Multi-scale Context Adaptation for Improving Child Automatic Speech Recognition in Child-Adult Spoken Interactions

Manoj Kumar[1], *Daniel Bone*[1], *Kelly McWilliams*[2], *Shanna Williams*[2], *Thomas D. Lyon*[2], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA
[2]Gould School of Law, University of Southern California, Los Angeles, CA

prabakar@usc.edu, http://sail.usc.edu

Abstract

The mutual influence of participant behavior in a dyadic interaction has been studied for different modalities and quantified by computational models. In this paper, we consider the task of automatic recognition for children's speech, in the context of child-adult spoken interactions during interviews of children suspected to have been maltreated. Our long-term goal is to provide insights within this immensely important, sensitive domain through large-scale lexical and paralinguistic analysis. We demonstrate improvement in child speech recognition accuracy by conditioning on both the domain and the interlocutor's (adult) speech. Specifically, we use information from the automatic speech recognizer outputs of the adult's speech, for which we have more reliable estimates, to modify the recognition system of child's speech in an unsupervised manner. By learning first at session level, and then at the utterance level, we demonstrate an absolute improvement of upto 28% WER and 55% perplexity over the baseline results. We also report results of a parallel human speech recognition (HSR) experiment where annotators are asked to transcribe child's speech under two conditions: with and without contextual speech information. Demonstrated ASR improvements and the HSR experiment illustrate the importance of context in aiding child speech recognition, whether by humans or computers.

Index Terms: dyadic interaction, children's speech

1. Introduction

Child automatic speech recognition (ASR) is highly challenging due to a number of factors. These include higher fundamental and formant frequencies due to shorter vocal tract length [1], a limited and less conventional vocabulary, mispronunciations caused by immature co-articulation skills, and increased durational and spectral variability [2]. Speech researchers have considered various techniques to address the challenges of children's speech recognition, such as age-dependent acoustic modeling and speaker normalization by frequency warping and spectral shaping [3], modeling pronunciation distortions via phone confusion matrices [4], and adaptation using existing children's speech data [5]. With the exception of a few recent works [5, 6, 7], most approaches have focused on read speech or restricted-vocabulary speech, which limits the child's spoken vocabulary, speaking rate, and spectral variability.

We consider a critically important real-world domain for which continuous children's ASR is useful. Generally, we are focused on spoken interaction between a child and an adult as it occurs in a semi-structured clinical or educational setting. The spontaneous and often highly variable nature of these dialogs further complicates ASR, and language modeling in particular. Specifically, the data that we work with consists of dialogs between trained interviewers and children who are victims or witnesses of maltreatment or abuse.

Such a dyadic exchange involves a two-way information flow that shapes the outcome of the interaction. There is a mutual influence between the behavioral patterns, including speech, of the interlocutors. Leveraging knowledge of this influence has shown to be beneficial in learning to predict non-verbal conversational behaviors such as head nods, back-channels [8, 9, 10] and emotional states [11]. In this paper we utilize contextual lexical information representing such mutual influence to improve the recognition of child's speech.

This study is part of a research approach which provides quantitative insights into human behavior–i.e., Behavioral Signal Processing (BSP [12]). In the context of child-adult spontaneous interactions with a specific end-goal (e.g., diagnostic evaluation), the adult–a clinician, teacher, or attorney–is expected to lead the conversation and elicit responses from the child using a semi-structured question-answer pattern. Studies have indicated that behavioral cues extracted from the adult's speech are influenced by, and are reflective of, the child's communicative intent and mental state. Bone et al. found that during autism assessment sessions, the interacting psychologist's back-channel use, conversational dominance (modeled as fraction of turn length), and intonational patterns provided discriminative cues about the child's autism symptom severity [13]. Kumar et al. encoded lexical information into psycho-linguistic norms, and noted that affect, emotional valence, and gender laddenness constructs from the psychologist's speech were significantly correlated with calibrated severity scores for autism [14]. Lastly, in a task to predict perceived engagement level of the child, Gupta et al. showed that functionals of prosodic features (pitch, shimmer, and intensity) derived from the interviewer provided significant engagement classification accuracies [15].

Our initial, but important, effort towards understanding the child's spoken behavior is in performing ASR of the child's speech, wherein we aim to improve word accuracy using two primary methods: domain adaptation as well as contextual information from the adult (at local and global levels). The availability of transcribed and time-aligned audio is an essential first step toward supporting subsequent analysis of domain related measures such as productivity and comprehension [16, 17]; Specifically, follow-up studies will analyze prosody and language from ASR-lattices. Towards improving ASR, we use a combination of acoustic and language model adaptation. First, we produce separate baseline systems for adult and children speech, followed by domain adaptation to remove acoustic and linguistic mismatch. Next, ASR hypotheses from the adult's speech are used to adapt the language models for the child's

ASR system at different levels of context. Lastly, we present a perceptual experiment to inform our analyses in which human speech recognition (HR) is performed with varying levels of contextual information (the interlocutor's adjacent turns). We investigate which direction of context is more beneficial to adaptation: causal (preceding) or anti-causal (upcoming). We find both ASR and HSR experiments show that context information from the interlocutor plays a significant role in recognizing children's speech. This general methodology can be extended to other similar, task-oriented interactions.

2. Datasets

Several child speech databases were used for building models and testing the methods in this paper, including those from the target domain of child-adult interactions in child maltreatment cases, which we refer to as "Forensic Interviews".

Data from a total of 30 children were selected for the testing experiments reported in this paper (Age: mean = 8.56 yrs, std. dev = 3.14yrs, Gender: 25 Female, 5 Male). Each child participated in a single session, and the average session length was 45 minutes. A typical session consists of three phases: The *instructions* phase, where the interviewer briefs the child about the rules of the session and elicits a promise to tell the truth. The second phase is the *rapport-building* phase. The interviewer asks the child open-ended questions about non-abusive events in order to elicit narrative responses, including questions about things the child likes to do and the child's last birthday. The purpose is to put the child at ease and to increase the child's verbal productivity. The final phase is the *allegations* phase, where questions relating to the occurrence and details of the alleged incident are posed.

Manual transcriptions are prepared at the session level and checked for accuracy by trained annotators. Transcripts include a time-stamp at every two minute interval. The timestamps are used to split the audio into two minute segments. Within each segment, the text is aligned with the audio at word level using SailAlign [18], an ASR-based recursive speech-text alignment algorithm. Combining information about utterance boundaries from the transcripts with the time-alignments, we further split the two-minute segments into time-aligned utterances. All utterances before the first time-stamp and after the last time-stamp in a session were ignored for this work since automatic alignments could not be obtained. Since the transcripts include utterance-level speaker labels, it was possible to separate the corpus into adult and children speech data. The latter is used as test data for the purpose of this paper.

Other Children's Speech Corpora

ASR systems for child speech have been shown to perform better when trained with children's speech rather than adults speech [19]. Three different child speech corpora were used for training child ASR models in this work, referred to as CUKids, CHIMP and OGI (Table 1). *CUKids* includes the CU Read, Prompted and CU Story Corpus [20] consisting of read speech (stories and summaries) from children of ages 6-11 years collected by the University of Colorado, and a subset of CSLR Prompted and Read Children's speech [20] consisting of children producing isolated words, sentences and short spontaneous stories. The *CHIMP* corpus (Children's Interactive Multimedia Project) [21] includes spontaneous speech data from 160 kids while playing a voice-activated video game controlled by an animated chimpanzee character.

Table 1: *Child speech datasets used in this work*

Dataset	# Utt	Size (hours)
Forensic Interviews	3795	3.83
CUKids	40000	40.34
CHIMP	25000	10.25
OGI	1100	30.46

The *OGI* corpus [22] contains prompted and spontaneous speech from 1100 children ranging from Kindergarten through 10th grade. The corpus consists of a single utterance per speaker. Since each utterance was of relatively long duration (μ=99.8s, σ=34.4s), speech-text alignment was performed using SailAlign to segment the audio into smaller utterances.

3. Experiments

To evaluate the performance of our speech recognition system, and demonstrate the improvement at each level of adaptation we report the results as follows: (i) Using out-of-domain data to train baseline systems, (ii) Perform domain adaptation separately for adult and child speech, and (ii) Perform session adaptation for the child speech by using recognition hypotheses of the adult speech as contextual information. The last step is performed at two different levels - utterance and session. The overall methodology is outlined in Figure 1. For steps (ii) and (iii) we perform cross-validation in a leave-one-subject manner - at each fold, speech from one child is used as test data while the rest are used as adaptation data. The cross-validation is repeated so as to ensure every child is included in the test data. Hypotheses from all folds are collected together for evaluation. the details pertaining to each setup are described in the following subsections.

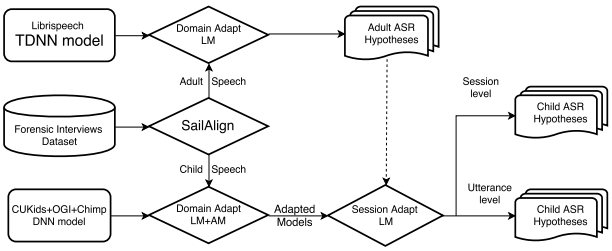

Figure 1: *An overview of the adaptation process at domain and session level*

3.1. Baseline systems

The ASR model for decoding adult speech is comprised of a time-delay neural network trained on a subset of Librispeech [23], a freely available large corpus of read English speech. Time-delay neural networks have been shown to be effective in capturing temporal contexts and inexpensive in terms of training time unlike recurrent networks [24]. A TDNN was trained on subset of Librispeech by concatenating 40 dimensional Mel Frequency Cepstral Coefficients with 100 dimensional i-vectors. 3 frames were spliced on either side to provide context at the input layer. More details of the acoustic model can be found in [25]. We used the pre-trained trigram language model [1] made available with the Librispeech corpus for obtaining the baseline results.

For the kids speech, a feed-forward DNN was trained by

[1] http://www.openslr.org/11/

pooling in data from CUKids, CHIMP and OGI datasets. First, speaker adaptive training (SAT) was used to estimate fMLLR transforms for the train data. 13-dimensional MFCC features were transformed using the estimated transforms and spliced with 5 frames on either side before feeding into the input layer. A total of 5 hidden layers were used with 1024 units per layer. The network contains approx 6.3M tunable parameters. Further, we built a trigram language model from the combined text of the corpora. We used the British English Example Pronunciation dictionary (BEEP) [26], since it has a larger vocabulary when compared to the CMU pronunciation dictionary, and was shown to perform better on children's speech [4].

The Kaldi [27] speech recognition toolkit was used for training the acoustic models and for decoding purposes, and the SRILM [28] toolkit was used for estimating the N-gram counts.

3.2. Domain Adaptation

To reduce the domain mismatch between the train and test corpora, we adapted the language models for adult speech and both acoustic and language models for the child speech. For the adult's speech, we use text from the adaptation sessions to build a small trigram language model with Witten-Bell smoothing for the unseen examples. The baseline language model is adapted to this domain model through linear interpolation of the word counts using a tunable weight. We decode the test data in two different ways: (i) Rebuilding the decoding graph, and (ii) Re-scoring lattices from the baseline step using the new (adapted) language model. Re-scoring lattices is a faster alternative to re-building the decoding graph, hence this was a reasonable choice given the leave-one-subject cross validation scheme. We experiment with both approaches in this work, but report results only for (i) since there was no significant difference in performance. The hypotheses obtained using the adapted model are kept aside for session adaptation with the child speech in the next step.

The LM adaptation process as outlined above is similarly repeated for the child speech using the child transcripts. However, since acoustic mismatch is a more profound problem for child speech in comparison to adult speech [19, 29, 30], we adapt the baseline acoustic model in addition to the language model. We retrain the neural network for an additional epoch using the word aligned audio from the adaptation sessions. Acoustic adaptation plays a significant role in WER reduction as seen from the results. Decoding on the test data is performed using the adapted acoustic and language models.

3.3. Session Adaptation

We exploit contextual information from in-session adult's speech by using the decoding graphs and hypotheses generated from the domain-adapted adult ASR. We adapt the child LM using the adult's hypotheses on top of the domain-adapted child decoding graphs. We perform a linear interpolation of the language models using tunable weights similar to the previous step. Hence the LM used for decoding child is adapted twice: using the child transcripts from adaptation sessions and adult hypotheses from test session. This setup captures the global context from the adult's speech of the entire session, hence more frequently used word counts are assigned higher weights during decoding phase.

Next, we restrict the context window for adaptation to a single utterance before and after the child utterance to be decoded. Hence, n-gram weights are adapted for each utterance in the test set, and lattice re-scoring is the only feasible option at this step. We expect this *local (utterance) adaptation* to be more selective of the useful information than *global (session) adaption*, and hence lead to greater improvements in performance.

3.4. Perceptual Experiments: Human Speech Recognition

To compare the performance of our ASR systems and demonstrate the similarity of our learning strategy with human speech recognition, we conducted certain perceptual evaluations. Three native English speakers were asked to transcribe the children's speech under two scenarios: (i) by listening once, only to the child utterance, and (ii) by listening once, to the context of conversation followed by the child utterance. We use the nearest preceding and following adult utterance for the context. The child speech data was equally divided among the three transcribers with care taken to avoid any repeatability of utterance which could potentially result in memorization by transcribers. Hence, a child utterance transcribed by annotator X for scenario (i) will not be transcribed by the same annotator for scenario (ii).

4. Results

For evaluation purposes, we use two measures - Word Error Rate (WER) and LM perplexity. WER is an end-to-end evaluation metric for ASR systems while perplexity is indicative of the predictive power of the language model alone. We present the results for baseline, domain adaptation, utterance-level (local) and session-level (global) adaptations as well as human recognition systems in Table 2.

Table 2: *Word Error Rate and perplexity measures at different levels of adaptation. HSR: Human Speech Recognition. NOTE: SRILM reports perplexity as negative exponent of the likelihood scaled by number of utterances [28]*

System	WER	Perplexity
Baseline	73.39	431
Domain adapt. (AM)	65.27	-
Domain adapt. (AM + LM)	62.47	247
Session adapt. (global)	61.04	207
Session adapt. (local)	52.69	193
HSR (isolated)	27.08	-
HSR (context)	22.49	-

The effect of domain adaptation is notable from the obtained WERs and perplexities, with acoustic adaptation alone contributing to 11% relative WER improvement and 15.1% in combination with LM adaptations. Using the adult's speech from the entire session provides only a modest improvement. We suspect this is an outcome of assigning equal weights to all words (and word counts) throughout session. Hence, words not directly related to the child's utterance tend to be assigned higher weights during the decoding process. The importance of context is highlighted when the adaptation is restricted to the neighboring utterances, resulting in an additional 15.6% relative WER improvement over the domain adapted ASR system. Although the perceptual experiment results in relatively high WERs for speech recognition by humans, addition of context accounts for 16.9% relative WER improvement, similar to ASR. Next, we discuss the effects of child age and the direction of context in learning from the adult's speech.

4.1. Performance across age groups

We observe that overall improvement in WER is consistent across ages (Figure 2). Speech from younger children is more difficult to recognize, as seen from the higher WER for younger

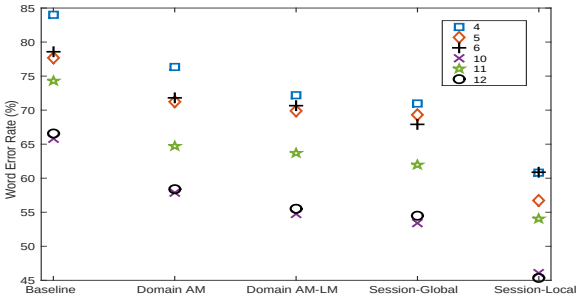

Figure 2: *WER measures across various adaptations as function of child age (in years). Session-Global and Session-Local indicate session-level and utterance-level LM adaptations, respectively.*

children. Relative WER improvement is also higher for older children (Table 3). This can possibly be explained by two reasons: Language development is yet to fully take place for younger children, hence it is more difficult to capture word counts using n-grams. This manifests as greater mismatch between linguistic structure of the adaptation and test data. Secondly, an adult's speech while interacting with younger children might be more difficult for an ASR system due to loosely defined linguistic structure and more out-of-vocabulary words (OOVs). This seems to be the case as seen from Table 3, where adult WERs are higher for younger-age children of ages 4 or 5.

Table 3: *Age-based WER analysis. For each age group, overall improvement in WER between the baseline and context-based session adaptation (local) is presented. Additionally, the WER of adult's speech from the conversation with children in each age group is presented*

Age (years)	4	5	6	10	11	12
% WER Improv. (ASR)	22.54	27.01	27.63	29.97	27.29	31.88
% WER Improv. (HSR)	25.76	-5.98	4.47	35.64	14.91	3.35
Adult WER	30.73	35.00	23.75	25.6	29.23	27.12

4.2. Causality in the use of context

Next, we investigate the following question: Is there one direction of information flow from the adult interlocutor that is more beneficial for recognizing child's speech? The *casual* component would imply that looking at the previous utterances of the adult assists recognition of the current child utterance. In general, this can happen when the child repeats the adult's question (or parts of it) or uses words semantically related to the adult's speech. On the other hand, *anti-causal* information flow would imply that succeeding adult utterances are useful for understanding the child utterance. An example is clarifications, when the adult confirms what is being spoken by the child. An illustration is shown in Figure 3. This phenomenon is expected to be more frequent during adult-child interactions when compared to adult-adult interactions due to the higher incidence of atypical and mispronunciations by the child.

We repeat the context-based session adaptation for the ASR system by conditioning on the neighboring adult utterances - preceding, succeeding and both. We also vary the number of the utterances used for learning to check the effect of context window. We use the optimized learning weights from previous experiments for linear interpolation of the language models.

From Table 4, we first observe that using both directions of information flow during learning results in intermediate performance between using only one direction of flow. Next, the

Adult: OK AND WHO DOES **<name>** SHARE A **ROOM** WITH?
Child: NOBODY **JUST COOKIE MONSTER**
Adult: **JUST COOKIE MONSTER** OK AND WHO DOES **<name>** SHARE A **ROOM** WITH?
Child: **NOBODY**
Adult: **NOBODY** SO HE HAS HIS **ROOM** ALL TO HIMSELF?
Child: YEAH
Adult: OK SO WHERE DOES **<name>** **SLEEP?**
Child: SHE **SLEEPS** AT HER **ROOM** IS DOWNSTAIRS

Figure 3: *A transcript excerpt showing directional information flow from the adult. Child phrases similar to contextual adult phrases are indicated in blue. Causal and anti-causal flows are indicated using green and red respectively*

Table 4: *WERs for Context adaptations conditioned on information flow direction*

# Utts	1	2	3	4
Causal	53.98	54.30	53.58	54.40
Anti-causal	50.78	49.41	49.15	47.47
Combined	52.69	50.80	49.66	49.55

causal flow is fairly independent of the context window, hence it does not matter how far we look behind the current child utterance. The anti-causal direction of learning however, performs better of the two and WER reduction is consistent as the context window increases. This implies that adult utterances following a child utterance provide useful information, and farther we look, more the information content. One implication is that the interviewer utilizes successive repetitions of the child's utterances, possibly to confirm details and assist the child with recollection of abusive incidents.

5. Conclusion

For adult-child spoken interactions, we have demonstrated that incorporation of context improves child ASR by conditioning on the adult interlocutor's speech. We approached a more difficult task (i.e., child ASR) by exploiting information from the prediction of an easier one (adult ASR). We analyze the benefits of scope and direction of context with respect to selection of the adult utterance hypotheses. We find that utterance level adaptation offers higher performance gains, and that succeeding adult utterances provide more information than the previous adult utterances. Results from the HSR experiment on the same data show recognition improvements consistent with the ASR systems, and that our proposed framework bears resemblance to the way context aids humans in recognizing speech. However, age-specific improvements due to context adaptation are inconsistent in HSR, as in the case of age 5 (unlike ASR).

In this work, we adapted language models with linear interpolation. This limits the scope of learning since we are constrained by the information contained in the choice of words and word counts. In the future, we would like to capture the semantic information from neighboring adult utterances and introduce words of similar meaning to improve the search space for potential child utterance words. Another method to incorporate semantic contextual information is to replace ASR hypotheses used in language adaptation with outputs from generative models for dialogue completion systems, which were recently proposed [31, 32]. However, absent of large amounts of domain-specific data, the importance of adaptation methods is increased when implementing children's speech recognition systems.

6. Acknowledgements

This work was supported by funds from the National Institutes of Health and the National Science Foundation.

7. References

[1] S. Lee, A. Potamianos, and S. 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.

[2] A. Potamianos and S. Narayanan, "Spoken dialog systems for children," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 197–200.

[3] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children." in *Eurospeech*, vol. 97, 1997, pp. 2371–2374.

[4] P. Gurunath Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling," in *Proceedings of Workshop on Child, Computer and Interaction (WOCCI 2014)*, Sep. 2014. [Online]. Available: www.wocci.org/2014/files/submissions/Shivakumar14-ISR.pdf

[5] S. S. Gray, D. Willett, J. Lu, J. Pinto, P. Maergner, and N. Bodenstab, "Child automatic speech recognition for US english: child interaction with living-room-electronic-devices," in *The 4st Workshop on Child, Computer and Interaction, WOCCI 2014, Singapore, September 19, 2014*, 2014, pp. 21–26.

[6] H. Liao, G. Pundak, O. Siohan, M. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Interspeech*, 2015.

[7] Y. Qian, X. Wang, K. Evanini, and D. Suendermann-Oeft, "Improving dnn-based automatic recognition of non-native children's speech with adult speech," in *Proceedings of Workshop on Child, Computer and Interaction (WOCCI 2016)*, 2016.

[8] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in english and japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177 – 1207, 2000.

[9] L.-P. Morency, I. de Kok, and J. Gratch, "Context-based recognition during human interactions: Automatic feature selection and encoding dictionary," in *Proceedings of the 10th international conference on Multimodal interfaces*. ACM, 2008, pp. 181–188.

[10] ——, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 70–84, 2010.

[11] C.-C. Lee, C. Busso, S. Lee, and S. S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions." in *Interspeech*, 2009.

[12] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[13] D. Bone, C.-C. Lee, A. Potamianos, and S. Narayanan, "An investigation of vocal arousal dynamics in child-psychologist interactions using synchrony measures and a conversation-based model," in *Proceedings of Interspeech*, Sep. 2014.

[14] M. Kumar, R. Gupta, D. Bone, N. Malandrakis, S. Bishop, and S. S. Narayanan, "Objective language feature analysis in children with neurodevelopmental disorders during autism assessment," in *Proceedings of Interspeech*, 2016.

[15] R. Gupta, K. Audhkhasi, S. Lee, and S. S. Narayanan, "Detecting paralinguistic events in audio stream using context in features and probabilistic decisions," *Computer, Speech, and Language*, vol. 36, pp. 72–92, Mar. 2016. [Online]. Available: www.sciencedirect.com/science/article/pii/S088523081500073X

[16] S. J. Andrews, E. C. Ahern, S. N. Stolzenberg, and T. D. Lyon, "The productivity of wh-prompts when children testify," *Applied Cognitive Psychology*, 2016.

[17] T. D. Lyon, L. Wandrey, E. Ahern, R. Licht, M. P. Sim, and J. A. Quas, "Eliciting maltreated and nonmaltreated children's transgression disclosures: Narrative practice rapport building and a putative confession," *Child development*, vol. 85, no. 4, pp. 1756–1769, 2014.

[18] A. Katsamanis, M. P. Black, P. Georgiou, L. Goldstein, and S. S. Narayanan, "Sailalign: Robust long speech-text alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Jan. 2011. [Online]. Available: sail.usc.edu/software/SailAlign/

[19] D. Elenius and M. Blomberg, "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children." in *Interspeech*, 2005, pp. 2749–2752.

[20] R. Cole, P. Hosom, and B. Pellom, "University of colorado prompted and read childrens speech corpus," Technical Report TR-CSLR-2006-03, University of Colorado, Tech. Rep., 2006.

[21] S. S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78 [IEEE Signal Processing Society Best Paper Award Wnner, 2005], Feb. 2002.

[22] K. Shobaki, J. Hosom, and R. A. Cole, "The OGI kids^2 speech corpus and recognizers," in *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16-20, 2000*, 2000, pp. 258–261.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[24] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 3214–3218.

[25] G. Chen, H. Xu, M. Wu, D. Povey, and S. Khudanpur, "Pronunciation and silence probability modeling for ASR," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 533–537.

[26] A. Robinson, "The british english example pronunciation (beep) dictionary," *Retrieved from World Wide Web: ftp://svrftp. eng. cam. ac. uk/pub/comp. speech/dictionaries/beep. tar. gz*, 1996.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[28] A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit." in *Interspeech*, vol. 2002, 2002, p. 2002.

[29] J. G. Wilpon and C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 349–352.

[30] M. Gerosa, D. Giuliani, and F. Brugnara, "Towards age-independent acoustic modeling," *Speech Communication*, vol. 51, no. 6, pp. 499–509, 2009.

[31] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," *arXiv preprint arXiv:1507.04808*, 2015.

[32] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan, "A neural network approach to context-sensitive generation of conversational responses," *arXiv preprint arXiv:1506.06714*, 2015.