieves the best WERs on the development data: 29.8% on Afrikaans and 62.1% on isiXhosa. We also, for the first time, show WERs on our held-out test set (third row, Table I). These correlate well with the scores on the development sets: on the test data, we respectively achieve a 13.8% and 21.4% absolute improvement for Afrikaans and isiXhosa over the base model trained only on the 5-minute sets. This combined best system (utilising only 5 minutes of labelled data) also comes within 12% of the topline system (utilising more than 3 hours of labelled in-domain data). For reference, the CER of our best 5-minute combined system is 17.3% and 17.4% on the Afrikaans development and test sets, respectively, and 26.1% and 22.6% on the isiXhosa sets.

V. Conclusion

We looked at strategies previously proposed for child ASR, but in a wider range of combinations for the unique setting of recognising oral narratives from children aged 4 to 5. Using only 5 minutes of in-domain data we found the following: Using out-of-domain adult data is more beneficial on isiXhosa than on Afrikaans, since isiXhosa is under-represented in the Whisper base model. Parameter-efficient fine-tuning is also not consistent on the two languages. Voice conversion helps, but only if it is applied to in-domain data. Semi-supervised learning helps on both languages. A combined system shows that the best strategies are complementary when applied together.

The ASR models developed here represent the first step towards the larger goal of automating oral narrative assessments. In the full assessment system, predicted transcripts will be fed into a subsequent model to score a child's narrative ability. Although our WERs are still high (including our topline results), previous studies have shown that insights can still be extracted from noisy transcripts despite high WERs [45], [46]. Future work will investigate this in a full oral narrative assessment system.

REFERENCES

[1] S. van Staden, K. Roux, and M. Tshele, "PIRLS 2021: South African children's literacy achievement," Centre for Evaluation and Assessment, 2023.

[2] Y. D. Chiu, "The simple view of reading across development: Prediction of Grade 3 reading comprehension from prekindergarten skills," *Remedial and Special Education*, 2018.

[3] H. N. Hjetland, E. I. Brinchmann, R. Scherer, C. Hulme, and M. Melby-Lervåg, "Preschool pathways to reading comprehension: A systematic meta-analytic review," *Educational Research Review*, 2020.

[4] S. Babayiğit, S. Roulstone, and Y. Wren, "Linguistic comprehension and narrative skills predict reading ability: A 9-year longitudinal study," *British Journal of Educational Psychology*, 2021.

[5] A. Schick and G. Melzi, "The development of children's oral narratives across contexts," *Early Education*, 2010.

[6] E. Reese, D. Leyva, A. Sparks, and W. Grolnick, "Maternal elaborative reminiscing increases low-income children's narrative skills relative to dialogic reading," *Early Education and Development*, 2010.

[7] J. V. Oakhill and K. Cain, "The precursors of reading ability in young readers: Evidence from a four-year longitudinal study," *Scientific Studies of Reading*, 2012.

[8] N. Gardner-Neblett and I. U. Iruka, "Oral narrative skills: Explaining the language-emergent literacy link by race/ethnicity and SES," *Developmental Psychology*, 2015.

[9] A. P. Kaiser and M. Y. Roberts, "Advances in early communication and language intervention," *Journal of Early Intervention*, 2011.

[10] K. L. Carson, G. T. Gillon, and T. M. Boustead, "Classroom phonological awareness instruction and literacy outcomes in the first year of school," *Language, Speech, and Hearing Services in Schools*, 2013.

[11] G. Gillon, B. McNeill, A. Scott, A. Denston, L. Wilson, K. Carson, and A. H. Macfarlane, "A better start to literacy learning: Findings from a teacher-implemented intervention in children's first year at school," *Reading and Writing*, 2019.

[12] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Acoustical Society of America*, 1999.

[13] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, 2007.

[14] H. Veeramani, N. B. Shankar, A. Johnson, and A. Alwan, "Towards automatically assessing children's picture description tasks," in *Proc. SLaTE*, 2023.

[15] A. Johnson, H. Veeramani, N. Balaji Shankar, and A. Alwan, "An equitable framework for automatically assessing children's oral narrative language abilities," in *Proc. Interspeech*, 2023.

[16] G. Yeung and A. Alwan, "On the difficulties of automatic speech recognition for kindergarten-aged children," in *Proc. Interspeech*, 2018.

[17] F. Claus, H. G. Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, "A survey about databases of children's speech," in *Proc. Interspeech*, 2013.

[18] Y. Zhang, Z. Yue, T. Patel, and O. Scharenborg, "Improving child speech recognition with augmented child-like speech," in *Proc. Interspeech*, 2024.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML.*, 2023.

[20] R. Fan, N. B. Shankar, and A. Alwan, "Benchmarking children's ASR with supervised and self-supervised speech foundation models," in *Proc. Interspeech*, 2024.

[21] Z. Song, J. Zhuo, Y. Yang, Z. Ma, S. Zhang, and X. Chen, "LoRA-Whisper: Parameter-efficient and extensible multilingual ASR," in *Proc. Interspeech*, 2024.

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

[23] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, "Voice conversion based data augmentation to improve children's speech recognition in limited data scenario," in *Proc. Interspeech*, 2020.

[24] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, 2002.

[25] N. Gagarina, D. Klop, S. Kunnari, K. Tantele, T. Välimaa, U. Bohnacker, and J. Walters, "MAIN: Multilingual assessment instrument for narratives – revised version," *ZAS Papers in Linguistics*, 2019.

[26] K. Cain, S. O'Carroll, J. Oakhill, D. Klop, M. Visser, A. Smith, and A. Swart, "Exploring the impact of a story-based teacher training programme on language and early literacy in 4- and 5-year-olds," Wordworks, Cape Town, 2024.

[27] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," http://www.praat.org/, 2022, version 6.2.09.

[28] W. Liu, Y. Qin, Z. Peng, and T. Lee, "Sparsely shared LoRA on Whisper for child speech recognition," in *Proc. ICASSP*, 2024.

[29] R. Jain, A. Barcovschi, M. Y. Yiwere, P. Corcoran, and H. Cucu, "Exploring native and non-native English child speech recognition with Whisper," *IEEE Access*, 2024.

[30] A. Babu *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, 2022.

[31] E. Barnard, M. Davel, C. van Heerden, F. Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proc. SLTU*, 2014.

[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.

[33] D. Elenius and M. Blomberg, "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children," in *Proc. Interspeech*, 2005.

[34] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving children's speech recognition through out-of-domain data augmentation," in *Proc. Interspeech*, 2016.

[35] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, "Adaptation algorithms for neural network-based speech recognition: An overview," *IEEE Open Journal of Signal Processing*, 2020.

[36] D. K. Singh, P. P. Amin, H. B. Sailor, and H. A. Patil, "Data augmentation using CycleGAN for end-to-end children ASR," in *Proc. EUSIPCO*, 2021.

[37] Z. Shuyang, M. Singh, A. Woubie, and R. Karhila, "Data augmentation for children ASR and child-adult speaker classification using voice conversion methods," in *Proc. Interspeech*, 2023.

[38] M. Baas and H. Kamper, "Voice conversion for stuttered speech, instruments, unseen languages and textually described voices," *Communications in Computer and Information Science*, 2023.

[39] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," in *Proc. Interspeech*, 2023.

[40] M. Baas and H. Kamper, "Voice conversion can improve ASR in very low-resource settings," in *Proc. Interspeech*, 2022.

[41] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Proc. ICASSP*, 2013.

[42] F. Weninger, F. Mana, R. Gemello, J. Andrés-Ferrer, and P. Zhan, "Semi-supervised learning with data augmentation for end-to-end ASR," in *Proc. Interspeech*, 2020.

[43] E. Wallington, B. Kershenbaum, O. Klejch, and P. Bell, "On the learning dynamics of semi-supervised training for ASR," in *Proc. Interspeech*, 2021.

[44] J. Wang, Y. Zhu, R. Fan, W. Chu, and A. Alwan, "Low resource German ASR with untranscribed data spoken by non-native children – Interspeech 2021 shared task SPAPL system," in *Proc. Interspeech*, 2021.

[45] Y. Tao, S. G. Mitsven, L. K. Perry, D. S. Messinger, and M.-L. Shyu, "Audio-based group detection for classroom dynamics analysis," in *Proc. ICDMW*, 2019.

[46] S. L. Pugh, S. K. Subburaj, A. R. Rao, A. E. B. Stewart, J. Andrews-Todd, and S. K. D'Mello, "Say what? Automatic modeling of collaborative problem solving skills from student speech in the wild," in *Proc. EDM*, 2021.