Challenges remain in Building ASR for Spontaneous Preschool Children Speech in Naturalistic Educational Environments

Conference Paper · September 2022

DOI: 10.21437/Interspeech.2022-555

CITATIONS
0

READS
17

7 authors, including:

Satwik Dutta
University of Texas at Dallas
11 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Rebecca Elizabeth Hacker
University of Kansas
7 PUBLICATIONS **29** CITATIONS
SEE PROFILE

Dwight Irvin
University of Kansas
33 PUBLICATIONS **335** CITATIONS
SEE PROFILE

John H. L. Hansen
University of Texas at Dallas
651 PUBLICATIONS **13,123** CITATIONS
SEE PROFILE

Interspeech 2022
18-22 September 2022, Incheon, Korea

Challenges remain in Building ASR for Spontaneous Preschool Children Speech in Naturalistic Educational Environments

Satwik Dutta[1], *Sarah Anne Tao*[1], *Jacob Reyna*[1], *Rebecca E. Hacker*[2], *Dwight Irvin*[2], *Jay Buzhardt*[2], *and John H.L. Hansen*[1*]

[1]Center for Robust Speech Systems, The University of Texas at Dallas, Richardson, Texas, USA
[2]Juniper Gardens Children's Project, The University of Kansas, Kansas City, Kansas, USA

satwik.dutta@utdallas.edu, dwirvin@ku.edu, john.hansen@utdallas.edu

Abstract

Monitoring child development in terms of speech/language skills has a long-term impact on their overall growth. As student diversity continues to expand in US classrooms, there is a growing need to benchmark social-communication engagement, both from a teacher-student perspective, as well as student-student content. Given various challenges with direct observation, deploying speech technology will assist in extracting meaningful information for teachers. These will help teachers to identify and respond to students in need, immediately impacting their early learning and interest. This study takes a deep dive into exploring various hybrid ASR solutions for low-resource spontaneous preschool (3-5yrs) children (with & without developmental delays) speech, being involved in various activities, and interacting with teachers and peers in naturalistic classrooms. Various out-of-domain corpora over a wide and limited age range, both scripted and spontaneous were considered. Acoustic models based on factorized TDNNs infused with Attention, and both N-gram and RNN language models were considered. Results indicate that young children have significantly different/developing articulation skills as compared to older children. Out-of-domain transcripts of interactions between young children and adults however enhance language model performance. Overall transcription of such data, including various non-linguistic markers, poses additional challenges.

Index Terms: early childhood, speech recognition, preschool children, low resource, speech/language delays, naturalistic environments.

1. Introduction

Early childhood is the formative years of a child's developmental skills, which include but are not limited to cognitive, motor, physiological, speech, and language development. On average, children acquire about 900 words by 24 months [1], and show rapid linguistic development thereafter based on speech production, vocabulary and grammar knowledge. A preschool classroom is a viable space for supporting overall development in young children. Speech/language development in preschool classrooms is reliant on various natural communication partners, including both peers and teachers. Children's speech sounds develop from their first babbles until mid-elementary school [2]. Throughout early childhood (birth to 8 yrs), typically developing children are expected to progressively acquire and improve production of speech sounds. Table 1 shows speech sounds that are expected to be developed in each stage of early childhood. When speech production skills are developing, children may omit, substitute or have inconsistency. Language planning is also evolving, so word selection and grammar may have issues. Not all children acquire these skills at a similar pace, especially those with speech/language developmental issues. Speaking traits can vary significantly from child to child who are typically developing, as well as those who might be atrisk (eg: Autism, Down syndrome, etc.). Early speech/language acquisition delays can also affect long term social and academic outcomes [3]. Using of direct observations[4] or video coding to support teachers working with young children with and without delays is not a sustainable, nor scalable, endeavor. Deploying sensor-based speech monitoring tools in classrooms can be of immense help to teachers in creating and maintaining a rich language environment for all children. Such tools could provide feedback to allow teachers to better identify children in need of further linguistic development and support.

Table 1: *Summary of speech sound development in early childhood (birth to 8 yrs) in ARPAbet format.*

Stage	Early	Middle	Late
Age (years)	1 to 3	3 to $6\frac{1}{2}$	5 to $7\frac{1}{2}$
Speech sounds expected to be developed for each stage with examples	M "mama" B "baby" Y "you" N "no" W "we" D "daddy" P "pop" HH "hi"	T "two" NG "running" K "cup" G "go" F "fish" V "van" CH "chew" JH "jump"	SH "sheep" S "see" TH "think" TH "that" R "red" Z "zoo" L "like" ZH "measure"

It is known that developing ASR systems for children is far more challenging than for adults [5], primarily due to various developing factors (e.g., articulation/pronunciation, physiology/motor skills, vocabulary, and grammar). Most prior research on child ASR [6, 7, 8, 9, 10, 11, 12, 13] has focused on older children (6-15 yrs), with more than 60 hours data collected in clean/controlled settings, with just one speaker using prompts or read stimuli, and limited spontaneous speech. To date, limited research has focused on developing speech processing systems for spontaneous adult-child interactions in naturalistic preschool settings (3-5 years) while they are involved in various activities throughout the day. Moreover, there is lack of publicly available young child speech corpora, thus low-resource. A recent study [14] also described various challenges in developing ASR systems for single-word utterances read aloud by kindergarten (5-6 years) children achieving a Word Error Rate (WER) of 25%. Our multi-disciplinary educational research project [15, 16] focuses on quantifying "learning" based on social engagement for use in classroom settings by teachers. In this study our primary focus is on developing a robust ASR system for preschool children taking into account their de-

*Work supported by the NSF Grant #1918032 & #1918012

10.21437/Interspeech.2022-555

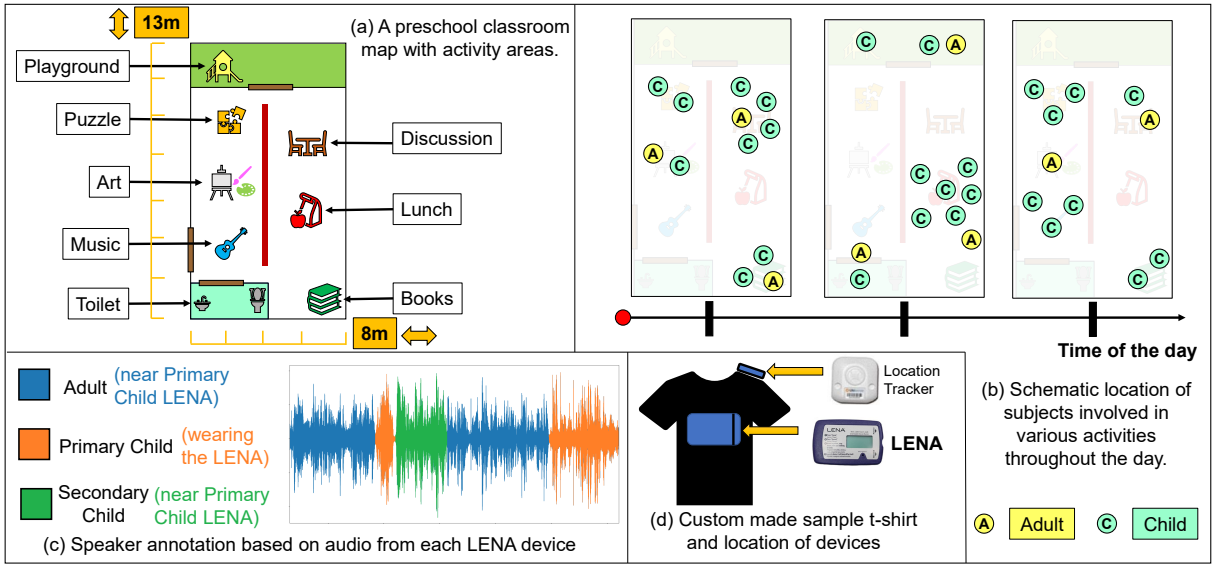

Figure 1: *Data collection of Preschool Child-Adult Interactions.*

veloping nature and developmental delays. This paper is structured as follows: Section 2 describes both in-house and out-of-domain corpora, Section 3 outlines various data augmentation approaches and both acoustic and language model development. In Section 4, ASR performance and analysis of error is shown followed by it's impact on developmental milestones. Finally, we conclude the study in Section 5.

2. Corpora

2.1. Primary Corpus: Preschool Children

Spontaneous child and adult speech was captured in preschool classrooms (Fig 1(a)), in a large urban community in a Southern state in US, using a light weight compact digital audio recorder (LENA[1]) attached to subjects (Fig. 1(d)). A total of 33 children aged 3 to 5 years with and without language or speech delays, and 8 adults teachers participated in this study. For a given session, multiple adults and children were involved in various activities throughout the day. Fig. 1(b) shows a schematic diagram of locations of the subjects through various timestamps of the day for a given session. Conversational speech was collected in multiple sessions over several days in different classrooms with different groups of children. The LENA unit data can be considered as individual audio streams and were tagged into three speaker (Fig. 1(c)) categories: Primary child (speech initiated by child wearing that LENA unit), Secondary child (speech originated by any other children within close proximity of primary child), and Adult (speech originated by any adult in close proximity). It is noted that for each LENA audio stream, there is only 1 Primary child and multiple Secondary Children and Adults (e.g., each LENA stream is associated with anonymous child ID). Out of all individual LENA audio streams, 40 streams were used for training ($\approx$ 18 hours) and remaining 8 for test ($\approx$ 4.5 hours). Care was taken to avoid overlap of speakers between train/test. Ground-truth was based on human transcriptions and only the segments spoken by both primary and secondary children (will be referred as 'Preschool') were considered for ASR assessment.

[1]https://www.lena.org/

2.2. Secondary Corpus: OGI, CMU Kids & CHILDES

OGI Kids corpus[17] ($\approx$ 60 hours) contains both prompted and spontaneous speech of 1100 children between Kindergarten and 10th grade, collected using head-mounted microphones while interacting with a computer using prompts. For the CMU Kids corpus[18] ($\approx$ 9 hours), speech is read aloud by 76 children for an age range of 6 to 11 years using head-mounted microphones. Transcripts from various corpora of the American English part of the CHILDES [19] project were used. These corpora in CHILDES, identified through a careful review with the goal of using only those conversations involving younger children (5 yrs or less) and in naturalistic scenarios, included: Braunwald, EllisWeismer, Gleason, Hall, HSLLD, MacWhinney, McMillan, Peters/Wilson, POLER-Controls, Sachs, Sawyer, Snow, and Sprott.

3. Experiment Setup

3.1. Data Augmentation

Both OGI and CMU corpora were used for speech data augmentation. Previous work using either one or both corpora [9, 11, 12, 14] for ASR only considered scripted and not spontaneous speech. For our study, two sets of OGI were considered: (i)'OGI Scripted': used only scripted speech from a random sample of children across all ages from both corpora, and (ii)'OGI Kindergarten': used both scripted and spontaneous speech of children in Kindergarten from OGI. All spontaneous speech segments in OGI were $\approx$2 mins duration each, so these were hand transcribed into shorter segments (10 to 15 secs) for ASR experiments. Since both OGI and CMU are clean, Musan[20] dataset was used to degrade the audio (in OGI & CMU).

3.2. Acoustic Model (AM) Development

All acoustic model training and decoding experiments were performed using Kaldi [21]. For the GMM-HMM systems, Mel-frequency cepstral coefficients (MFCC) were extracted for every 25 ms window and 10 ms overlap. 13 MFCCs along with their Δ and $\Delta\Delta$ features were used as front-end features.

The GMM-HMM systems were trained to provide frame-to-phone alignments for the DNN based systems. Various acoustic model adaptation techniques such as: linear discriminant analysis, maximum likelihood linear transformation estimation and speaker adaptive training were also included in training the triphone GMM-HMM systems for better alignment. The input features to the DNN-HMM models included a 40-D high resolution MFCCs of current and neighbouring frames and a 100-D i-vector of the current frame. The i-vectors were calculated by generating speed-perturbed training data. In addition, the high-resolution MFCCs were also replaced with 40-D Mel-frequency Filter Banks Energies (MFBE) by Inverse DCT. Factorized time-delay neural networks (TDNN-F)[22], originally proposed as a data-efficient alternative to TDNN for enhancing ASR performance of low-resource languages with less than 100 hours of data, were primarily used as hidden layers for our hybrid DNN-HMM acoustic models. Apart from TDNN-F layers, CNN and LSTM layers were also deployed. A time-restricted self-attention [23, 24] mechanism (with multiple heads) was also deployed. Another data augmentation approach called SpecAugment [25] was applied directly to MFBEs. It consisted of warping the features, masking blocks of frequency channels, and masking blocks of time steps. Vocal Tract Length Normalization (VTLN) [26], to compensate for varying vocal tract lengths of speakers and previously used in developing various ASR systems for children [6, 7], was also performed.

3.3. Language Model (LM) Development

In this study, both N-gram and RNN-based LMs were used. All N-gram LMs were trained using SRILM toolkit [27] and the RNN-based using PyTorch. Four 3-gram LMs were trained from scratch using the training text: (i) only Preschool, (ii) Preschool, CMU and OGI-Scripted, (iii) Preschool and OGI-Kindergarten, and (iv) Preschool and CHILDES. Pre-trained 3-gram and 4-gram LibriSpeech [28] LMs were also used. For the RNN-based LMs, we used 2-layer LSTMs of 650 embedding size, and 650 hidden dimension. Dropout was considered to overcome overfitting. Lattice rescoring[29], which has shown better performance than N-best rescoring, was used to decode the RNN-based LM. CMU Pronouncing Dictionary was used in this study. Various non-linguistic markers included: laugh, cough, scream, gasp, breath, babble, cry, loud music, crowd and play noise, and any other distinct noise.

4. Results & Discussion

4.1. Child ASR Performance

Selected ASR experiments and results are summarized in Table 2, reporting WER on Preschool test speech data. Exp# A1 shows a triphone GMM-HMM AM trained on Preschool speech generate a very high WER of 90.28% for pre-trained 3-gram LibriSpeech LM. Using an 11-layer TDNN-F based AM, 40 MFCC features and speed-perturbed i-vector (of factor 3) in Exp# A2, a much lower WER of 63.66% was achieved using the same LM than Exp# A1. However, using a pre-trained 4-gram LibriSpeech LM, a minor improvement is reported. Overall, higher N-grams did not reduce WER significantly, so the results based on only 3-gram were reported for all future experiments. Similarly, an increased speed perturbation factor of 5 also didn't improve the WER much.

In Exp# B1 (similar to A1 except LM), we notice that using an in-domain LM, WER drops to 78.39% as compared to 90.28% in Exp# A1. Again in Exp# B2 (similar to A2 except LM), we notice a significant drop of WER to 49.02% as compared to 63.66% in Exp# A2. Interpolation (without any pruning) of both the above LMs and rescoring did not improve WERs. Using LM trained on in-domain shows a significant improvement in our study than using pre-trained LibriSpeech LM, as compared to previous studies [9, 11] for older children speech where Librispeech LM worked fine. This signifies that young children do not follow the grammar/language structure in spoken English or those similar to adults, while they are still developing such skills the sentences produced by preschool children will contain various errors such as incorrect grammar, repetitions, etc. In Exp# B3 by replacing MFCCs with MFBEs and increasing the number of TDNN-F layers to 17, WER further improves to 47.02%. However in Exp# B4, using VTLN shows no improvement in WER (47.17%) for DNN-HMM systems compared to Exp# B3 (previous research using VTLN has only shown improvements for GMM-HMM systems). In Exp# B5 by adding SpecAugment layer to MFBEs, and an AM using a 6-layers of CNN and 9-layers of TDNN-F, WER further reduces to 43.03%. But in Exp# B6 by adding 1-layer of TDNN-F and LSTM, WER increases to 44.59%. In Exp# B7, by replacing the last TDNN-F+LSTM layer with multi-head Attention, WER reduces to 42.00%. Previous research [24] has achieved improvements by replacing TDNN+LSTM layers with attention for larger datasets. By lattice rescoring of an LSTM-based LM, WER (42.67%) does not improve. RNN-based LMs are data hungry, and it seems that our Preschool data does not have enough text.

Similar to Exp# B7, in Exp#s C1 by augmenting older children speech (CMU, OGI Scripted) to Preschool speech WER of 43.57% is achieved. By augmenting both scripted and spontaneous Kindergarten children speech (OGI Kindergarten), however does not improve WERs as shown in Exp#s D1. These results show that: (i) age is an important factor while developing children ASR, (ii) young children have developing articulation skills (impacting AM performance), and (iii) developing grammar/language skills (impacting LM performance). Finally by adding the CHILDES transcripts to Preschool in Exp# E1, for training an LSTM-based LM and by lattice rescoring we achieve the lowest WER of 39.52% across all test subjects. In Exp# E1A, we report WERs of 36.88% and 60.28% for test subjects with and without speech/language delays. For the same ASR engine, children with delays show higher WER.

4.2. Child ASR Error Analysis

WER, measured on the best model in Exp# E1, constituted of 25% substitution and 12% deletion w.r.t. the total words in test set. The total % of errors, due to substitution and deletion, and classified by part of speech, consisted of: 45% Nouns, 12% Verbs, 10% Pronouns, 6% Prepositions, 6% Adverbs, 4% Ad-

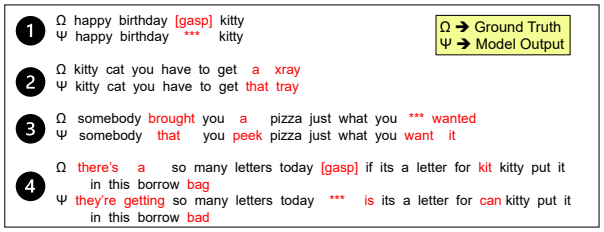

Figure 2: *Various error scenarios of model output vs. ground-truth.*

Table 2: *Child ASR Performance.*

#	Features♦	Acoustic Model Training Data♠	Acoustic Model	Language Model Training Data♠	Language Model	WER (%) of Preschool Test
A. Using Preschool (3-5 yr) child speech and pre-trained adult LM						
A1	MΔ	PS	GMM-Tri3	L	3-gram	90.28
A2	MΔ + I3	PS	TDNN-F(11)	L	3-gram vs. 4-gram	63.66 vs. 61.26
B. Using only Preschool (3-5 yr) child speech						
B1	MΔ	PS	GMM-Tri3	PS	3-gram	78.39
B2	MΔ + I3	PS	TDNN-F(11)	PS	3-gram	49.02
B3	E + I3	PS	TDNN-F(17)	PS	3-gram	47.02
B4	E + I3	PS_{VTLN}	TDNN-F(17)	PS	3-gram	47.14
B5	E_S + I3	PS	CNN(6) + TDNN-F(9)	PS	3-gram	43.03
B6	E_S + I3	PS	CNN(6) + TDNN-F(10) + LSTM(1)	PS	3-gram	44.59
B7	E_S + I3	PS	CNN(6) + TDNN-F(9) + Attn(1)	PS	3-gram vs. LSTM	42.00 vs. 42.67
C. Augmenting out-domain children speech over a wide age range (5-15 yrs)						
C1	E_S + I3	PS + CM + OS	CNN(6) + TDNN-F(9) + Attn(1)	PS + CM + OS	3-gram	43.57
D. Augmenting out-domain kindergarten (5-6 yrs) children speech						
D1	E_S + I3	PS + OK	CNN(6) + TDNN-F(9) + Attn(1)	PS + OK	3-gram	42.32
E. Using out-domain naturalistic conversations of young children (5 yrs or less) and adults for LM training						
E1	E_S + I3	PS	CNN(6) + TDNN-F(9) + Attn(1)	PS + CH	LSTM	39.52
E1A	Test subjects WITHOUT speech/language DELAYS vs. subjects WITH speech/language DELAYS					36.88 vs. 60.28
♦ MΔ → MFCC & Δ & ΔΔ, E/E_S → Filter-Bank Energy (/with SpecAugment), I3/I5 → 3/5* Speed pert. iVector ♠ PS → Preschool, L → LibriSpeech, CM → CMU, CH → CHILDES, OS → OGI Scripted, OK → OGI Kindergarten						

jectives, 2% WH-words (what, who, etc.), and 15% others. Out of all substitution errors, 80% were Monosyllabic words and remaining were Multi-syllabic. While for deletions, 90% were Monosyllabic words and remaining were Multi-syllabic. Out of all substitution errors, 38% words contained at least 1 middle stage speech sound (refer Tab.1), and 43% words with at least 1 late stage speech sound. Similarly for deletion errors, 37% words had at least 1 middle and 29% words had at least 1 late stage speech sounds. Errors arise due to various non-linguistic markers (e.g: [gasp]), shown in Fig.2(1,4), which otherwise do not impact the sentence(meaning). Shown in Fig.2(2,3,4), words pairs like 'x-ray' and 'tray', or 'bag' and 'bad' have very similar pronunciations. Similarly, Fig.2(3) also shows an error scenario where 'wanted' was predicted as 'want it'. In the audio for Fig.2(3), while the child was trying to pronounce 'pizza', they did utter 'peek' before and thereby it can considered as a transcription error.

4.3. Beyond ASR: Impact on Learning Milestones

Various developmental milestones, from 2 months to 5 years, outlined by the American Academy of Pediatrics[30, 31], can not only assist parents but provide valuable information to preschool teachers if supported by speech technology. Table 3 provides a subject-wise evaluation of the impact of Child ASR performance on tracking few prominent language learning milestones[2] which includes: (i) verbs, (ii) WH-words (who, what, where, etc.) , and (iii) sentences.

Table 3: *Impact of Child ASR on Language Learning Milestones.*

Child ID and type	% correctly identified by ASR		
	Verbs	WH-words	Sentences*
#1 + P	75	95	52
#2 + P	77	81	35
#3 + P	69	82	45
#4 + P	66	41	33
#5 + P	58	54	27
#6 + P (delayed)	51	37	28
#7 + S	64	77	38
#8 + S	69	80	49
#9 + S	69	66	40
#10 + S	69	82	43
#11 + S	61	86	36
#12 + S	61	50	39
#13 + S	66	100	41
#14 + S (delayed)	50	50	26
P = Primary, S = Secondary *Recognized without any insertion/deletion/substition			

5. Conclusions

Developing ASR systems for children is difficult, and even more challenging for younger children, especially in naturalistic classrooms scenarios. It is not possible to relate the performance of adult or older children ASR performance to young children ASR since young children have evolving speech production and language skills. Augmenting scripted older children speech, and both scripted and spontaneous speech of kindergarten children does not aid the performance of the ASR engine. However, naturalistic conversations between young children and adults help to strengthen RNN-based language model. A major challenge also is transcribing young children speech, due to speech intelligibility, thus requiring more time and subjective judgement for transcribers to comprehend preschool children speech. Often, the transcribers have to rely on their best guess. Our investigation shows that although high WERs of 39.52% occur, this confirms that to develop robust acoustic and language models for educational applications of preschool children, more naturalistic data of conversations between younger children and adults in such scenarios is needed. Future work will emphasize on collection of similar data to help strengthen the ASR engine, and also merging the real-time location information with the ASR output for fruitful feedback to teachers to help adapt their teaching methods for diverse students.

[2]https://www.asha.org/public/speech/development/chart

6. References

[1] J. Huttenlocher, W. Haight, A. Bryk, M. Seltzer, and T. Lyons, "Early vocabulary growth: relation to language input and gender." *Developmental psychology*, vol. 27, no. 2, p. 236, 1991.

[2] L. D. Shriberg, "Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 1, pp. 105–140, 1993.

[3] A. P. Kaiser and M. Y. Roberts, "Advances in early communication and language intervention," *Journal of early intervention*, vol. 33, no. 4, pp. 298–309, 2011.

[4] D. W. Irvin, S. A. Crutchfield, C. R. Greenwood, R. L. Simpson, A. Sangwan, and J. H. Hansen, "Exploring classroom behavioral imaging: Moving closer to effective and data-based early childhood inclusion planning," *Advances in Neurodevelopmental Disorders*, vol. 1, no. 2, pp. 95–104, 2017.

[5] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 10-11, pp. 847–860, 2007.

[6] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth, "Acoustic normalization of children's speech," in *Eighth European Conference on Speech Communication and Technology*. Citeseer, 2003.

[7] P. G. Shivakumar, A. Potamianos, S. Lee, and S. S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling." in *WOCCI*, 2014, pp. 15–19.

[8] R. Tong, L. Wang, and B. Ma, "Transfer learning for children's speech recognition," in *2017 International Conference on Asian Language Processing (IALP)*. IEEE, 2017, pp. 36–39.

[9] F. Wu, L. P. García-Perera, D. Povey, and S. Khudanpur, "Advances in automatic speech recognition for child speech using factored time delay neural network." in *Interspeech*, 2019, pp. 1–5.

[10] 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.

[11] G. Yeung, R. Fan, and A. Alwan, "Fundamental frequency feature normalization and data augmentation for child speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6993–6997.

[12] L. Rumberg, H. Ehlert, U. Lüdtke, and J. Ostermann, "Age-invariant training for end-to-end child speech recognition using adversarial multi-task learning," *Proc. Interspeech 2021*, pp. 3850–3854, 2021.

[13] R. Gretter, M. Matassoni, D. Falavigna, A. Misra, C. Leong, K. Knill, and L. Wang, "ETLT 2021: Shared Task on Automatic Speech Recognition for Non-Native Children's Speech," in *Proc. Interspeech 2021*, 2021, pp. 3845–3849.

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

[15] J. H. Hansen, M. Najafian, R. Lileikyte, D. Irvin, and B. Rous, "Speech and language processing for assessing child–adult interaction based on diarization and location," *International Journal of Speech Technology*, vol. 22, no. 3, pp. 697–709, 2019.

[16] P. V. Kothalkar, S. Datla, S. Dutta, J. H. Hansen, Y. Seven, D. Irvin, and J. Buzhardt, "Measuring frequency of child-directed wh-question words for alternate preschool locations using speech recognition and location tracking technologies," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 414–418.

[17] K. Shobaki, J.-P. Hosom, and R. Cole, "The ogi kids' speech corpus and recognizers," in *Proc. of ICSLP*, 2000, pp. 564–567.

[18] M. Eskenazi, J. Mostow, and D. Graff, "The cmu kids corpus ldc97s63," *Linguistic Data Consortium database*, 1997.

[19] B. MacWhinney, *The CHILDES project: Tools for analyzing talk*. Psychology Press, 2014.

[20] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[22] 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. Interspeech 2018*, 2018, pp. 3743–3747. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1417

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[24] D. Povey, H. Hadian, P. Ghahremani, K. Li, and S. Khudanpur, "A time-restricted self-attention layer for asr," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5874–5878.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[26] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1. IEEE, 1996, pp. 346–348.

[27] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[29] K. Li, D. Povey, and S. Khudanpur, "A parallelizable lattice rescoring strategy with neural language models," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6518–6522.

[30] R. J. Gerber, T. Wilks, and C. Erdie-Lalena, "Developmental milestones: motor development," *Pediatrics in review*, vol. 31, no. 7, pp. 267–277, 2010.

[31] J. M. Zubler, L. D. Wiggins, M. M. Macias, T. M. Whitaker, J. S. Shaw, J. K. Squires, J. A. Pajek, R. B. Wolf, K. S. Slaughter, A. S. Broughton *et al.*, "Evidence-informed milestones for developmental surveillance tools," *Pediatrics*, vol. 149, no. 3, 2022.