Odyssey 2020 The Speaker and Language Recognition Workshop
1-5 November 2020, Tokyo, Japan

Assessing Child Communication Engagement via Speech Recognition in Naturalistic Active Learning Spaces

Rasa Lileikyte[1], *Dwight Irvin*[2], *John H. L. Hansen*[1]

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

rasa.lileikyte@utdallas.edu, dwirvin@ku.edu, john.hansen@utdallas.edu

Abstract

The ability to assess children's conversational interaction is critical in determining language and cognitive proficiency for typically developing and at-risk children. The earlier at-risk child is identified, the earlier support can be provided to reduce the social impact of the speech disorder. To date, limited research has been performed for young child speech recognition in classroom settings. This study addresses speech recognition research with naturalistic children's speech, where age varies from 2.5 to 5 years. Data augmentation is relatively under explored for child speech. Therefore, we investigate the effectiveness of data augmentation techniques to improve both language and acoustic models. We explore alternate text augmentation approaches using adult data, Web data, and via text generated by recurrent neural networks. We also compare several acoustic augmentation techniques: speed perturbation, tempo perturbation, and adult data. Finally, we comment on child word count rates to assess child speech development.

1. Introduction

Assessing children's conversational engagement is important in determining language and cognitive proficiency for typically developing and at-risk children (e.g. speech or language delayed). It is known that automatic speech recognition (ASR) for child conversational speech is more challenging than for adults, specifically because of the developing language planning, physiology, and motor skills of young speakers. Moreover, there is a lack of available child speech corpora. The diversity of child speech causes issues as well, since speaking traits can vary significantly from child to child who are typically developing, as well as those that might be at-risk. Children's speech structure within the age range of 2 - 6 years differs significantly from 6 - 18 year old speakers. Most prior child speech recognition efforts have focused on an older children group. Children in the age range of 2 - 6 have reduced vocal system physiologies, they are still developing their speech motor skills, pronunciation, and vocabulary. Young children do not necessarily follow adult grammar rules and proper linguistic structure. In studies [1, 2], the language interaction traits such as a child word count rate was shown to be important in the early stages of language development. The relation between word count rates and early signs of Autism has also been addressed in [3, 4, 5, 6, 7].

The motivation of this study is as follows:

- Investigate young child (age from 2.5 to 5 years) naturalistic speech recognition, when speech was recorded in active learning spaces;
- Assess the effect of data augmentation techniques for child speech;
- Explore if word count rates can provide insight to help separate at-risk and typically developing children.

While research has considered child ASR in the past, most of the studies focus on the 6 - 18 age group [8, 9]. Only a few studies have explored preschoolers speech recognition, using words, phrases, and structured human-computer interaction scenario [10, 11, 12, 13]. In our study, the scenario is based on naturalistic conversational interaction between child-adult and child-child in the daycare spaces, where children and adults are mobile with attached LENA[1] recorders. We investigate young child ASR, where age varies from 2.5 to 5 years. The extensive work from the LENA Foundation has investigated naturalistic speech of preschoolers. However, it has not considered ASR, since their language assessment strategy only estimates word count based on phoneme change sequence [14, 15, 16, 17, 18].

The corpus used in our study is comprised of 15 hours of transcribed children audio for training. Our task is very challenging, it is common to experience high word error rates (WERs) for such ASR conditions. For example, on the large 2000 hours corpus adult conversational speech recognition yields to 11% WER [19]. Meanwhile, systems with 3 hours and 40 hours training sets, achieve about 52% and 42% WER, respectively [20, 21, 22].

We explore data augmentation techniques for child naturalistic speech recognition. Data augmentation has shown to consistently improve performance of adult ASR systems. However, it has not been extensively studied for child speech. To cope with a limited amount of child training data, previous studies have explored the use of adult speech (e.g. in [9, 10, 23]). Children's speech between 2.5 to 5 years differs significantly from adult speech. As such, migrating adult based speech technologies towards this young child population is significantly more challenging. In our work, we explore (1) language model augmentation via text generated by recurrent neural networks (RNNs) [24], (2) acoustic model augmentation using speed and tempo perturbation [25]. These techniques have been used for adult speech augmentation [24, 25]. However, to the best of our knowledge, our work is the first to study these approaches for child speech. We compare these techniques and the use of adult data.

Finally, we investigate word count rates to assess child speech development of typically developing, as well as those children that might be at-risk. The word counts are estimated based on the hypothesis of our ASR system. Word count estimation could provide insight in the assessment of child language engagement in learning spaces, and identify which child might need more teacher attention.

[1] http://www.lenafoundation.org/

10.21437/Odyssey.2020-56

2. Related work

State-of-the-art speech recognition systems are usually trained on large data sets. Large quantities of in-domain data are not always available, especially for young children's speech due in part to IRB/privacy issues, as well as child speech skills diversity.

Most corpora containing children's speech focus on the 6 - 18 age group. The data sets of this age range consist of isolated words, read and prompted speech e.g. CU kids' [13], CID [26], CMU KIDS [27], TIDIGITS [28], and PF-STAR [29]. Spontaneous children-machine dialogues are collected in corpora such as NICE fairy-tale [30], CU kids' summarized stories [13], child-robot interaction AIBO [31] and PF-STAR [29], child-machine interaction [32], and Wizard-of-Oz [33]. The CHILDES corpora [34] comprise child-human conversational speech. Most of these data sets were collected in the relatively quiet settings. For the preschoolers up to 6 year age there are only few data sets. The speech of preschoolers appears in the subsets of CU kids' [13] and PF-STAR [29] corpora (4 - 6 years). The recordings contain isolated words, sentences, short spontaneous story telling, and child-robot interaction. The corpus of LENA Natural Language [35] comprises very young children (1 to 4 years) naturalistic speech. It is based on child-adult speech interaction in a naturalistic home environment. The CHILDES corpora [34] also contain naturalistic speech of young children ranging in age from 1 to 6 years.

Some automatic speech recognition studies have been performed for preschoolers within the age range of 2 - 6 years, while usually such systems have investigated older children. Isolated word and phrase recognition for 3 to 6 years children with speech disorders is analyzed in [10, 11, 13]. When 6 - 18 age group is explored, besides isolated word and sentence recognition, the studies also include continuous child speech [9, 36, 37, 38, 39, 40, 23, 41, 42, 43, 44].

Most of the previous studies explore older children's speech recorded in a relatively quiet environment, where spontaneous speech is based on children-machine dialogues. Our work focuses on very young children's speech (2.5 to 5 years), when spontaneous recordings are obtained from naturalistic conversations between child-adult and child-child during daily activities in the noisy daycare spaces.

3. Data

All experiments reported in this study use American English child spontaneous conversations captured in a high quality childcare learning center in the United States. Data was collected from 33 children of age 2.5 to 5 years, and from 4 adults/teachers (3 females and 1 male). Based on actual diagnosis eight of the children are at-risk (e.g. speech or language delayed). The speech data was gathered in three inclusive early childhood classrooms during naturally occurring morning and afternoon activities. Teachers were told to go about their typical morning activities and routines. The classrooms operated within a center-based program in a large urban community in a Southern state. As illustrated in Figure 1, the recordings contain speech in various environments such as science, art, books, music, dining space, indoor and outdoor playground. The learning spaces are open and noisy, resulting in distractions such as crowd/babble noise, and competing speech. The data was gathered wearing LENA recording units (see Figure 2), which are light-weight compact audio recorders, that cause minimal self-awareness for the speakers, allowing voice capture during naturalistic conversations. The LENA system consists of an audio recording device and speech recognition software [45]. In our work only LENA recorders are used.

The child training corpus contains about 15 hours manually transcribed audio, where transcripts have 120K word tokens. Adult data consists of 23 hours of manually transcribed audio, with 300K words in the transcripts. In addition to data gathered from the childcare learning center, an out-of-domain conversational-like Web text corpus [46] was also used, consisting of 2.6 million word tokens. All results are reported from 3 hours test data set of child speech. For development, a 1.5 hour data set was used. No speaker appeared simultaneously in the training and test sets.

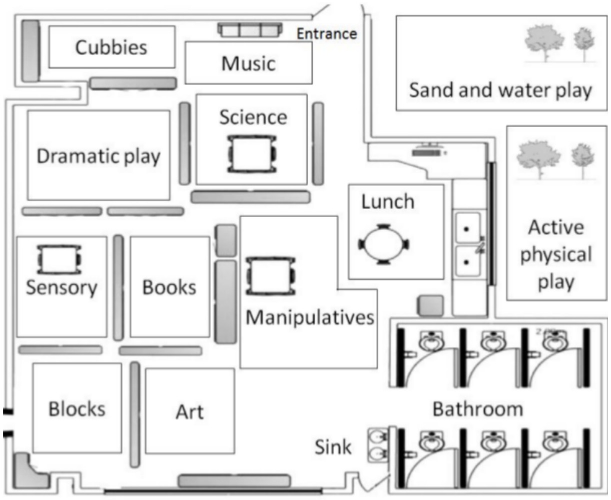

Figure 1: A typical high quality childcare learning center.

Figure 2: LENA recording device.

4. Baseline recognition system

In our experiments, an ASR system is constructed using 15 hours of transcribed conversational child speech within an age range of 2.5 - 5 years, as described in Section 3. Acoustic models are tied-state, left-to-right 3-state HMMs with Gaussian mixture observation densities. Also, triphone-based models are word position-dependent. The acoustic models are trained on 39-dimensional Mel-Frequency Cepstral Coefficients (MFCC). The features are 9 frame spliced and projected into 40 dimensions using linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT). Next, speaker adaptive training (SAT) is performed using a single feature-space maximum likelihood linear regression (fMLLR).

The 3-gram back-off language model is built using manual transcriptions from the child corpus (more details in Section 3). The lexicon is from [46], which consists of the most frequent 150K words found in the Web corpus. To build the system, we

Table 1: Results for GMM-HMM contrastive language model training conditions: manual children transcriptions (trs), manual adult transcriptions (adult), Web-texts (web), RNN generated texts based on child transcripts (rnn). In all experiments, acoustic models are based on children transcribed audio.

Language model	#Tokens	Ppx	% WER
trs (baseline)	120K	68	71.77
trs + adult	420K	60	73.24
trs + web	2.6M	56	73.33
trs + rnn	30M	66	71.80
trs + adult + web + rnn	33M	57	71.68

Table 2: Results for GMM-HMM, DNN-HMM contrastive acoustic model training conditions: manually transcribed children audio (trs), copies of children training set with different speed perturbation factors (speed perturbed), with different tempo perturbation factors (tempo perturbed), adult transcribed audio (adult). Different amount of hours for training is used (#Hrs).

Acoustic model	Perturbation factors	#Hrs	% WER	
			GMM	DNN
trs	-	15	71.68	66.26
trs + adult	-	38	76.21	66.82
trs + speed perturb	0.9, 1.1	45	71.37	63.78
trs + tempo perturb	0.9, 1.1	45	71.96	65.36
trs + speed perturb	0.8, 0.9, 1.1, 1.2	75	72.63	64.86
trs + tempo perturb	0.8, 0.9, 1.1, 1.2	75	72.07	66.85
trs + speed perturb + tempo perturb	0.8, 0.9, 1.1, 1.2	135	72.17	66.11
trs + speed perturb + tempo perturb + adult	0.8, 0.9, 1.1, 1.2	158	71.42	63.74

use the Kaldi speech recognition toolkit [47]. ASR performance is measured with WER.

This baseline model was used to decode the core open test data set, resulting in a WER of 71.77% (Table 1). A relatively high WER is expected, given the spontaneous young-child multi-speaker conversational language environment.

5. DNN system

A deep-neural network (DNN) system is trained to estimate the HMM state likelihoods [48]. The DNN uses the same features as our SAT GMM-HMM system described in Section 4: features are spliced using a context of 9 frames, followed by LDA+MLLT+fMLLR. Alignments are produced by the SAT GMM-HMM system. In the experiments with original child training data set, we use DNN topology: 2 hidden layers, 2048 neurons per layer, and the output layer is based on softmax. Sequence-discriminative training is applied with sMBR objective [49]. The learning rate is 1e-5, and the number of epochs is 5. In this study, we perform acoustic model augmentation experiments using DNN systems, with training data ranging from 15 to 158 hours. For all experiments, the same DNN topology is used, but a different number of hidden layers is employed. When the audio data set is augmented from 38 to 75 hours, 4 hidden layers are used. Furthermore, when increasing the quantity of training data, we expand the DNN to 6 hidden layers.

6. Data augmentation

The constraint given for this ASR task is that the quantity of texts and available transcribed audio data for spontaneous child speech is limited. In this section, alternate data augmentation approaches are analyzed for both language and acoustic models enhancement.

6.1. Language model augmentation

To improve the language model, three alternate data augmentation techniques are investigated: adding adult data, Web texts, and producing additional texts via RNNs [24]. The language model is estimated using supplemental text resources and interpolated with the original baseline language model. The expectation maximization (EM) algorithm is used for interpolation to minimize the perplexity of the development set.

Adult data usage. The use of manually annotated adult transcriptions is investigated for data augmentation. All conversational alike adult data was recorded in childcare center, as described in Section 3.

Web data usage. Extra conversational-like Web text data is explored to improve the language model (see in Section 3).

RNN based text generation. We also investigate additional text generation using an RNN as proposed in [24]. The RNN has 2 hidden layers and 512 units per layer. We randomly shuffled the training transcripts and split into five non-overlapping subsets. For each split, the RNN was trained using four sets and the fifth set used for validation. The RNN finds long contextual regularities, produces quite meaningful sentences, and maintains the same vocabulary.

To assess the improvement derived from the use of supplemental text resources, contrastive experiments are performed with alternate language models. From Table 1 it is observed that word perplexity is improved using all data augmentation techniques. The WER improvement is achieved only with the language model incorporating adult training transcripts, Web texts, and RNN generated texts, resulting in texts with 33M word tokens (bottom entry). In this case, the perplexity is reduced by 11 points (68 vs 57), with a corresponding tiny gain of 0.09% absolute WER over the baseline (71.77% vs 71.68%).

6.2. Acoustic model augmentation

Acoustic data augmentation is assessed via three alternate approaches: speed, tempo perturbation as described in [25], and adult data set use. We investigate the impact of different perturbation coefficients and alternate number of copies of the original child data set (15 hours).

Speed perturbation. Speed perturbation emulates both pitch and tempo variations in the speech signal. Speed modification

Table 3: Word counts of each child in test set: the number of words in references (WC ref), in hypothesis (WC hyp).

#ID child	WC ref	WC hyp	#ID child	WC ref	WC hyp
1 (at-risk)	726	605	5	3794	3296
2 (at-risk)	2780	2510	6	3825	3514
3 (at-risk)	3126	2887	7	7989	7289
4 (at-risk)	3634	3428			

is achieved by resampling the signal. We used the *speed* command of *sox*[2] tool to modify the speed of the signal. We explore augmentation of the training data set by changing the speed of the audio signal, resulting in four versions of the original child training data with speed factors of 0.8, 0.9, 1.1, and 1.2.

Tempo perturbation. The tempo of the signal is modified, while the pitch and spectral envelope of the signal is not changed. To perform tempo perturbation, we used the *sox* with *tempo* command. The training data set was enlarged by creating four additional copies of the original child training data by modifying the tempo factors to 0.8, 0.9, 1.1, and 1.2.

Adult data usage. We joined child and adult training data sets. The adult data set is comprised of 23 hours of transcribed audio, where most speakers are females. All data was recorded in childcare center (see Section 3).

Acoustic model augmentation results are provided in Table 2. In the experiments, we use a language model where child training transcriptions are interpolated with adult, Web, and RNN generated texts.

Table 2 shows that for GMM-HMM system, the highest WER improvement is obtained by incorporating two copies of child transcribed audio with speed factors 0.9, 1.1. In this case, 45 hours of training data is used, with WER improved by 0.31% absolute compared to original child training audio set (71.68% vs 71.37%).

The performance of DNN-HMM systems is also summarized in Table 2. The top line indicates that with the original children transcribed audio set, improvement of 5.42% absolute is obtained using DNN-HMM training over GMM-HMM. Comparing DNNs performance with different acoustic model sets, it can be observed that an absolute WER reduction of 2.48% is achieved using 45 hours data set which incorporates speed perturbed audio signals with 0.9, 1.1 factors (66.26% vs 63.78%). No improvement is obtained adding tempo perturbation with factors varying from 0.8 to 1.2. Other perturbation combinations are beneficial compared to the original child training audio set, but not better than using two copies of the speed perturbation. Finaly, we investigate the 158 hours data set that additionally includes transcribed adult data. In this case, the highest improvement of 8.03% is achieved over the baseline (71.77% vs 63.74%).

7. Word count estimation

The environment of the early childhood classroom settings is important for child speech learning. There is a need to identify which children are at-risk for low language/communication engagement, and these children should receive more teacher support during learning activities.

In this section, we assess children's speech development using word count rates. The word counts are estimated based on the hypothesis of our best ASR system. The results of word count estimation for children's speech are provided in Table 3. Comparing the word counts from references (Table 3, column WC ref) with counts in hypothesis (column WC hyp), it can be seen that even if there are ASR system errors, it is still possible to establish which children have low conversational interaction and are at-risk (child#1, child#2, and child#3). The system reorders child#4 and child#5 based on word counts in the hypothesis. Due to challenges in this naturalistic child-child and adult-child learning space, the word counts are not completely accurate, however they are consistent, and we are still able to establish which children have low conversational interaction. These children should get more teacher support in social and pre-academic learning during activities in the daycare center.

8. Conclusions

This research has investigated the benefits of applying data augmentation techniques for young child (age from 2.5 to 5 years) in assessing child naturalistic engagement through speech recognition. We explored several data augmentation techniques to advance language and acoustic models, and showed which provided gains in ASR performance. We also explored assessment of child language development via word count rates. The results showed that even lower performing ASR systems can contribute to effective conversation engagement assessment.

Alternate text augmentation approaches were investigated to increase the limited amount of original transcribed conversational child speech using: (i) adult data, (ii) Web data, and (iii) texts generated by RNN. Interpolating these texts collectively leads to a perplexity improvement of 11 points, but unfortunately there is little/no WER gain observed over the original baseline.

Next, acoustic augmentation techniques for child speech were explored based on: (i) speed perturbation, (ii) tempo perturbation, and (iii) adult data. The experiments were performed with training data varying from 15 to 158 hours. Both speed and tempo perturbation were shown to improve WER, with speed perturbation factors of 0.9, 1.1 to be the most beneficial. The greatest WER reduction of 8.03% absolute was achieved over the baseline after incorporating all augmented audio data sets, the improved language model, and using DNN system.

Conversational interaction using word counts was explored to assess children's speech engagement. The system helped to establish a relative rank ordering of children's conversational interaction, and therefore served to provide a separation grade between at-risk and typically developing children within such child-adult active learning spaces.

9. Acknowledgements

This study was supported in part by NSF Grant #1918032 (UT-Dallas CRSS), and the Kentucky Race to the Top Early Learning Challenge Grant #S412A130045. The authors wish to thank Dr. Beth S. Rous, Dr. Dwight Irvin, and Ying Luo at Univ. of Kentucky for establishing the protocol and collecting the child database used in this study.

[2] http://sox.sourceforge.net/

10. References

[1] Laura M Justice, Anita S McGinty, Tricia Zucker, Sonia Q Cabell, and Shayne B Piasta, "Bi-directional dynamics underlie the complexity of talk in teacher–child play-based conversations in classrooms serving at-risk pupils," *Early Childhood Research Quarterly*, vol. 28, no. 3, pp. 496–508, 2013.

[2] Betty Hart and Todd R Risley, *Meaningful differences in the everyday experience of young American children.*, Paul H Brookes Publishing, 1995.

[3] Dongxin Xu, Jill Gilkerson, Jeffrey Richards, Umit Yapanel, and Sharmi Gray, "Child vocalization composition as discriminant information for automatic autism detection," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 2518–2522.

[4] Monica Delano and Martha E Snell, "The effects of social stories on the social engagement of children with autism," *Journal of Positive Behavior Interventions*, vol. 8, no. 1, pp. 29–42, 2006.

[5] Connie Kasari, Amanda C Gulsrud, Connie Wong, Susan Kwon, and Jill Locke, "Randomized controlled caregiver mediated joint engagement intervention for toddlers with autism," *Journal of autism and developmental disorders*, vol. 40, no. 9, pp. 1045–1056, 2010.

[6] Dwight W Irvin, Stephen A Crutchfield, Charles R Greenwood, Richard L Simpson, Abhijeet Sangwan, and John HL 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.

[7] John HL Hansen, Maryam Najafian, Rasa Lileikyte, Dwight Irvin, and Beth 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.

[8] Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos, "A review of asr technologies for children's speech," in *Proceedings of the 2nd Workshop on Child, Computer and Interaction*. ACM, 2009, p. 7.

[9] Joachim Fainberg, Peter Bell, Mike Lincoln, and Steve Renals, "Improving children's speech recognition through out-of-domain data augmentation.," in *Interspeech*, 2016, pp. 1598–1602.

[10] Daniel Smith, Alex Sneddon, Lauren Ward, Andreas Duenser, Jill Freyne, David Silvera-Tawil, and Angela Morgan, "Improving child speech disorder assessment by incorporating out-of-domain adult speech.," in *Interspeech*, 2017, pp. 2690–2694.

[11] Prasanna Kothalkar, Johanna Rudolph, Christine Dollaghan, Jennifer McGlothlin, Thomas Campbell, and John HL Hansen, "Fusing text-dependent word-level i-vector models to screenat riskchild speech," in *Interspeech*, 2018, pp. 1681–1685.

[12] Prasanna V Kothalkar, Johanna Rudolph, Christine Dollaghan, Jennifer McGlothlin, Thomas F Campbell, and John HL Hansen, "Automatic screening to detectat riskchild speech samples using a clinical group verification framework," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 4909–4913.

[13] Andreas Hagen, Bryan Pellom, and Ronald Cole, "Childrens speech recognition with application to interactive books and tutors," in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, St. Thomas, USA*, 2003, pp. 186–191.

[14] Jill Gilkerson, Jeffrey A Richards, Steven F Warren, Judith K Montgomery, Charles R Greenwood, D Kimbrough Oller, John HL Hansen, and Terrance D Paul, "Mapping the early language environment using all-day recordings and automated analysis," *American journal of speech-language pathology*, vol. 26, no. 2, pp. 248–265, 2017.

[15] Jessica R Dykstra, Maura G Sabatos-DeVito, Dwight W Irvin, Brian A Boyd, Kara A Hume, and Sam L Odom, "Using the language environment analysis (lena) system in preschool classrooms with children with autism spectrum disorders," *Autism*, vol. 17, no. 5, pp. 582–594, 2013.

[16] A Sangwan, JHL Hansen, DW Irvin, S Crutchfield, and CR Greenwood, "Studying the relationship between physical and language environments of children: Who's speaking to whom and where?," in *2015 IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE)*. IEEE, 2015, pp. 49–54.

[17] Melanie Soderstrom and Kelsey Wittebolle, "When do caregivers talk? the influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments," *PloS one*, vol. 8, no. 11, pp. e80646, 2013.

[18] Carrie L Ota and Ann M Berghout Austin, "Training and mentoring: Family child care providers use of linguistic inputs in conversations with children," *Early Childhood Research Quarterly*, vol. 28, no. 4, pp. 972–983, 2013.

[19] Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur, "End-to-end speech recognition using lattice-free mmi.," in *Interspeech*, 2018, pp. 12–16.

[20] Rasa Lileikytė, Lori Lamel, Jean-Luc Gauvain, and Arseniy Gorin, "Conversational telephone speech recognition for lithuanian," *Computer Speech & Language*, vol. 49, pp. 71–82, 2018.

[21] Rasa Lileikytė, Thiago Fraga-Silva, Lori Lamel, Jean-Luc Gauvain, Antoine Laurent, and Guangpu Huang, "Effective keyword search for low-resourced conversational speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5785–5789.

[22] William Hartmann, Roger Hsiao, and Stavros Tsakalidis, "Alternative networks for monolingual bottleneck features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5290–5294.

[23] Romain Serizel and Diego Giuliani, "Deep neural network adaptation for children's and adults' speech recognition," in *Italian Computational Linguistics Conference (CLiC-it)*, 2014.

[24] Tomas Mikolov and Geoffrey Zweig, "Context dependent recurrent neural network language model.," *SLT*, vol. 12, pp. 234–239, 2012.

[25] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[26] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

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

[28] R Gary, "Leonard and george doddington," in *TIDIGITS speech corpus*. Texas Instruments, Inc, 1993.

[29] Anton Batliner, Mats Blomberg, Shona D'Arcy, Daniel Elenius, Diego Giuliani, Matteo Gerosa, Christian Hacker, Martin Russell, Stefan Steidl, and Michael Wong, "The pf_star children's speech corpus," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[30] Linda Bell, Johan Boye, Joakim Gustafson, Mattias Heldner, Anders Lindström, and Mats Wirén, "The swedish nice corpus–spoken dialogues between children and embodied characters in a computer game scenario," in *Interspeech 2005-Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*. ISCA, 2005, pp. 2765–2768.

[31] Anton Batliner, Christian Hacker, Stefan Steidl, Elmar Nöth, Shona D'Arcy, Martin J Russell, and Michael Wong, """ you stupid tin box"-children interacting with the aibo robot: A cross-linguistic emotional speech corpus.," in *Lrec*, 2004.

[32] Benfang Xiao, Cynthia Girand, and Sharon Oviatt, "Multimodal integration patterns in children," in *Seventh International Conference on Spoken Language Processing*, 2002.

[33] Shrikanth Narayanan and Alexandros Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, 2002.

[34] Brian MacWhinney, *The CHILDES project: Tools for analyzing talk, Volume II: The database*, Psychology Press, 2014.

[35] Jill Gilkerson and Jeffrey A Richards, "The lena natural language study," *Boulder, CO: LENA Foundation. Retrieved March*, vol. 3, pp. 2009, 2008.

[36] Syed Shahnawazuddin, KT Deepak, Gayadhar Pradhan, and Rohit Sinha, "Enhancing noise and pitch robustness of children's asr," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5225–5229.

[37] Alexandros Potamianos and Shrikanth Narayanan, "Robust recognition of children's speech," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 603–616, 2003.

[38] Jay G Wilpon and Claus N Jacobsen, "A study of speech recognition for children and the elderly," in *1996 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1996, vol. 1, pp. 349–352.

[39] Khairun-nisa Hassanali, Su-Youn Yoon, and Lei Chen, "Automatic scoring of non-native children's spoken language proficiency.," in *SLaTE*, 2015, pp. 13–18.

[40] Rong Tong, Nancy F Chen, and Bin Ma, "Multi-task learning for mispronunciation detection on singapore childrens mandarin speech," *Interspeech*, pp. 2193–2197, 2017.

[41] Marco Matassoni, Roberto Gretter, Daniele Falavigna, and Diego Giuliani, "Non-native children speech recognition through transfer learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6229–6233.

[42] HP Chapa Sirithunge, MA Viraj J Muthugala, AG Buddhika P Jayasekara, and DP Chandima, "A wizard of oz study of human interest towards robot initiated human-robot interaction," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 515–521.

[43] Andreas Hagen, Bryan Pellom, and Ronald Cole, "Highly accurate childrens speech recognition for interactive reading tutors using subword units," *speech communication*, vol. 49, no. 12, pp. 861–873, 2007.

[44] Prashanth Gurunath Shivakumar, Alexandros Potamianos, Sungbok Lee, and Shrikanth Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling.," in *WOCCI*, 2014, pp. 15–19.

[45] D Xu, U Yapanel, and S Gray, "Reliability of the lenatm language environment analysis system in young childrens natural language home environment," *Boulder, CO: LENA Foundation. Retrieved from http://www. lenafoundation. org/TechReport. aspx Find this author on*, 2009.

[46] Anthony Rousseau, Paul Deléglise, and Yannick Estève, "Enhancing the ted-lium corpus with selected data for language modeling and more ted talks.," in *LREC*, 2014, pp. 3935–3939.

[47] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[48] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2012.

[49] Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks.," in *Interspeech*, 2013, pp. 2345–2349.