Preschool Children Speech Recognition for Early Childhood Intervention: Motivation and Challenges

Satwik Dutta and **John H.L. Hansen**
Center for Robust Speech Systems
The University of Texas at Dallas, Richardson, Texas, USA
satwik.dutta@utdallas.edu, john.hansen@utdallas.edu

Dwight Irvin
Juniper Gardens Children's Project
The University of Kansas, Kansas City, Kansas, USA
dwirvin@ku.edu

Abstract

Monitoring child development in terms of speech/language skills has a long-term impact on their overall growth. As student diversity continue to expand in US classrooms, there is a growing need to benchmark social engagement, both from a teacher-student perspective, as well as student-student content. Given various challenges with direct observation, deploying speech technology can 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 hybrid ASR solutions for low-resource spontaneous preschool (3-5yrs) children (with and without developmental delays) speech, being involved in various activities, and interacting with teachers and peers in naturalistic classrooms. For the purpose of data augmentation, various out-of-domain corpora over a wide and limited age range, both scripted and spontaneous were considered. Acoustic models based on factorized time-delay neural networks, and both N-gram and neural 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 enhances language model performance. Overall transcription of such data, including various non-linguistic markers, poses additional challenges.

1 Introduction

Early childhood (Britto et al., 2017) is the formative years of a child's developmental skills, which include but are not limited to cognitive, motor, physiology, speech, and language development. On average, children acquire about 900 words by 24 months (Huttenlocher et al., 1991), and show rapid linguistic development thereafter based on speech production, vocabulary and grammar knowledge. Preschool classrooms are viable spaces for supporting language 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 (Shriberg, 1993). 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 issues. Early speech/language acquisition delays can affect long term social and academic outcomes (Kaiser and Roberts, 2011). Using of direct observations (Irvin et al., 2017) or manual video coding to support teachers working with young children with speech 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 lan-

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"

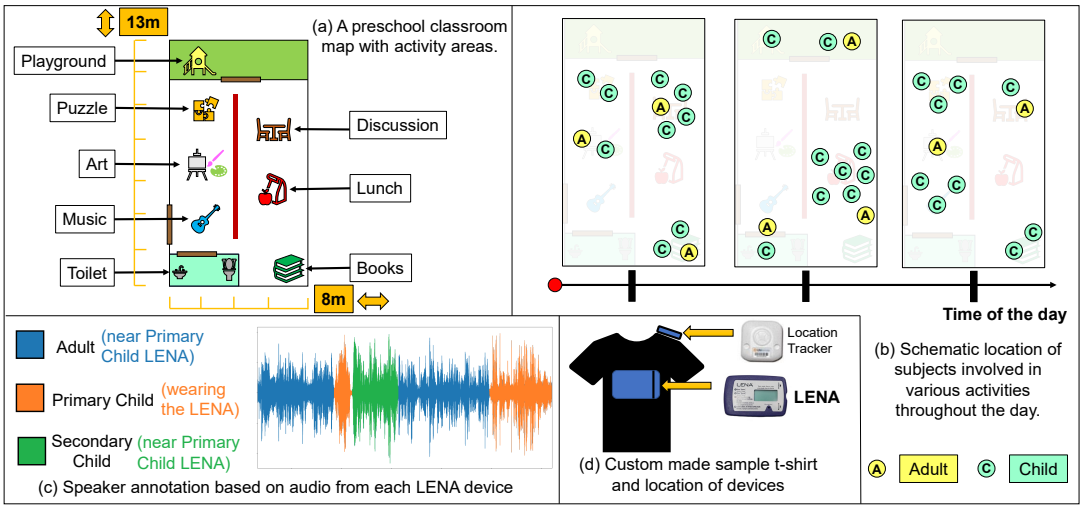

Figure 1: Data collection of Preschool Child-Adult Interactions.

guage environment for all children. Such tools could provide feedback to allow teachers to better identify children in need of further linguistic development and support. It is known that developing Automatic Speech Recognition (ASR) systems for children is far more challenging than for adults (Gerosa et al., 2007), primarily due to various developing factors (e.g., articulation/pronunciation, physiology/motor skills, vocabulary, and grammar). Most prior research on child ASR (Stemmer et al., 2003; Shivakumar et al., 2014; Tong et al., 2017; Wu et al., 2019; Shivakumar and Georgiou, 2020; Yeung et al., 2021; Rumberg et al., 2021; Gretter et al., 2021) has focused on older children (6-15 yrs), with more than 70 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 adult-child interactions in naturalistic preschool settings (3-5 yrs) while they are involved in various activities throughout the day. Moreover, there is lack of publicly available young child speech corpora (primarily due to privacy/regulations). A recent study (Yeung and Alwan, 2018) also described various challenges in developing ASR systems for single-word utterances read aloud by kindergarten (5-6 yrs) children achieving a Word Error Rate (WER) of 25%. Our multi-disciplinary educational research project focuses on quantifying "learning" based on social engagement for use in classroom settings by teachers. It is based on spontaneous interactions between multiple teachers and preschool children (3 to 5 years) within naturalistic noisy preschool classroom environments. Prior work related to this project has primarily considered 'speaker group' diarization, by classifying speech segments from adult vs. children, collectively across multiple classrooms and also for child-directed speech via an adult ASR approach. In this study our primary focus is on developing a robust ASR system for preschool children taking into account their developing nature and developmental delays.

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 speech/language 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 stream and was 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

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

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 the same group of children 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. All participants consented to the use of de-identified data for analysis. This study was approved by the Institutional Review Board of both KU and UTD for analysis.

2.2 Secondary Corpus: OGI, CMU Kids & CHILDES

OGI Kids corpus (Shobaki et al., 2000) ($\approx$ 60 hours) contains both prompted and spontaneous speech of 1100 children between Kindergarten and 10^{th} grade, collected using head-mounted microphones while interacting with a computer using prompts. For the CMU Kids corpus (Eskenazi et al., 1997) ($\approx$ 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 CHILDES (MacWhinney, 2014) project were also used. These corpora in CHILDES, which were 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 (Wu et al., 2019; Yeung et al., 2021; Rumberg et al., 2021; Yeung and Alwan, 2018) for ASR only considered scripted and not spontaneous speech. For our study, two sets of OGI were considered for augmentation: (i)'OGI Scripted': used only scripted speech from a random sample of children across all ages, and (ii)'OGI Kindergarten'($\approx$ 5 hours): used both scripted and spontaneous speech of children in Kindergarten from OGI. Spontaneous speech segment/speaker 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, compared to our Preschool data, Musan (Snyder et al., 2015) 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 (Povey et al., 2011). For the GMM-HMM systems, Mel-frequency cepstral coefficients (MFCCs) (Young, 1996) 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 (Hansen and Hasan, 2015) of the current frame. The i-vectors were calculated by generating speed-perturbed training data with 3 (0.9,1.0,1.1) speed factors. In addition, the high-resolution MFCCs were also replaced with 40-dimensional Mel-frequency Filter Banks Energies (MFBE) (Paliwal, 1999) by Inverse Discrete Cosine Transform. Factorized time-delay neural networks (TDNN-F)(Povey et al., 2018a), 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 the hybrid DNN-HMM acoustic models. Apart from TDNN-F layers, CNN and LSTM layers were also deployed. A time-restricted self-attention (Vaswani et al., 2017; Povey et al., 2018b) mechanism (with multiple heads) was also deployed. Another data augmentation approach called SpecAugment(Park et al., 2019) was applied directly to MFBEs. It consists of warping the features, masking blocks of frequency channels, and masking blocks of time steps. Vocal Tract Length Normalization (VTLN) (Eide and Gish, 1996), a speaker normalization technique to compensate for varying vocal tract lengths of

speakers and previously used in developing various child ASR systems(Stemmer et al., 2003; Shivakumar et al., 2014), was also performed.

3.3 Language Model (LM) Development

In this study, both N-gram and RNN-based language models (LM) were used. All N-gram LMs were trained using SRILM toolkit (Stolcke, 2002) and the RNN-based using PyTorch. Four types of LMs were trained from scratch using the training text: (i) only Preschool data (Sec.2.1), (ii) Preschool, CMU and OGI-Scripted, (iii) Preschool and OGI-Kindergarten, and (iv) Preschool and CHILDES. Pre-trained 3-gram and 4-gram LibriSpeech (Panayotov et al., 2015) LMs were also used. For the RNN-based LMs, we used 2-layer LSTMs of 605 embedding size and 650 hidden dimension. Dropout was considered to overcome overfitting. Lattice rescoring (Li et al., 2021) was used to decode the RNN-based LM. CMU Pronouncing Dictionary[2] was used. Various non-linguistic markers included: laugh, cough, scream, gasp, breath, babble, cry, loud music, crowd and play noise, and other noise.

4 Results & Discussion

4.1 Child ASR Performance

Selected ASR experiment 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. Replacing the 3-gram with 4-gram LibriSpeech LM, a minor improvement is reported in Exp# A3. Overall, higher N-grams did not impact WER significantly across all experiments, so only 3-grams were reported all future experiments. Similarly, an increased speed perturbation factor of 5 (Exp# A3) also didn't improve the WER much. In Exp# B1 (similar to A1 except LM), we notice that by 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 of both the above LMs and rescoring did not improve WERs. Using LM trained on spontaneous conversations of preschool children shows a significant improvement in our study than using pre-trained LibriSpeech LM, as compared to previous studies (Wu et al., 2019; Yeung et al., 2021) 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 a 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%. 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 (Povey et al., 2018b) had achieved improvements by replacing TDNN+LSTM layers with attention layers for large datasets. Again in Exp# B7, by lattice decoding of an LSTM-based LM, WER (42.44%) does not improve. LSTM-based LMs are data hungry, and it seems 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% was achieved. By augmenting both scripted and spontaneous Kindergarten children speech (OGI Kindergarten), also did 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 respectively. For the same ASR engine, children with delays show

[2]http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/

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	63.66
A3	MΔ + I3 vs. I5	PS	TDNN-F(11)	L	4-gram	62.00 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. 42.44
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. i-vector ♠ PS → Preschool, L → LibriSpeech, CM → CMU, CH → CHILDES, OS → OGI Scripted, OK → OGI Kindergarten						

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% adjectives, 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), otherwise do not change the meaning. Shown in Fig.2(2,4), words pairs like 'x-ray' and 'tray', or 'bag' and 'bad' have very similar pronunciations. Similarly, Fig.2(3) shows an error where 'wanted' was predicted as 'want it'. In the original audio for Fig.2(3), while the child was trying to pronounce 'pizza', they did utter 'peek' before and thereby it can be considered as a transcription error.

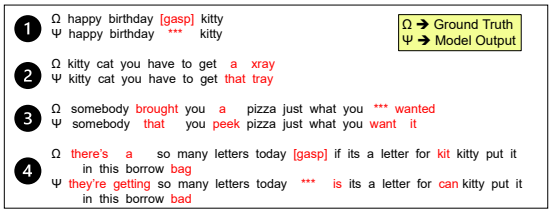

Figure 2: Various error scenarios of model output vs. ground-truth.

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 ASR performance for adults or older children with young children since young children have evolving speech production and language skills. Augmenting scripted older children speech, and both scripted and spontaneous speech of kindergarten children did not aid the ASR performance for preschool children. However, naturalistic conversations between young children and adults help to strengthen the language model slightly. A major challenge is transcribing speech of young children, due to speech intelligibility, thus requiring more time and subjective judgement for transcribers to comprehend such speech in noisy settings. Often, the transcribers have to rely on their best guess. Our investigation shows that although high WERs of $\approx$ 40% occur, to develop robust ASR models for educational applications of preschool children, more similar naturalistic data of younger children in such scenarios is needed. Our future work will emphasize on collection of similar data, focus on strengthening the ASR model, and also merging location information with ASR output for feedback to teachers.

Acknowledgements

Work supported by the National Science Foundation Grant #1918032 award to Hansen.

References

Pia R Britto, Stephen J Lye, Kerrie Proulx, Aisha K Yousafzai, Stephen G Matthews, Tyler Vaivada, Rafael Perez-Escamilla, Nirmala Rao, Patrick Ip, Lia CH Fernald, et al. 2017. Nurturing care: promoting early childhood development. *The Lancet*, 389(10064):91–102.

Ellen Eide and Herbert Gish. 1996. A parametric approach to vocal tract length normalization. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 346–348. IEEE.

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

Matteo Gerosa, Diego Giuliani, and Fabio Brugnara. 2007. Acoustic variability and automatic recognition of children's speech. *Speech Communication*, 49(10-11):847–860.

R. Gretter, Marco Matassoni, D. Falavigna, A. Misra, C.W. Leong, K. Knill, and L. Wang. 2021. ETLT 2021: Shared Task on Automatic Speech Recognition for Non-Native Children's Speech. In *Proc. Interspeech 2021*, pages 3845–3849.

John HL Hansen and Taufiq Hasan. 2015. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal processing magazine*, 32(6):74–99.

Janellen Huttenlocher, Wendy Haight, Anthony Bryk, Michael Seltzer, and Thomas Lyons. 1991. Early vocabulary growth: relation to language input and gender. *Developmental psychology*, 27(2):236.

Dwight W Irvin, Stephen A Crutchfield, Charles R Greenwood, Richard L Simpson, Abhijeet Sangwan, and John HL Hansen. 2017. Exploring classroom behavioral imaging: Moving closer to effective and data-based early childhood inclusion planning. *Advances in Neurodevelopmental Disorders*, 1(2):95–104.

Ann P Kaiser and Megan Y Roberts. 2011. Advances in early communication and language intervention. *Journal of early intervention*, 33(4):298–309.

Ke Li, Daniel Povey, and Sanjeev Khudanpur. 2021. A parallelizable lattice rescoring strategy with neural language models. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6518–6522. IEEE.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk*. Psychology Press.

Kuldip K Paliwal. 1999. On the use of filter-bank energies as features for robust speech recognition. In *ISSPA'99. Proceedings of the Fifth International Symposium on Signal Processing and its Applications (IEEE Cat. No. 99EX359)*, volume 2, pages 641–644. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. 2018a. Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proc. Interspeech 2018*, pages 3743–3747.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Daniel Povey, Hossein Hadian, Pegah Ghahremani, Ke Li, and Sanjeev Khudanpur. 2018b. A time-restricted self-attention layer for asr. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878. IEEE.

Lars Rumberg, Hanna Ehlert, Ulrike Lüdtke, and Jörn Ostermann. 2021. Age-invariant training for end-to-end child speech recognition using adversarial multitask learning. *Proc. Interspeech 2021*, pages 3850–3854.

Prashanth Gurunath Shivakumar and Panayiotis Georgiou. 2020. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer speech & language*, 63:101077.

Prashanth Gurunath Shivakumar, Alexandros Potamianos, Sungbok Lee, and Shrikanth S Narayanan. 2014. Improving speech recognition for children using acoustic adaptation and pronunciation modeling. In *WOCCI*, pages 15–19.

Khaldoun Shobaki, John-Paul Hosom, and Ronald Cole. 2000. The ogi kids' speech corpus and recognizers. In *Proc. of ICSLP*, pages 564–567.

Lawrence D Shriberg. 1993. Four new speech and prosody-voice measures for genetics research and

other studies in developmental phonological disorders. *Journal of Speech, Language, and Hearing Research*, 36(1):105–140.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Georg Stemmer, Christian Hacker, Stefan Steidl, and Elmar Nöth. 2003. Acoustic normalization of children's speech. In *Eighth European Conference on Speech Communication and Technology*. Citeseer.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Rong Tong, Lei Wang, and Bin Ma. 2017. Transfer learning for children's speech recognition. In *2017 International Conference on Asian Language Processing (IALP)*, pages 36–39. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Fei Wu, Leibny Paola García-Perera, Daniel Povey, and Sanjeev Khudanpur. 2019. Advances in automatic speech recognition for child speech using factored time delay neural network. In *Interspeech*, pages 1–5.

Gary Yeung and Abeer Alwan. 2018. On the difficulties of automatic speech recognition for kindergarten-aged children. *Interspeech 2018*.

Gary Yeung, Ruchao Fan, and Abeer Alwan. 2021. 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)*, pages 6993–6997. IEEE.

Steve Young. 1996. A review of large-vocabulary continuous-speech. *IEEE signal processing magazine*, 13(5):45.