Computer Speech & Language 63 (2020) 101101

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Leveraging Linguistic Context in Dyadic Interactions to Improve Automatic Speech Recognition for Children

Manoj Kumar*[,a], So Hyun Kim[b], Catherine Lord[c], Thomas D. Lyon[d], Shrikanth Narayanan[a]

[a] Signal Analysis and Interpretation Lab, University of Southern California United States
[b] Center for Autism and the Developing Brain, Weill Cornell Medicine United States
[c] Semel Institute of Neuroscience and Human Behavior, University of California Los Angeles United States
[d] USC Child Interviewing Lab and Mobile Center, University of Southern California United States

ARTICLE INFO

Article History:
Received 18 February 2019
Revised 14 February 2020
Accepted 30 March 2020
Available online 16 April 2020

Keywords:
Child speech
Automatic speech recognition
Autism spectrum disorder
Forensic interviews

ABSTRACT

Automatic speech recognition for child speech has been long considered a more challenging problem than for adult speech. Various contributing factors have been identified such as larger acoustic speech variability including mispronunciations due to continuing biological changes in growth, developing vocabulary and linguistic skills, and scarcity of training corpora. A further challenge arises when dealing with spontaneous speech of children involved in a conversational interaction, and especially when the child may have limited or impaired communication ability. This includes health applications, one of the motivating domains of this paper, that involve goal-oriented dyadic interactions between a child and clinician/adult social partner as a part of behavioral assessment. In this work, we use linguistic context information from the interaction to adapt speech recognition models for children speech. Specifically, spoken language from the interacting adult speech provides the context for the child's speech. We propose two methods to exploit this context: lexical repetitions and semantic response generation. For the latter, we make use of sequence-to-sequence models that learn to predict the target child utterance given context adult utterances. Long-term context is incorporated in the model by propagating the cell-state across the duration of conversation. We use interpolation techniques to adapt language models at the utterance level, and analyze the effect of length and direction of context (forward and backward). Two different domains are used in our experiments to demonstrate the generalized nature of our methods - interactions between a child with ASD and an adult social partner in a play-based, naturalistic setting, and in forensic interviews between a child and a trained interviewer. In both cases, context-adapted models yield significant improvement (upto 10.71% in absolute word error rate) over the baseline and perform consistently across context windows and directions. Using statistical analysis, we investigate the effect of source-based (adult) and target-based (child) factors on adaptation methods. Our results demonstrate the applicability of our modeling approach in improving child speech recognition by employing information transfer from the adult interlocutor.

*Corresponding author.
E-mail address: prabakar@usc.edu (M. Kumar).

https://doi.org/10.1016/j.csl.2020.101101
0885-2308/

1. Introduction

Automatic speech recognition (ASR) systems have become ubiquitous in our daily lives with increasing applications in smartphones and voice-activated personal assistants. Recent advances in deep learning have contributed to significant improvements in speech recognition accuracy in the last decade (Hinton et al., 2012; Graves et al., 2013; Graves and Jaitly, 2014). However, improving ASR performance for child speech continues to be more challenging problem than for adult speech due to the inherent heterogeneity in child speech signal (Potamianos and Narayanan, 1998; Lee et al., 1999; 2014). While a number of approaches have been developed over the years to tackle these issues (Burnett and Fanty, 1996; Potamianos and Narayanan, 2003; Shivakumar et al., 2014; Gray et al., 2014), most of them have been confined to read or prompted speech in their evaluation.

In this work, we study robust child ASR in a specific spoken interaction setting, namely semi-structured, goal oriented interactions between a child and an adult. These are interpersonal interactions between a child and a trained professional who navigates through a series of topics in order to gain an understanding of the child's social communication levels, intent and/or mental state and elicit the required response towards achieving the interaction goals. Specifically, we consider two real-world application domains: forensic interviews and play-based, interactive sessions of children with ASD with an adult social partner. In *forensic interviews* (Lamb et al., 2007; Hershkowitz et al., 2007), a trained interviewer questions a child about suspected criminal victimization, usually child sexual or physical abuse. The interviews are held outside a courtroom setting, and are aimed at reducing potential trauma while maximizing incident recall as well as reducing suggestive and leading question types. Observations of autism symptoms in clinical settings often involve play-based, semi-structured contexts created by an adult social partner (e.g., clinician) such as the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000) and the Brief Observation of Social Communication Change (BOSCC; Grzadzinski et al., 2016). These sessions are conducted in an interactive manner and the child's responses are evaluated based on multiple socio-communicative categories, which are further combined to quantify the level of autism symptom severity. Both domains contain spontaneous child speech which are typically characterized by short utterance durations and produced either under significant cognitive load (forensic interviews) or social demand (ADOS and BOSCC sessions, which probe for social communication and interaction characteristics).

We introduce a modeling framework for exploiting context from the spoken language of the adult interlocutor during such sessions. We explore incorporating semantic contextual information using a novel methodology. Specifically, we train a neural network based conversational system directly on the word representations. We employ a sequence-to-sequence model (Sutskever et al., 2014) which is trained using the adult's speech as encoder inputs and child's response as decoder outputs. At test time, contextual adult utterances are fed into the network to produce hypotheses for the target child utterance, which are incorporated into a dynamic language model (In the rest of this work, an utterance is defined operationally as a speaker turn). We also make use of context using lexical repetitions, and study the combination of both systems. We investigate the effect of length and direction of context on the ASR performance. Finally, we study the effect of external factors, namely age, utterance duration and word error rate (WER) of adult hypotheses on the improvement of our adaptation methods.

2. Background

Child speech is inherently different from adult speech and more complex from the viewpoint of computational processing. A number of properties have been studied, namely high acoustic variability within and across age-groups due to a growing vocal tract in children (Gerosa et al., 2006; Vorperian et al., 2005; Mugitani and Hiroya, 2012), an under-developed vocabulary leading to pronunciation and grammatical errors (Hazan and Barrett, 2000; Potamianos and Narayanan, 1998) and overall high temporal and spectral speech variability (Lee et al., 1999) across children of different age groups and gender. Research in child ASR has incorporated various techniques to address specific aspects of the above challenges. The relative difficulty in obtaining large annotated child speech corpora (more than few tens of hours) when compared to adult speech corpora led to early studies focusing on adapting acoustic models trained on adult speech to variabilities associated with child speech (Burnett and Fanty, 1996; Claes et al., 1998; Das et al., 1998; Narayanan and Potamianos, 2002). Later, (Shivakumar and Georgiou, 2018; Shivakumar et al., 2014; Elenius and Blomberg, 2005) trained acoustic models using only child speech corpora or augmented with adult speech. They observed that mixed data augmentation provided better performance than using only either children or adult speech. In order to capture pronunciation differences between child and adult speech, and variabilities within child speech, Shivakumar et al. (2014) employed phone confusion matrices to generate pronunciation alternatives during the decoding stage. Since formant frequency locations shift across the child's linguistic development, vocal tract length normalization (VTLN) (Stemmer et al., 2003; Serizel and Giuliani, 2014; 2017; Elenius and Blomberg, 2005) was used as a speaker normalization technique. VTLN compensates for different vocal tract lengths of speakers by warping the frequency axis using a speaker-specific normalization factor. Previous works also explored acoustic model adaptations such as maximum likelihood linear regression (MLLR) (Gray et al., 2014; Potamianos and Narayanan, 2003) and language model (LM) adaptations using linear interpolation (Gray et al., 2014; Kumar et al., 2017) resulting in modest improvement in WER. In a recent approach, (Shivakumar and Georgiou, 2018) explored adaptation in the feature space using a hybrid DNN-HMM acoustic model (AM). Specific layers of a DNN trained on adult speech were re-trained to account for idiosyncrasies of child speech - acoustic variability (layers close to input) and mispronunciations (layers close to output). In practice, a combination of the above methods can be employed to achieve the best performance.

2.1. Spontaneous child speech

Spontaneous speech, especially in an interaction setting, tends to include a richer vocabulary and increased variability in speaking style and background conditions than prompted speech. Previous works on analyzing spontaneous child speech have been limited to conversational agents as animated characters (Bell et al., 2005; Narayanan and Potamianos, 2002; Hagen et al., 2003), in the role of instructor (Perez-Marn and Pascual-Nieto, 2013; Johnson et al., 2000; Serholt and Barendregt, 2016), data collected from search engines (Liao et al., 2015) or in computer-based reading assessment/training applications (Black et al., 2011; Mostow et al., 1994; Cole et al., 1999). Analytic differences were observed in the child speech signal collected in spontaneous versus prompted manner. Specifically, children were found to exhibit increased disfluencies, significantly decreased vowel duration, shorter utterance durations and higher speaking rates which can possibly be attributed to higher cognitive load (Gerosa et al., 2009). Designing child ASR systems in such cases has traditionally focused on task-specific pronunciation and language models to tackle the unique vocabulary, often making use of corpus-specific transcripts.

2.2. Interaction context during dyadic conversations

Computational modeling of dyadic interactions offers possibilities in supporting and enhancing the behavior modeling and outcome prediction, complementing human judgment across many applications including health and education (Narayanan and Georgiou, 2013; Bone et al., 2017). Robust ASR is an essential step in enabling such possibilities. In both domains considered in this work, we demonstrate that performance of child speech recognition can be significantly improved by using contextual information from the adult speech. Specifically, we borrow information by solving a relatively easier problem (adult ASR) to improve the performance of the harder problem (child ASR). Utilizing context from one interlocutor to study the other's behavioral state has been explored in the past, especially with para-linguistic analyses. (Ward and Tsukahara 2000) and (Ward 1996) showed that the presence of a specific prosodic feature (low pitch value late in utterance) was an important predictor for verbal backchannels (um, oh, etc) in two separate corpora of telephone conversations involving English and Japanese speakers. Building upon (Ward and Tsukahara 2000; Morency et al. 2008) performed head gesture recognition (e.g., head nods) using latent dynamic conditional random fields trained on multimodal features from the interlocutor, namely timing (pause information, utterance duration), eye gaze, prosodic (pitch slope, continuing intonation, rapid energy changes in speech) and lexical information (unigrams). Later, (Lee et al. 2011) used a dynamic Bayesian network to model emotional states (both categorical and continuous) using acoustic-prosodic features from current and past utterances of the interlocutor. Follow-up works on the IEMOCAP corpus (Metallinou et al. 2012a,b) proposed a generic framework for modeling emotion dynamics using motion-capture (MOCAP) and prosodic features. The authors studied the evolution of dynamics both within a turn (current utterances of two speakers) and across turns (current and past utterances of both speakers). In case of the latter, they utilize recurrent neural network based architectures to model an arbitrarily long context length. Multimodal cues from the interlocutor were also shown to predict the participants' body language (Yang et al., 2014), with the strength of coordination depending on the nature of interaction (friendly or conflicting).

2.3. Language model adaptations for ASR

Language model adaptations for ASR have been studied previously in the context of human-machine interactions, especially those involving conversational agents (Narayanan and Potamianos, 2002). A majority of studies consider topics/classes - either a broad categorization of word units, or task-specific groupings (e.g., music, travel, etc). A large, out-of-domain class-independent language model is first trained and (typically linearly) interpolated with smaller class specific models. The interpolation weights can be set using hyper-parameter search, estimated using a development set (Solsona et al., 2002) or predicted using a deep neural network trained to minimize perplexity on a development set (Raju et al., 2018). To mitigate the availability of class-labeled transcriptions, (Visweswariah and Printz 2001) included unlabeled data for training class-specific models through iterative re-estimation. However, training multiple disjoint models often leads to data fragmentation. To tackle this, (Gruenstein et al. 2005) trained context-based dynamic classes for language model adaptation by borrowing the dialog state information from the conversation agent. An initial class-based language model is still required, where semantically similar units (e.g., names of different airlines) share n-gram statistics. In the context of this work, the difficulties associated with training such a model are two-fold: (1) Labeled, spontaneous child speech corpora are scarce, let alone with class-labeled transcripts, and (2) Concept of classes can be very different for child speech than adult speech (i.e., class distributions and labels might differ) and has not been explored yet.

2.3.1. Using semantic information for LM adaptation

Semantics refers to the meaning or concepts conveyed in language, rather than surface words. Semantic information in an utterance can be represented using parsers, such as a hierarchical grouping constructed using semantic tags from the word level to the sentence level (full parser) (Erdogan et al., 2005). In the above work, a full parser or shallow parser (single level of hierarchy - one semantic tag per word) is first used to obtain a language model score for every decoded utterance. This is then combined with scores computed using statistical word-level (n-gram) language model using maximum entropy modeling. Combining information from semantic and lexical sources, the authors demonstrated WER improvements for a spoken dialog system in three different domains. Semantic information adaptation can also be achieved using topic models (Hofmann, 2001)

wherein a training data set is used to estimate a latent set of topics. Topic-dependent language models can then be adapted from a base model using interpolation. Alternatively, semantic relations between words can be represented using shared occurrence/proximity in a latent space (Latent Semantic Analysis Gildea and Hofmann, 1999). However, both topic-models and LSA require substantial in-domain training corpora to reliably estimate topics.

2.4. Context for ASR adaptation

A majority of approaches in the above works have treated 'context' as speech from both speakers during the interaction, a fraction of them included external sources such as time of the day[1] (Michaely et al., 2017; Williams et al., 2018; Patel et al., 2018). This has been possible since context sources are reliable and fairly easy to obtain. For instance, obtaining n-best ASR hypotheses can be reliable for adult speech especially in controlled conditions, while state information for agent dialog (during human-agent conversations) are already available within an application. In contrast, child-adult interactions are often collected in diverse environments and estimating topics/classes from ASR hypotheses for child speech can prove unreliable. Moreover the child speech may be produced under significant communication difficulty or impairment. For the above reasons, we restricted the context source to only the adult speech.

3. Datasets for evaluation

3.1. Forensic interviews

A forensic interview (FI) is a semi-structured conversation between a certified interviewer and a child who is a suspected victim of/witness to abuse. FIs are structured to maximize child productivity and minimize suggestiveness and trauma (Lamb et al., 2007; La Rooy et al., 2015). Children's performance during an FI is likely to be adversely affected by their reticence and suggestibility, particularly when asked closed-ended questions (Lamb et al., 2007; Lyon, 2014), their typically close relationship to the perpetrator (Hershkowitz et al., 2014), and their shame and embarrassment in describing abuse (Morrison et al., 2018). Efficient and accurate transcription of FIs can increase the speed and accuracy of legal and social services interventions to protect children against abuse.

Understanding child speech during FIs is an important component towards this goal; automated extraction of characteristics of how a child speaks (e.g., acoustic prosodic patterns) and what a child speaks (words) can aid in the process of speech understanding. Automating this requires robust ASR. Child ASR can be especially challenging in this scenario considering the spontaneous nature of interaction, and the cognitive and affective load on the child due to the complex topics addressed and overall stressful nature of the interview. As part of this study, we looked at forensic interviews for 30 children (each child had one FI conducted, one interview = one session). Time-stamps at every two minutes were marked in the transcripts and used to split the session into smaller segments. Speech-to-text alignment is performed within each segment and this information is used to break the segments into utterances. All utterances were manually checked for errors before being used in the experiments of this work.

3.2. Play-based, naturalistic sessions for children with ASD with an adult social partner

Autism Spectrum Disorder (ASD) refers to a group of neuro-developmental disorders characterized by social-communication deficits along with restricted, repetitive behaviors. ASD diagnosis is obtained primarily using behavioral reports or observations made during sessions between the child and a trained clinician. In the backdrop of rising reported prevalence of autism among children in the US at 1 in 59 children diagnosed with ASD (Kogan et al., 2018), computational modeling can provide objective descriptions of the child behavior and insights into the relation between child behavior and clinically-relevant behavioral codes provided by psychologists. Building upon previous studies which have reported association between para-linguistic features extracted from the child and clinician (Bone et al., 2014) and association between the adult's lexical (Kumar et al., 2016) features with the child's symptom severity, we explore whether the interacting social partner's speech contains information useful for decoding the child speech utterance. We consider a dataset of 21 Brief Observation of Social Communication Change (BOSCC; (Grzadzinski et al., 2016)) and 1 Autism Diagnostic Observation Schedule (ADOS; (Lord et al., 2000)) session obtained from 4 different clinical sites. The ADOS is a diagnostic measure based on a play-based, semi-structured context designed to examine social communication and repetitive and restricted behaviors for a diagnosis of ASD. The ADOS takes about 40–60 min to complete. The BOSCC is a treatment outcome measure designed to capture subtle changes in social communication in children with ASD over the course of intervention. The BOSCC takes about 12 min to complete. Both ADOS and BOSCC create a naturalistic context to examine social communication for the children with ASD while they interact with an adult social partner. The sessions were manually annotated for both speaking times and transcripts. Similar to FI, the transcript segments were broken down into utterances using speech-to-text alignment and manually checked for errors. Further details are presented in Table 1.

We note that in both domains, ground truth from speech activity detection and speaker diarization are assumed available. This is not the case with most recordings in natural real-world settings; however in this paper we have chosen this setup in order to exclusively analyze ASR errors.

[1] Note that 'context' here denotes information external to the decoded utterance, and hence does not include spliced frames.

Table 1
Statistics ($\mu \pm \sigma$) for child speech in Forensic Interview (FI) sessions and Autism Spectrum Disorder (ASD) sessions. *Session duration for ASD is averaged across the ADOS (n = 1, usually lasting 40–60 min) and BOSCC sessions (n = 21, usually lasting 12 min).

	ASD	FI
Session Duration (min)	16.50 ± 6.52*	45.00 ± 17.71
Age (yrs)	9.28 ± 3.12	8.56 ± 3.14
Utterances/session	29.45 ± 16.55	37.3 ± 16.28

4. Methods

In this section, we describe the methods used to incorporate context from the adult interlocutor's speech. We borrow the contextual information using two methods - *lexical repetitions* and *semantic response predictions*. Both methods are used to generate a collection of n-grams which serve as hypotheses for the child speech. These n-grams are used to adapt a base language model into context-adapted models, which are in turn used to decode the child speech utterance. Mathematically, consider a child-adult conversation $\{(P_1, C_1), (P_2, C_2), \ldots, (P_T, C_T)\}$ with 2T utterances, where P_t and C_t represent the adult utterance and child utterance at turn-pair t respectively. The goal is to improve the ASR performance for C_t using the information from $P_t, P_{t-1}, \ldots$ and/or $P_{t+1}, P_{t+2}, \ldots$. An illustration of a child-adult session along with operational assumptions are provided in Fig. 1.

4.1. Lexical repetitions

In this method, we look for exact repetitions of context words while decoding the child speech. Repetitions form an important component in reflective listening or active listening, so as to confirm and/or clarify what was spoken. Conversational excerpts showing repetitions are presented in Fig. 2. Note that repetitions occur both ways; for example, the child confirms what was asked and the adult confirms what was said. Hence, we hypothesize that there are benefits to searching for context in both forward and backward directions. From the selected context window and direction, we extract n-grams (up to trigrams) and modify the base language model as explained in Section 4.3. The adapted language model is used for decoding the target child speech utterance.

4.2. Semantic response prediction

Lexical repetitions may not provide all the context information that is useful in decoding the child utterance. Semantics (referring to concept/meaning of a sentence) offers an important broader context and for a coherent conversation, represents a large portion of the inter-dependence between utterances. As mentioned in Section 2.3.1, semantic information can be captured using semantic parsers or LSA. In the context of this work, the shortcomings of previous approaches are two-fold: (1) It is often non-trivial to define a discrete set of concepts even within a limited time window during an open conversation. Further, language

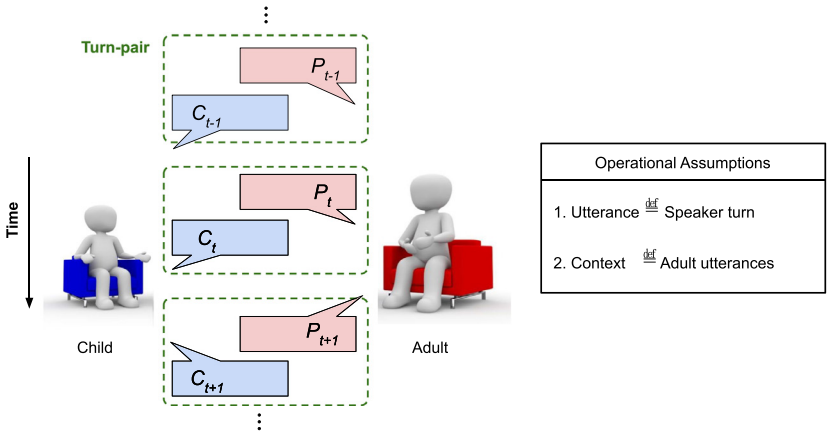

Fig. 1. Illustrating a child-adult conversation including operational assumptions in this work. A *session* S = $\{(P_1, C_1), (P_2, C_2), \ldots, (P_T, C_T)\}$ consists of alternating speaker turns. An *utterance* is defined as a speaker turn. A *turn-pair* (P_t, C_t) consists of one adult utterance and one child utterance. The *context* for improving ASR for child utterance C_t is derived from adult utterances before ($P_t, P_{t-1}, \ldots$) and/or after ($P_{t+1}, P_{t+2}, \ldots$).

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

Child: AND ITS GOING TO BE A TRAP A **BOOBY TRAP**
Psych: A **BOOBY TRAP** ALRIGHT WE ARE GOING TO PUT IT AWAY THREE TWO ONE
Child: YAAY
Psych: WHAT DID YOU SAY YOUR BROTHER'S **NAME** WAS I FORGOT
Child: HIS **NAME** IS **<name>**
Psych: **<name>** OH
Child: **HE IS AT SCHOOL**
Psych: **HE IS AT SCHOOL** RIGHT NOW HOW **OLD** IS YOUR BROTHER?
Child: **HE IS** EIGHT YEARS **OLD**
Psych: SO **HE IS** ONE YEAR OLDER THAN YOU

Fig. 2. Transcript excerpts (top: Forensic Interviews; bottom: ASD Session) illustrating information flow between the speakers. Child phrases similar to contextual adult phrases are indicated in blue. Directional flows from adult-to-child and child-to-adult are indicated using green and red respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

peculiarities in child speech such as overregularization (Marcus et al., 1992) and neologism (Volden and Lord, 1991) compound the difficulty in learning topic models. (2) Topic-models and LSA ignore the sequence information within a sentence/document (in this case, order of words in utterance) which can often prove important in conveying the meaning.

One class of neural network models which have recently shown success in learning from sequence information among multiple applications are sequence-to-sequence (seq2seq) (Sutskever et al., 2014) models. A seq2seq network maps from a variable length input sequence to another variable length output sequence. The choice of seq2seq networks for modeling semantic information is motivated by their success in varied natural language tasks such as machine translation (Sutskever et al., 2014), learning multi-lingual sentence representations (Firat et al., 2016; Yu et al., 2018) and conversational agents/dialog systems (Vinyals and Le, 2015). In the above applications, seq2seq models capture the underlying meaning of the input in the learnt representation; while translating a spoken utterance or engaging in a meaningful conversation (often with humans). This phenomenon was illustrated in Palangi et al. (2016), where the neural encoder architecture used in seq2seq models was found to activate the same cells for words with similar meanings when trained with weak supervision. In Conneau et al. (2018), various classification tasks of semantic nature such as tense, subject count and object count were used to demonstrate the semantic information captured by embeddings trained using a seq2seq model. Further, seq2seq models have been shown to parse a sentence directly from text into its abstract meaning representation (AMR) format (Konstas et al., 2017; Song et al., 2019), a directed graph which represents the semantics of a sentence invariant of its syntactic form (Banarescu et al., 2013). These works illustrate the suitability of seq2seq models for learning the semantic information between speakers in child-adult interactions.

Our application of seq2seq models is similar to that of the conversational agent, where the model takes an input speech utterance and provides an utterance in response. Specifically, we generate transcript hypotheses for the child utterance given the context adult ASR transcript. These hypotheses are used to estimate n-grams for context adaptation. We first provide a description of the seq2seq model including the LSTM unit in Section 4.2.1. In Section 4.2.2 we illustrate the application of seq2seq model for improving child ASR including training for response prediction and generating multiple hypotheses during inference. We provide an extension to seq2seq models for incorporating contextual utterances beyond the current turn-pair in Section 4.2.3.

4.2.1. Seq2seq model structure

A seq2seq model consists of two components. The *encoder* maps from a variable length input sequence to a fixed dimensional vector (*thought* vector), while the *decoder* generates a variable length output sequence using the thought vector. Encoder input and decoder output are typically sequences of vectors, where each vector can be a word/character representation. Both encoder and decoder consist of layers of recurrent neural networks (RNN), where each layer consists of a recurrently connected unit. At the encoder, word representations from the adult speech utterance are input to the unit at every timestep (a timestep corresponds to a word in this work). The child speech utterance is inferred at the outputs from the decoder unit at every timestep. Since the vanilla RNN unit is known to suffer from the vanishing gradient issue (Bengio et al., 1994), a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) or Gated Recurrent Unit (GRU) (Cho et al., 2014) is typically used instead. In this work, we use the LSTM unit similar to the original seq2seq model in Sutskever et al. (2014).

An LSTM unit consists of three "gates" which regulate the flow of information: an input gate, output gate and forget gate. A gate is a sigmoidal activation function whose output lies between 0 and 1, and regulates information flow into and out of the unit. The input to the LSTM unit can be denoted as the sequence $\{x_n\}, n \in \{1 \dots N\}$. At each timestep n, an LSTM unit receives three vectors: the input (x_n), cell-state from previous timestep (c_{n-1}) and output from previous timestep (h_{n-1}). The cell-state serves as

long-term memory for the unit and transmits information across timesteps. The input gate regulates the amount of new information from x_n that will be used to store in the cell-state. Formally,

$$i_n = \sigma(W_i\,[h_{n-1}, x_n] + b_i)$$
$$\tilde{C_n} = \tanh(W_c\,[h_{n-1}, x_n] + b_c)$$
$$c_{new} = i_n \odot \tilde{C_n}$$

where $\odot$ and $[\cdot\,,\,\cdot]$ represent element-wise product and vector concatenation respectively. c_{new} represents the component from the input to be stored in the cell-state. Next, the forget gate regulates the information to disregard from the previous timestep.

$$f_n = \sigma\left(W_f[h_{n-1}, x_n] + b_f\right)$$
$$c_f = f_n \odot c_{n-1}$$

Here, c_f represents the component that will be retained from the previous cell-state. The new cell-state c_n is the sum of above components.

$$c_n = c_{new} + c_f$$

Finally, the output gate filters the information from the cell-state to output for current timestep.

$$o_n = \sigma(W_o\,[h_{n-1}, x_n] + b_o)$$
$$h_n = o_n \odot \tanh(c_n)$$

During training, the weights (W_c, W_i, W_f, W_o) and biases (b_c, b_i, b_f, b_o) are learnt using gradient descent. For more details of the LSTM unit, refer to (Hochreiter and Schmidhuber 1997; Sutskever et al. 2014) and (Graves 2012).

4.2.2. Using seq2seq model for improving child ASR

Model training: We train the seq2seq network with the adult utterance P_t as input and the child utterance C_t as output, i.e., the network is trained to predict the child's responses from the context. We first convert the words into a one-hot representation using a vocabulary built with the most frequent words. Assuming that the vocabulary contains V words, each word is converted into a V dimensional binary vector $\overrightarrow{w}$ containing zeros in all dimensions except the one uniquely identifying the word. Thus each utterance is represented by a matrix of dimension $V \times N$, where N is the number of words in that utterance. The LSTM unit in the encoder takes word representations from the adult utterance as input. The cell-state provided to the first timestep is a vector of zeros (except during context-training, see Section 4.2.3) and the cell-state from the final timestep is the thought vector $\mathbf{v}$, which is presumed to contain information from the adult speech utterance. The thought vector is fed into the decoder as cell-state at first timestep, along with a special start-token as the input vector (see Fig. 3a). Successive decoder timesteps take as input word representations from child utterance while the cell-state is propagated throughout the utterance in the decoder. At every decoder timestep, the output vector is passed through softmax activation to produce the predicted word vector. Cross-entropy (CE) loss is computed between the predicted and true word vectors from the child utterance. The overall loss for utterance C_t is:

$$L(C_t) = \sum_{n=1}^{N_C} CE\left(\overrightarrow{w}_{C_{t,n}}, softmax(h_n)\right) \tag{1}$$

where h_n represents the output from the nth decoder timestep and $w_{C_{t,n}}$ represents the nth word from C_t.

Model inference: During inference the adult utterance is input to the encoder to obtain the thought vector. Similar to training, the start-token is fed to the first decoder timestep along with the thought vector. At the output of each decoder timestep, we randomly sample from the most likely words (known as top-K random sampling (Fan et al., 2018; Ippolito et al., 2019)) to feed as input into the next decoder timestep. This is continued until the end-token is encountered or the maximum sequence length is reached (Fig. 3b). Unlike greedy decoding (using only the most likely word at every timestep) sampling enables us to obtain multiple hypotheses for the child speech.

Training parameters: Maximum timesteps of encoder and decoder were fixed at 20 which corresponds to ≈95th percentile of the utterance word counts, while longer sentences were truncated. The cell-state dimension was fixed at 1000, following (Sutskever et al. 2014). The vocabulary size was fixed separately for each corpus (1520 for ASD, 8400 for FI), consisting of all words with a minimum frequency of 5. All remaining words were replaced with an out-of-vocabulary token. The number of layers was fixed at 1 for both encoder and decoder, considering the limited size of the data in this work. All learnable weights were initialized uniformly between −0.1 and 0.1. Adam optimizer (β_1 = 0.9, β_2 = 0.999) was used for training, with an initial learning rate of 0.0001. The batch size was fixed at 32 samples, where each sample is a turn pair (P_t, C_t). Training was carried out for 20 epochs. TensorFlow toolkit (Abadi et al., 2016) was used to perform all seq2seq model experiments.

4.2.3. Incorporating context in seq2seq model

It is often meaningful to incorporate the history of conversation while predicting the current utterance. Previous approaches to incorporate long-term context in seq2seq models include concatenating multiple utterances in the encoder (Yan et al., 2016) and using a two-step (hierarchical) model (Serban et al., 2017) where the first network encodes individual utterances, and a second network is trained to model previous utterances to the decoder output. In this work, we use a variant of the recently proposed context-aware training (Christensen et al., 2018) where the decoder cell-states are propagated through a conversation.

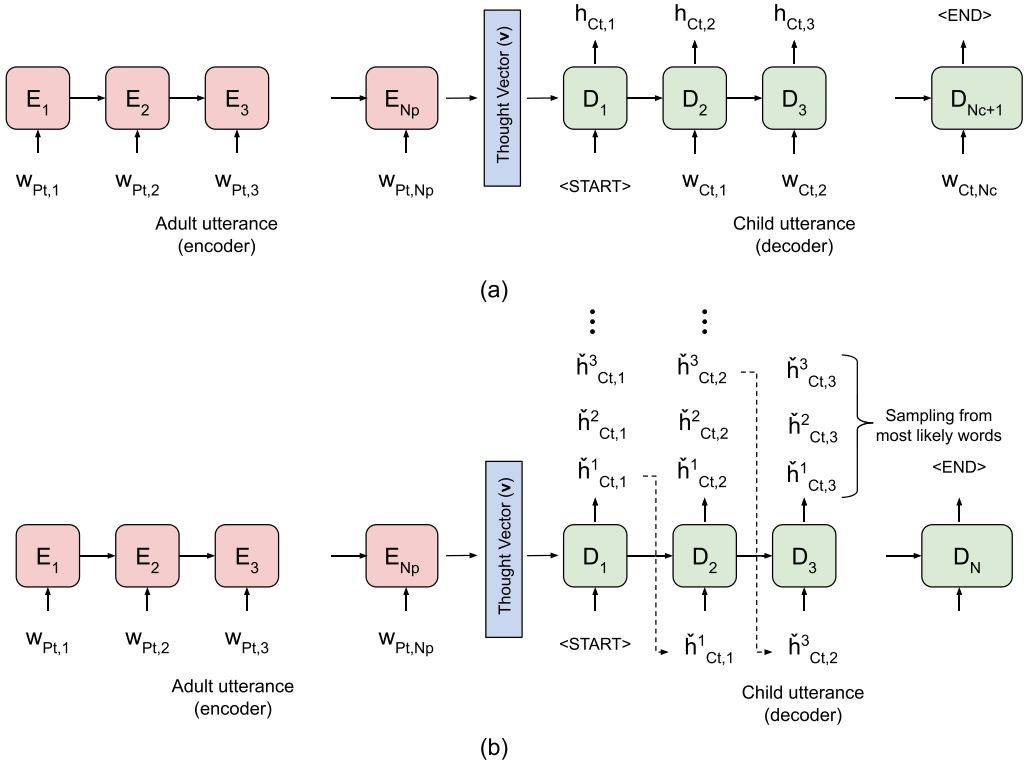

Fig. 3. (a) Training a seq2seq model with a single adult utterance at the encoder and target child utterance at the decoder. At each decoder timestep, words from child utterance are used to compute cross-entropy loss with the outputs ($h_{C_t,n}$) (b) Inference using a trained seq2seq model. At each decoder timestep, the top hypotheses for next word are sampled and fed into successive decoder timesteps until the $<$ END $>$ token is encountered.

Specifically, we use the same architecture in Fig. 3, but feed the cell-state from the final decoder timestep to the following utterance's first encoder timestep. We use the following representation of a training step:

$$(P_t, x) \rightarrow (C_t, y) \tag{2}$$

where P_t represents an encoder input sequence, x is cell-state to first encoder timestep, C_t is decoder output sequence and y is the cell-state from final decoder timestep. We illustrate below the proposed iterative context-aware training for a conversation defined earlier as $\{P_1, C_1, P_2, C_2, \ldots, P_T, C_T\}$.

- **Context-Independent Training:** During the first step of training, x is a zero vector Ø; representing no information from previous utterances.

  $(P_t, \emptyset) \rightarrow (C_t, e_{t,0}) \quad \forall\, t \in \{1, 2, ., T\}$

  Here, $e_{t,0}$ denotes the cell-state from final decoder timestep.

- **Single-Context Training:** At the next training step, decoder cell-states from context-independent training are fed as cell-state inputs to the first encoder timestep while training successive utterances:

  $(P_t, e_{t-1,0}) \rightarrow (C_t, e_{t,1}) \quad \forall\, t \in \{2, ., T\}$

  Note that $e_{t,1}$ contains information from turn-pairs t and $t-1$ through P_t and $e_{t-1,0}$ respectively

By feeding in $e_{t,1}$ as cell-state input to the first encoder timestep of the following utterance and so on, the above steps can be continued by increasing the context available. We note that although each input-output pair (P_t, C_t) is seen multiple times by the network, different amounts of context are made available through the first encoder cell-state each time. In this work, we train the seq2seq model with up to 3 context utterances ($e_{t,2}$). We note that longer context windows can possibly be used, however the propagated context information might diminish across utterances. During inference, we do not have access to neighboring child utterances to propagate the cell-state across the conversation. Hence, we generate hypotheses in a context-independent manner. We generate multiple hypotheses for the current child speech utterance C_t using top-K random sampling. Similar to lexical repetitions, we extract n-grams from these hypotheses to adapt a base language model.

4.3. Linear interpolation

Consider a fairly large n-gram based language model L_{base} estimated using out-of-domain text and a small context model $L_{context}$ estimated using adult context utterances or child utterance hypotheses from the seq2seq network. Both L_{base} and $L_{context}$ are assumed to share a common pronunciation. The goal is to estimate an adapted model that includes context information while at the same time does not overfit. We create an adapted model L_{adapt} as follows:

$$P(w|L_{adapt}) = \begin{cases} \lambda P(w|L_{base}) + (1-\lambda)P(w|L_{context}) & w \in L_{base} \cap L_{context} \\ \lambda P(w|L_{base}) & w \notin L_{context} \\ (1-\lambda)P(w|L_{context}) & w \notin L_{base} \end{cases}$$

where w represents an n-gram and λ is the interpolation weight.

5. Experiments

We present our experimental framework in three different steps:

- **Baseline Child ASR:** In Section 5.1, we describe the steps involved in building a robust baseline ASR for child speech, including corpus selection, noise and reverb augmentation and training a hybrid DNN-HMM acoustic model. We validate our model using an out-of-domain child speech corpus, namely CIDMIC (Lee et al., 1999).
- **Domain Adaptation:** We build separate models for both FI and ASD in Section 5.3 by fine-tuning the baseline models to domain specific acoustic conditions and vocabulary. All adaptation results are presented against domain-adapted models.
- **Context Adaptation:** We analyze adaptation to the adult's utterances at global (using all utterances from session) and local level (using only neighbouring utterances) in Section 5.4. In case of the latter, we experiment with two different types of context: lexical repetitions (referred to as *lexical*) and semantic response predictions (referred to as *semantic*). We further analyze the amount of context by controlling for the number of utterances (Section 6.4). Finally we examine the effect of external factors on adaptation results using statistical analysis in Section 6.5.

An overview of the entire system is presented in Fig. 4.

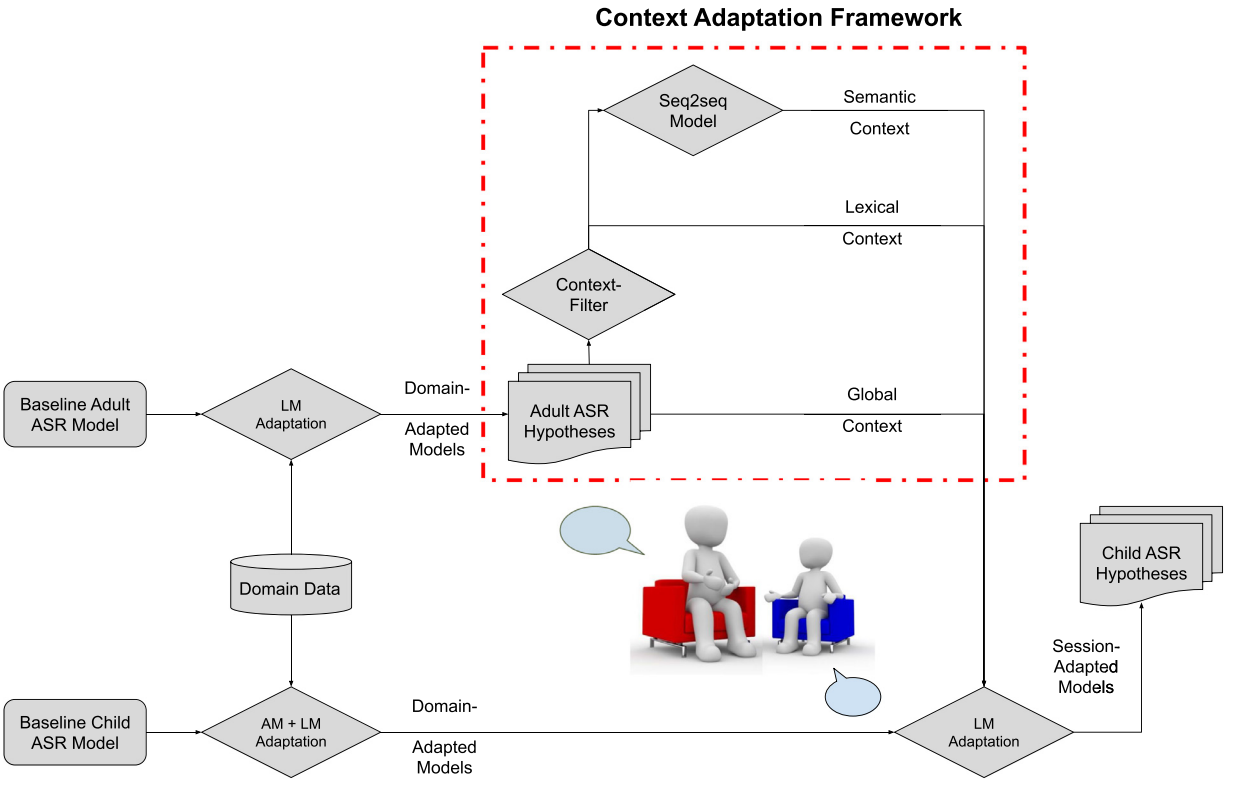

Fig. 4. Proposed methodology. The context adaptation framework is the primary contribution of this work.

Table 2
Details of child speech corpora used in training and evaluation of baseline ASR model. *CSLU statistics are computed after speech-to-text alignment.

Corpus	# Kids	Age Range (yrs)	Speaking Style	# Utts	Size (hrs)
CHIMP (Narayanan and Potamianos, 2002)	130	4–11	Spontaneous	4457	3.2
CMU Kids (Eskenazi et al., 1997)	73	6–11	Prompted	2120	3.1
CSLU* (Shobaki et al., 2000)	1093	5–15	Read, Prompted	14,238	14,8
CU Kids (Cole et al., 2006)	1084	6–11	Read	32,776	45.9
TIDIGITS (Leonard, 1984)	101	6–15	Prompted	3371	2.8
CIDMIC (Lee et al., 1999)(eval)	419	7–17	Prompted	4800	3.27

5.1. Baseline child ASR

As outlined in Section 2, child ASR models perform better when trained using children speech. We trained a baseline model based on deep neural networks that is further adapted separately at domain and session level. We use multiple child speech corpora (Table 2) spanning a wide range of age groups, background conditions, speaking styles (read, prompted, spontaneous), topics (video games, educational, generic) and utterance durations (isolated digits to stories).

Except TIDIGITS, all the corpora used here were specifically designed for developing child ASR systems. In the case of CSLU, audio recordings were of relatively long duration (μ = 99.8 s, σ = 34.4 s) and hence speech-to-text alignment was performed to segment them into smaller utterances. Only successfully aligned words were used for training purposes. We use the CIDMIC corpus as evaluation data to select our baseline model.

We resample audio from all corpora to 16 kHz sampling frequency. All utterances shorter than 2 s or longer than 20 s were removed to aid training. Utterances that included noise or silence labels in the transcripts were also removed. Further, corpus-specific disfluencies and para-linguistic labels were mapped onto a common out-of-vocabulary (OOV) unit. Details of each corpus after pre-processing steps are presented in Table 2. Since the combined size of our training corpora is only $\approx$ 70 h, we experimented with two augmentation techniques in this work.

First, we perturb the fundamental frequency (F0) of the audio without modifying the phonetic component. This is motivated by the fact that children have smaller vocal folds than adults which is manifested by higher F0. This difference is minimized above the age of 11–12 years (Lee et al., 1999). Hence, introducing modest pitch variations can potentially mimic a larger sample of children. For every utterance in the training set, we randomly choose a factor between [0.9,1.1] to scale the F0 for the entire utterance. We repeat this process 9 times to generate 700 h of clean speech. Next, we randomly introduce noise and reverberation conditions following Ko et al. (2017). A variety of real and simulated room impulse responses (RIRs) are added to the audio to simulate various background conditions, a complete description of this augmentation method can be found in Ko et al. (2017). We generate approximately 2000 h of reverberation augmented data this way. Since pre-DNN training steps typically saturate in performance with the size of training data, we include the noise and reverb-augmented data only during DNN training.

5.1.1. Pre-DNN training

We extract 13-dimensional Mel Frequency Cepstral Coefficients (frame width 25 ms, frame shift 10 ms) as our front end representations. Using the Kaldi (Povey et al., 2011) speech recognition toolkit, we train the acoustic model beginning with monophone GMM system through hybrid DNN-HMM in an iterative manner. At each step, alignments obtained using the previous model are used to initialize the current model. We used adaptation techniques such as linear discriminant analysis (LDA) and speaker adaptive training (SAT) using feature-space maximum likelihood linear regression (fMLLR) transforms. We use the SAT system to obtain alignments for the entire data for use in DNN training.

5.1.2. DNN Training

We train a hybrid HMM-DNN system where a time-delay neural network (TDNN) is used to estimate posterior probabilities for context-dependent HMM states. TDNNs have been shown to model long-term temporal dependencies as effectively as recurrent architectures with significantly less training times (Peddinti et al., 2015). We use 6 hidden layers in our network with 512 units at each layer. At the input to the network, bidirectional context of 2 frames is appended, followed by splicing at offsets {0,0}, {-1,1}, {-2,2}, {-3,3}, {-3,3}, {-6,0} (refer to Fig. 1 in Peddinti et al., 2015) at each successive layer. The network is trained for 10 epochs using stochastic gradient descent updates. In addition, we train another network where the pre-final TDNN layer is replaced with self-attention (Povey et al., 2018). This method is referred to as *Self-Attention* in this work.

5.1.3. Language modeling

We estimate a tri-gram language model using the CMU dictionary as pronunciation lexicon and Witten-Bell smoothing (Witten and Bell, 1991) for unseen units. Since transcripts from entire collection of child speech corpora consisted of only 800K tokens, we experimented with additional transcripts from adult speech corpora (Librispeech, TEDLIUM, ICSI, Fisher and WSJ) containing $\approx$ 35M tokens. We attempted interpolating the two models, but found that child speech transcripts negatively impacted performance.

Table 3
Word error rates (%) for baseline models applied on CIDMIC and child speech portion from Forensic Interviews (FI) and Autism diagnosis session (ASD).

Model	TDNN	Self-Attention
CIDMIC	29.50	29.76
FI	63.33	62.53
ASD	76.69	76.23

We evaluate the TDNN and Self-Attention systems on the CIDMIC corpus and provide the results in Table 3. We note that the WER is comparable to other recent systems evaluated on a subset of CIDMIC in previous studies (Shivakumar et al., 2014; Shivakumar and Georgiou, 2018).

5.2. Baseline adult ASR

We use the ASPIRE model[2] as an off-the-shelf ASR system for adult speech. The model is trained on an augmented version of the Fisher English corpus, using an augmentation method similar to Section 5.1. The model uses a time-delay network with intermediate BLSTM layers as part of its acoustic model. We observed a WER of 26.18% (Forensic Interviews) and 33.15% (Autism Sessions) using this model, and use the obtained hypotheses for adaptation purpose.

5.3. Domain adaptations

For the first adaptation step, we extend both the adult and child baseline ASR models to their respective domains. Acoustic adaptation adapts baseline models to domain-specific channel conditions while language model adaptation assigns higher weights to domain-specific vocabulary. Similar to our previous work on domain adaptation (Kumar et al., 2017), we perform LM adaptation for the adult speech model and AM+LM adaptation for the child speech model. During LM adaptation, we estimate a new language model by interpolating the base model with an in-domain model following Section 4.3. N-gram weights from the interpolated model are used to re-score lattices generated during baseline decoding. This amounts to re-ranking lattice paths generated using the baseline model against computing a new set of paths from the audio and decoding graph. We choose this method since constructing the decoding graph can be a time-intensive process especially during multiple folds of validation. For the case of AM adaptation we train the baseline hybrid DNN-HMM model for a single epoch using the pre-trained network as an initialization point. Training is restricted to a single epoch to avoid over-fitting. Although alternative stopping criteria can be explored, we did not experiment with them since this was not the focus of the current study. We implement domain adaptation using a leave-one-session-out manner, wherein a particular session is used for evaluation and all other sessions are treated as adaptation data during model training and estimating optimal interpolation weight λ (Section 4.3). For the purpose of this work, domain adapted models represent a fine-tuned comparison against which we present our context adaptation improvements.

5.4. Context adaptation

We use the adult interlocutor's language within every session as the context source. We categorize context adaptation as *global:* use of all utterances in the session, and *local:* use of only utterances neighboring the current child utterance to be decoded. Global context helps understand the influence (if any) of commonly occurring phrases/concepts throughout the session, while local context makes a trade-off between the amount and quality of context. We group local context adaptation by the type: lexical and semantic (see Sections 4.1 and 4.2). For each type, we investigate the direction (forward/backward). Context direction helps understand the relative importance between child repetitions (forward) and clarifications by the adult (backward). For all of our experiments, we perform the adaptation using the interpolation method outlined in Section 4.3. We present our results using word-error-rate (WER) and language model perplexity.

6. Results and discussions

6.1. Domain adaptation

From Table 4, we note an overall improvement in word error rate and perplexity across the two domains considered. However, there exist a few differences between the domains. First, child speech from ADOS and BOSCC sessions has a significantly higher baseline WER and perplexity. The difference can be partly explained by the fact that idiosyncratic speech is a well-known

[2] http://kaldi-asr.org/models/m1 .

Table 4
Word-error rates and perplexity scores for domain adaptation. *The AM-adapted model was discarded and baseline model used instead for all further experiments.

Model	Forensic		ASD	
	WER (%)	PPL	WER (%)	PPL
Baseline	62.53	217.09	76.23	335.08
Domain AM	59.10	–	77.52*	–
Domain AM+LM	**56.10**	**156.16**	**73.94**	**234.06**

observation in speech of children with ASD, for example neologism (Volden and Lord, 1991) and atypical prosody (Grossman et al., 2010). Hence, spoken language abnormalities are an additional complexity over segmental acoustic variabilities associated with children's speech. Next, utterances from FI are significantly longer than ASD ($p < 0.001$) and with significantly higher signal-to-noise ratio (in dB, $p < 0.001$). The above reasons also explain why acoustic adaptation using audio from other sessions does not provide any improvement in WER for ASR. However, there exists a clear improvement in the case of FI which are relatively free of language abnormalities. LM adaptation provides a clear WER improvement, with the absolute increase more for FI. This is also reflected in considerable perplexity reduction. The performance difference between baseline and LM-adapted models are also reflective of adult speech corpora being used to estimate language models in the former.

6.2. Context adaptation

For both FI and ASD, although global context offers significant WER improvements over the baseline (Table 5), that is not the case with respect to domain-adapted models (Table 4). Hence, using all the adult's utterances in a session during adaptation does not provide any significant gains over domain knowledge. However, significant gains are observed in perplexity values for both FI and ASD domains. This suggests that while contextual language from the entire session offers an informative channel, the acoustic model might be unable to account for degraded audio conditions and hence not result in significant WER improvements. Alternatively, the trained speech acoustic model might be insufficient to capture variations in speech of children with pathological conditions. To get an estimate of upper bound to improvement from global context, we replace the adult ASR hypotheses used during adaptation with the ground truth transcripts (Session LM - Oracle). We do not observe significant WER improvement for this case either, suggesting that the ASR hypotheses are robust enough for the global adaptation.

6.3. Effect of context type

We recall that lexical repetitions uses n-grams directly from ASR hypotheses of adult context utterances while in semantic response generation, the context utterances are passed through a seq2seq model and n-grams from decoder hypotheses are used for adaptation. While analyzing the type of context adaptation in Table 6, we fix the number of context utterances in each direction at 3.

6.3.1. Context adaptation using lexical repetitions

We observe that local context adaptation consistently outperforms the out-of-domain baseline as well as fine-tuned domain adapted model. Both perplexity and WER show significant improvement for all directions in both FI and ASD. However, similar to domain adaptations, the magnitude of improvement is higher in FI, which suggests that context adaptation is still dependent on baseline performance and data complexity, i.e., a better performing baseline results in larger improvements from adaptation. Bidirectional adaptation (where adult ASR hypotheses from both directions are used for LM adaptation) outperforms individual directions in terms of both perplexity and WER. However, we note that this improvement is not statistically significant over any individual direction. Moreover the amount of context available is twice than either forward/backward, hence a straightforward

Table 5
Word-error rates and perplexity scores for global context adaptation using ASR hypotheses (Session LM) and ground truth transcripts (Session LM - Oracle).

Model	Forensic		ASD	
	WER (%)	PPL	WER (%)	PPL
Baseline	62.53	217.09	76.23	335.08
Session LM	55.79	141.18	73.79	204.86
Session LM - Oracle	**55.51**	**131.11**	**73.42**	**182.31**

Table 6
Word-error rates and perplexity scores for utterance-level adaptation. For each method and corpus, results are reported for forward (F), backward (B) and bidirectional (Bi) directions of context adaptations. GT-Oracle represents adaptation using ground truth transcripts.

Method	Forensic						ASD					
	WER (%)			PPL			WER (%)			PPL		
	F	B	Bi	F	B	Bi	F	B	Bi	F	B	Bi
Baseline		62.53			217.09			76.23			335.08	
AM+LM Domain Adapted		56.10			156.16			73.94			234.06	
GT-Oracle	52.13	52.28	51.60	129.49	140.8	125.70	70.31	70.34	69.86	190.33	199.28	176.32
Lexical	52.32	52.38	51.82	133.43	142.63	131.45	70.54	70.47	70.32	202.66	202.26	187.89
Semantic	52.13	52.26	–	148.87	148.59	–	70.29	70.75	–	217.35	214.89	–
Combined	52.05	52.02	–	134.12	143.49	–	70.31	70.45	–	195.47	206.29	–

comparison may not be appropriate. This suggests that both repetitions by the child (response) and adult (clarifications, follow-up, etc) are equally important when it comes to effect on adaptation.

6.3.2. Context Adaptation using Semantic Response Generation:

Using decoder outputs from the seq2seq model shows improvement similar to, but slightly less than the lexical adaptation for both perplexity and WER, and both FI and ASD. We note two important differences in the way multiple context utterances are handled by the lexical and semantic adaptations. First, lexical adaptation uses the entirety of context utterances and in an order-agnostic manner (since only n-gram counts are borrowed). The seq2seq model is restricted by the maximum number of time-steps in the encoder when multiple utterances are concatenated. For instance, during inference for child utterance C_t using forward context adaptation of length 3, the first word from utterance P_{t-2} is input at the first encoder timestep. Utterances P_{t-1} and P_t are concatenated to P_{t-2} and input to successive encoder timesteps. Words from this concatenated utterance after the 20th timestep will be truncated. Hence, the amount of context actually seen by the seq2seq model may be smaller than the context length. Second, the current seq2seq architecture cannot incorporate both directions of context in the encoder in an unbiased manner, hence rendering bidirectional adaptation not possible. Nevertheless, semantic adaptation provides significant improvement over the domain-adapted models.

To investigate any complementary information from the lexical and semantic models, we combine them at the hypothesis level. We augment the adult ASR hypotheses (lexical) with the outputs from seq2seq model (semantic) and present results for the combined method in Table 6. We observe marginal improvement from combined adaptation in majority of cases, while even exceeding the oracle inputs with respect to WER for FI. Hence, while combined adaptation may not always be optimal, preliminary findings hold promise.

6.4. Effect of context size

We study the effect of amount of context used for adaptation, measured using the number of context utterances. A larger context window can contain both useful and non-useful information, and hence may not necessarily be useful for adaptation. In Fig. 5, we provide results for absolute perplexity and WER for lexical adaptation. As mentioned above in Section 6.3, increasing context during semantic adaptation has a high possibility of truncating utterances closer to the target, hence we do not present results for the semantic or combined adaptations. In all cases, we vary the number of context utterances from 1 to 10, again noting that the bidirectional case receives twice the amount of context compared to forward/backward directions. For comparison, we repeat the experiments using oracle transcripts.

We observe that the perplexity for FI shows a slight degradation with increase in context, suggesting that the noisy information dominate during adaptation. This is however not the case with backward adaptation, where there is no clear increase/decrease. Both oracle transcripts and ASR hypotheses result in the same effects, although the former results in reduced perplexity in most cases. A small context window (2 or smaller) seems to be optimal for both measures - perplexity and WER. In case of ASD, we notice that a window length of 2 provides the best perplexity values, either saturating or worsening with larger context windows.

6.5. Effect of external factors

We perform a statistical analysis on the adaptation results obtained from Section 5.4. We analyze the effect of external factors such as duration of child utterance, age of child and WER of corresponding adult context utterances on the adaptation results. For each factor, we compare the change in child ASR performance (perplexity/WER) against the baseline using one-way analysis-of-variance (ANOVA) with the null hypothesis representing no change in performance distribution. We encode the dependent variable into three levels, representing increase, no-change and decrease respectively. We categorize the age into two levels - young ($<$ 8yrs) and old ($>$ 8yrs). Results are presented in the form of significant factors according to p-values for each adaptation method in Table 7.

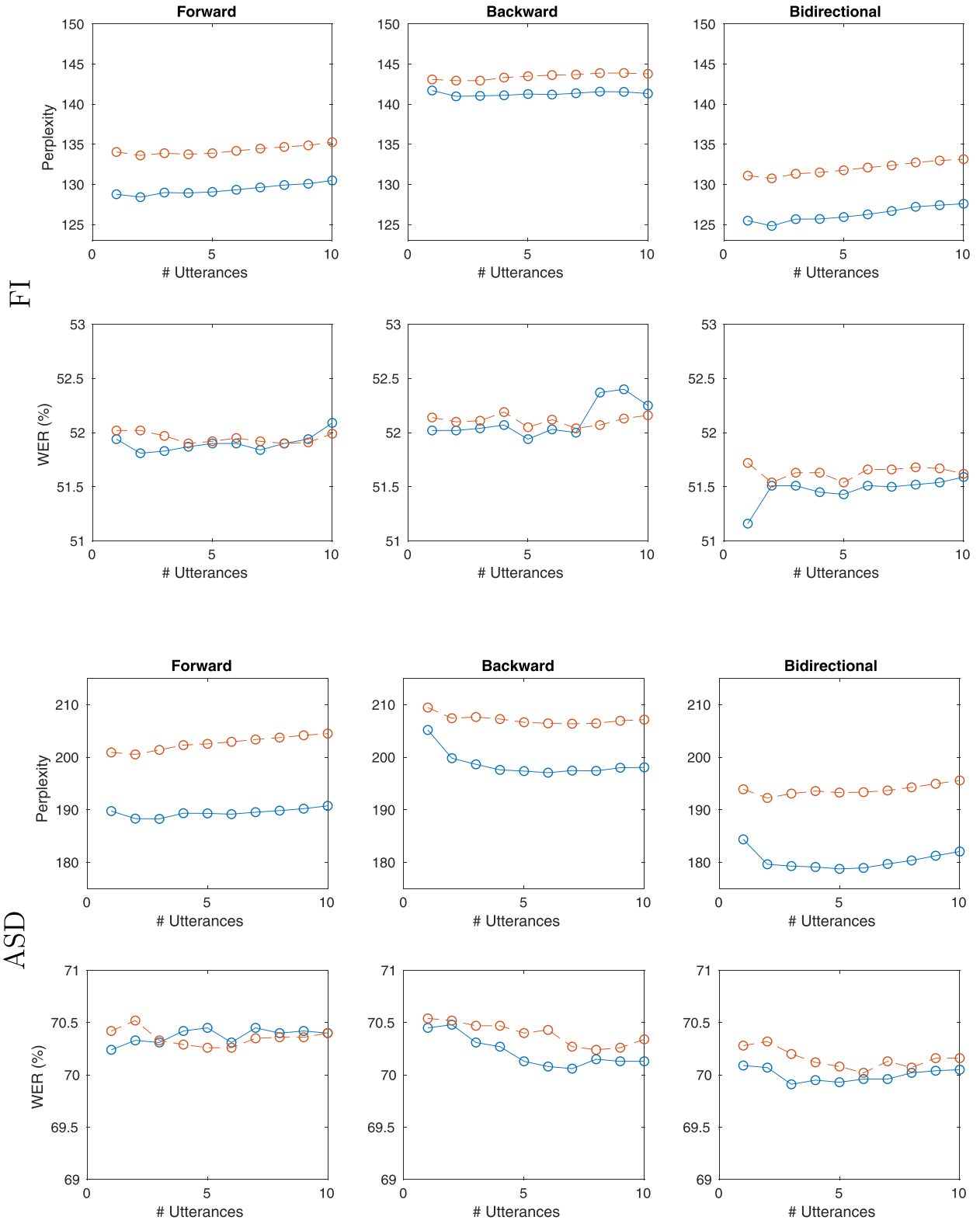

Fig. 5. Effect of number of context utterances on the perplexity and WER. For each case, upto 10 context utterances are used. Results are provided using both oracle transcripts (Blue, continuous) and ASR hypotheses (Red, dashed).

Table 7
Effect of utterance duration (U), child age (A) and adult WER (W) on the adaptation performance measured using WER and perplexity. Each entry presents the statistically significant factors ($p < 0.1$, $^*p < 0.01$) as determined by ANOVA.

Direction	Type	WER		Perplexity	
		FI	ASD	FI	ASD
–	Domain	U*,A	U*,A	W*	–
	Session-Global	U*,A	U*,W	–	U,W
Forward	Lexical	U*,A	U*	U	U*
	Semantic	U*,W	U*	U	U*,A
	Combined	U*	U*	U	U*,A
Backward	Lexical	U*	U*	U	U*
	Semantic	U*,A	U*	U	U*
	Combined	U*,A	U*	U	U*

Across different adaptation methods and domains, the duration of child utterance is a significant factor in perplexity improvement and strongly significant in WER. Overall, shorter utterances resulted in degradation of performance when compared to longer utterances. This raises a significant issue, since speech utterances by the child are typically shorter and consisting of fewer words than their adult interlocutors. Both age and adult WER do not play a dominant role for utterance adaptation, appearing only in selected cases under semantic or combined adaptation. A notable exception is during domain adaptation. WER of adult transcripts plays a significant role for improving perplexity and not WER. In this case, source factors (interlocutor: adult WER) dominate for perplexity improvement while target factors (child: utterance duration, child age) dominate during WER improvement. This observation follows from the fact that target factors directly influence acoustic variability, which is captured through WER and not perplexity.

7. Conclusions

In this work, we show that the adult interlocutor's spoken language is useful in improving child speech recognition accuracy in a child-adult dyadic interaction setting. We make use of two semi-structured but spontaneous child speech application domains to motivate and evaluate the proposed context modeling - forensic interviews and play-based, interactive sessions for children with ASD. Traditionally considered a challenging problem, we describe the development of a robust child ASR system built on top of state-of-the art models designed for adult speech. We demonstrate methods to extract lexical and semantic contextual information from the adult speech hypotheses extracted using an ASR system. We show that even few utterances from the immediate vicinity of the target utterance provide significant gains in performance as compared to session-level context. We further investigate the effect of direction and number of context utterances, noting that the seq2seq model is limited by the number of encoder timesteps. Combining hypotheses from the seq2seq model with lexical repetitions results in highest performance for majority of test conditions. We do not observe significant difference between adaptations using oracle transcripts and ASR hypotheses, emphasizing the robustness of our models to transcription errors from the adult ASR system. We also consider the effect of source-based factors (originating from interlocutor: adult WER) and target-based factors (originating from the child: utterance duration, chronological age) separately on the performance improvement using statistical analysis. We find that while utterance duration is a dominant factor during majority of adaptation conditions, improvements during domain adaptation are found to be influenced by target-based factors and source-based factors, respectively.

In the future, we are interested in automatically learning adaptation weights (possibly unique to each n-gram) to minimize WER without using a held-out set, considering limited availability of in-domain data. In the case of semantic response generation, it would be useful to learn to select context adult utterances relevant for each target child utterance. This would require both incorporating long-term context (using longer encoder timesteps or hierarchical networks) and including an attention mechanism in the seq2seq network. Considering the high baseline WER, further work will also continue to focus on developing a robust generic child ASR.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the Simons Foundation SFARI Research Award RFA #345327 awarded to Catherine Lord, Shrikanth Narayanan and So Hyun Kim.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. Tensorflow: A system for large-scale machine learning. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265–283.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N., 2013. Abstract Meaning Representation for Sembanking. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. Association for Computational Linguistics, Sofia, Bulgaria, pp. 178–186.

Bell, L., Boye, J., Gustafson, J., Heldner, M., Lindström, A., Wirén, M., 2005. The Swedish NICE corpus—spoken dialogues between children and embodied characters in a computer game scenario. In: Proceedings of the Eurospeech, 9th European Conference on Speech Communication and Technology, pp. 2765–2768.

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. 5 (2), 157–166.

Black, M.P., Tepperman, J., Narayanan, S.S., 2011. Automatic prediction of children's reading ability for high-level literacy assessment. IEEE Trans. Audio Speech Lang. Process. 19 (4), 1015–1028.

Bone, D., Lee, C., Chaspari, T., Gibson, J., Narayanan, S., 2017. Signal processing and machine learning for mental health research and clinical applications [perspectives]. IEEE Signal Process. Mag. 34 (5), 195–196.

Bone, D., Lee, C., Potamianos, A., Narayanan, S.S., 2014. An investigation of vocal arousal dynamics in child-psychologist interactions using synchrony measures and a conversation-based model. In: Proceedings of Interspeech, 15th Annual Conference of the International Speech Communication Association, pp. 218–222.

Burnett, D.C., Fanty, M., 1996. Rapid unsupervised adaptation to children's speech on a connected-digit task. In: Proceedings of the 4th International Conference on Spoken Language Processing. ICSLP '96, 2, pp. 1145–1148.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp. 1724–1734.

Christensen, S., Johnsrud, S., Ruocco, M., Ramampiaro, H., 2018. Context-aware sequence-to-sequence models for conversational systems. arXiv:1805.08455

Claes, T., Dologlou, I., ten Bosch, L., van Compernolle, D., 1998. A novel feature transformation for vocal tract length normalization in automatic speech recognition. IEEE Trans. Speech Audio Process. 6, 549–557.

Cole, R., Hosom, P., Pellom, B., 2006. University of Colorado Prompted and Read Childrens Speech Corpus. Technical Report. Technical Report TR-CSLR-2006-02, University of Colorado.

Cole, R., Massaro, D.W., Villiers, J.d., Rundle, B., Shobaki, K., Wouters, J., Cohen, M., Baskow, J., Stone, P., Connors, P., et al., 1999. New tools for interactive speech and language training: using animated conversational agents in the classroom of profoundly deaf children. MATISSE-ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., Baroni, M., 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, pp. 2126–2136.

Das, S., Nix, D., Picheny, M., 1998. Improvements in children's speech recognition performance. In: Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, 1, pp. 433–436.

Elenius, D., Blomberg, M., 2005. Adaptation and normalization experiments in speech recognition for 4 to 8 year old children. In: Proceedings of Eurospeech, 9th European Conference on Speech Communication and Technology, pp. 2749–2752.

Erdogan, H., Sarikaya, R., Chen, S.F., Gao, Y., Picheny, M., 2005. Using semantic analysis to improve speech recognition performance. Comput. Speech Lang. 19 (3), 321–343.

Eskenazi, M., Mostow, J., Graff, D., 1997. The CMU kids corpus. Linguist. Data Consort..

Fan, A., Lewis, M., Dauphin, Y., 2018. Hierarchical neural story generation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, pp. 889–898.

Firat, O., Cho, K., Bengio, Y., 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, pp. 866–875.

Gerosa, M., Giuliani, D., Narayanan, S., 2006. Acoustic analysis and automatic recognition of spontaneous children's speech. In: Proceedings of the 9th International Conference on Spoken Language Processing, pp. 1886–1889.

Gerosa, M., Giuliani, D., Narayanan, S., Potamianos, A., 2009. A review of ASR technologies for children's speech. In: Proceedings of the 2nd Workshop on Child, Computer and Interaction, pp. 1–8.

Gildea, D., Hofmann, T., 1999. Topic-based language models using EM. In: Proceedings of the 6th European Conference on Speech Communication and Technology, pp. 2167–2170.

Graves, A., 2012. Supervised Sequence Labelling. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 5–13.

Graves, A., Jaitly, N., 2014. Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of the 31st International Conference on Machine Learning, 32, pp. 1764–1772.

Graves, A., Mohamed, A., Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In: Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649.

Gray, S.S., Willett, D., Lu, J., Pinto, J., Maergner, P., Bodenstab, N., 2014. Child automatic speech recognition for US English: child interaction with living-room-electronic-devices. In: Proceedings of the 4th Workshop on Child, Computer and Interaction, pp. 21–26.

Grossman, R.B., Bemis, R.H., Skwerer, D.P., Tager-Flusberg, H., 2010. Lexical and affective prosody in children with high-functioning autism. J. Speech Lang. Hear. Res. 53 (3), 778–793.

Gruenstein, A., Wang, C., Seneff, S., 2005. Context-sensitive statistical language modeling. In: Proceedings of the 9th European Conference on Speech Communication and Technology, pp. 17–20.

Grzadzinski, R., Carr, T., Colombi, C., McGuire, K., Dufek, S., Pickles, A., Lord, C., 2016. Measuring changes in social communication behaviors: preliminary development of the brief observation of social communication change (BOSCC). J. Autism Dev. Disord. 46 (7), 2464–2479.

Hagen, A., Pellom, B., Cole, R., 2003. Children's speech recognition with application to interactive books and tutors. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 186–191.

Hazan, V., Barrett, S., 2000. The development of phonemic categorization in children aged 6–12. J. Phon. 28 (4), 377–396.

Hershkowitz, I., Fisher, S., Lamb, M.E., Horowitz, D., 2007. Improving credibility assessment in child sexual abuse allegations: the role of the NICHD investigative interview protocol. Child Abuse Neglect 31 (2), 99–110.

Hershkowitz, I., Lamb, M.E., Katz, C., 2014. Allegation rates in forensic child abuse investigations: comparing the revised and standard NICHD protocols.. Psychol. Public Policy Law 20 (3), 336–344.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. 29 (6), 82–97.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 42 (1), 177–196.

Ippolito, D., Kriz, R., Sedoc, J., Kustikova, M., Callison-Burch, C., 2019. Comparison of diverse decoding methods from conditional language models. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, pp. 3752–3762.

Johnson, W.L., Rickel, J.W., Lester, J.C., 2000. Animated pedagogical agents: face-to-face interaction in interactive learning environments. Int. J. Artif. Intell. Educ. 11, 47–78.

Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., Khudanpur, S., 2017. A study on data augmentation of reverberant speech for robust speech recognition. In: Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5220–5224.
Kogan, M.D., Vladutiu, C.J., Schieve, L.A., Ghandour, R.M., Blumberg, S.J., Zablotsky, B., Perrin, J.M., Shattuck, P., Kuhlthau, K.A., Harwood, R.L., Lu, M.C., 2018. The prevalence of parent-reported autism spectrum disorder among US children. Pediatrics 142 (6). https://doi.org/10.1542/peds.2017-4161.
Konstas, I., Iyer, S., Yatskar, M., Choi, Y., Zettlemoyer, L., 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada, pp. 146–157.
Kumar, M., Bone, D., McWilliams, K., Williams, S., Lyon, T.D., Narayanan, S.S., 2017. Multi-scale context adaptation for improving child automatic speech recognition in child-adult spoken interactions. In: Proceedings of Interspeech, 18th Annual Conference of the International Speech Communication Association, pp. 2730–2734.
Kumar, M., Gupta, R., Bone, D., Malandrakis, N., Bishop, S., Narayanan, S.S., 2016. Objective language feature analysis in children with neurodevelopmental disorders during autism assessment. In: Proceedings of Interspeech, 17th Annual Conference of the International Speech Communication Association, pp. 2721–2725.
La Rooy, D., Brubacher, S.P., Aromäki-Stratos, A., Cyr, M., Hershkowitz, I., Korkman, J., Myklebust, T., Naka, M., Peixoto, C.E., Roberts, K.P., et al., 2015. The NICHD protocol: a review of an internationally-used evidence-based tool for training child forensic interviewers. J. Criminol. Res. Policy Pract. 1 (2), 76–89.
Lamb, M.E., Orbach, Y., Hershkowitz, I., Esplin, P.W., Horowitz, D., 2007. A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: areview of research using the NICHD investigative interview protocol. Child Abuse & Neglect 31 (11–12), 1201–1231.
Lee, C.-C., Mower, E., Busso, C., Lee, S., Narayanan, S., 2011. Emotion recognition using a hierarchical binary decision tree approach. Speech Commun. 53 (9), 1162–1171.
Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of childrens speech: developmental changes of temporal and spectral parameters. J. Acoust. Soc. Am. 105 (3), 1455–1468.
Lee, S., Potamianos, A., Narayanan, S., 2014. Developmental acoustic study of american english diphthongs. J. Acoust. Soc. Am. 136 (4), 1880–1894.
Leonard, R., 1984. A database for speaker-independent digit recognition. In: Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, 9, pp. 328–331.
Liao, H., Pundak, G., Siohan, O., Carroll, M., Coccaro, N., Jiang, Q.-M., Sainath, T.N., Senior, A., Beaufays, F., Bacchiani, M., 2015. Large vocabulary automatic speech recognition for children. In: Proceedings of Interspeech, 16th Annual Conference of the International Speech Communication Association, pp. 1611–1615.
Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Leventhal, B.L., DiLavore, P.C., Pickles, A., Rutter, M., 2000. The autism diagnostic observation schedule—Generic: astandard measure of social and communication deficits associated with the spectrum of autism. J. Autism Dev. Disord. 30 (3), 205–223.
Lyon, T.D., 2014. Interviewing children. Ann. Rev. Law Soc. Sci. 10, 73–89.
Marcus, G.F., Pinker, S., Ullman, M., Hollander, M., Rosen, T.J., Xu, F., Clahsen, H., 1992. Overregularization in language acquisition. Monogr. Soc. Res. Child Dev. I–178.
Metallinou, A., Katsamanis, A., Narayanan, S., 2012. A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs. In: Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 2401–2404.
Metallinou, A., Wollmer, M., Katsamanis, A., Eyben, F., Schuller, B., Narayanan, S., 2012. Context-sensitive learning for enhanced audiovisual emotion classification. IEEE Trans. Affect. Comput. 3 (2), 184–198.
Michaely, A.H., Zhang, X., Simko, G., Parada, C., Aleksic, P., 2017. Keyword spotting for Google assistant using contextual speech recognition. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, pp. 272–278.
Morency, L.-P., de Kok, I., Gratch, J., 2008. Context-based recognition during human interactions: automatic feature selection and encoding dictionary. In: Proceedings of the 10th international conference on Multimodal interfaces. ACM, pp. 181–188.
Morrison, S.E., Bruce, C., Wilson, S., 2018. Childrens disclosure of sexual abuse: a systematic review of qualitative research exploring barriers and facilitators. J. Child Sex. Abus. 27 (2), 176–194.
Mostow, J., Roth, S.F., Hauptmann, A.G., Kane, M., 1994. A prototype reading coach that listens. In: Proceedings of the 12th National Conference on Artificial Intelligence, 1, pp. 785–792.
Mugitani, R., Hiroya, S., 2012. Development of vocal tract and acoustic features in children. Acoust. Sci. Technol. 33 (4), 215–220.
Narayanan, S., Georgiou, P., 2013. Behavioral signal processing: deriving human behavioral informatics from speech and language. Proc. IEEE 101 (5), 1203–1233.
Narayanan, S., Potamianos, A., 2002. Creating conversational interfaces for children. IEEE Trans. Speech Audio Process. 10 (2), 65–78.
Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., Ward, R., 2016. Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (4), 694–707.
Patel, A., Li, D., Cho, E., Aleksic, P., 2018. Cross-lingual phoneme mapping for language robust contextual speech recognition. In: Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5924–5928.
Peddinti, V., Povey, D., Khudanpur, S., 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In: Proceedings of Interspeech, 16th Annual Conference of the International Speech Communication Association, pp. 3214–3218.
Perez-Marn, D., Pascual-Nieto, I., 2013. An exploratory study on how children interact with pedagogic conversational agents. Behav. Inf. Technol. 32 (9), 955–964.
Potamianos, A., Narayanan, S., 1998. Spoken dialog systems for children. In: Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, 1. IEEE, pp. 197–200.
Potamianos, A., Narayanan, S., 2003. Robust recognition of children's speech. IEEE Trans. Speech Audio Process. 11 (6), 603–616.
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The kaldi speech recognition toolkit. In: Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding.
Povey, D., Hadian, H., Ghahremani, P., Li, K., Khudanpur, S., 2018. A time-restricted self-attention layer for ASR. In: Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5874–5878.
Raju, A., Hedayatnia, B., Liu, L., Gandhe, A., Khatri, C., Metallinou, A., Venkatesh, A., Rastrow, A., 2018. Contextual language model adaptation for conversational agents. In: Proceedings of Interspeech, 19th Annual Conference of the International Speech Communication Association, pp. 3333–3337.
Serban, I.V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A.C., Bengio, Y., 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 3295–3301.
Serholt, S., Barendregt, W., 2016. Robots tutoring children: Longitudinal evaluation of social engagement in child-robot interaction. In: Proceedings of the 9th Nordic Conference on Human-Computer Interaction, pp. 1–10.
Serizel, R., Giuliani, D., 2014. Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. In: Proceedings of the IEEE Spoken Language Technology Workshop (SLT), pp. 135–140.
Serizel, R., Giuliani, D., 2017. Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. Nat. Lang. Eng. 23 (3), 325–350.
Shivakumar, P., Potamianos, A., Lee, S., Narayanan, S., 2014. Improving speech recognition for children using acoustic adaptation and pronunciation modeling. In: Proceedings of the Workshop on Child, Computer and Interaction, pp. 15–19.
Shivakumar, P. G., Georgiou, P., 2018. Transfer learning from adult to children for speech recognition: evaluation, analysis and recommendations. arXiv:1805.03322.
Shobaki, K., Hosom, J.-P., Cole, R.A., 2000. The OGI kids speech corpus and recognizers. In: Proceedings of the 6th International Conference on Spoken Language Processing, pp. 258–261.
Solsona, R.A., Fosler-Lussier, E., Kuo, H.J., Potamianos, A., Zitouni, I., 2002. Adaptive language models for spoken dialogue systems. In: Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, 1, pp. 37–40.
Song, L., Gildea, D., Zhang, Y., Wang, Z., Su, J., 2019. Semantic neural machine translation using AMR. Trans. Assoc. Comput.Linguist. 7, 19–31.

Stemmer, G., Hacker, C., Steidl, S., Nöth, E., 2003. Acoustic normalization of children's speech. In: Proceedings of Eurospeech, 8th European Conference on Speech Communication and Technology, pp. 1313–1316.
Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. Advances in neural information processing systems, pp. 3104–3112.
Vinyals, O., Le, Q., 2015. A neural conversational modelarXiv:1506.05869.
Visweswariah, K., Printz, H., 2001. Language models conditioned on dialog state. In: Proceedings of Eurospeech, 7th European Conference on Speech Communication and Technology, pp. 251–254.
Volden, J., Lord, C., 1991. Neologisms and idiosyncratic language in autistic speakers. J. Autism Dev. Disord. 21 (2), 109–130.
Vorperian, H.K., Kent, R.D., Lindstrom, M.J., Kalina, C.M., Gentry, L.R., Yandell, B.S., 2005. Development of vocal tract length during early childhood: a magnetic resonance imaging study. J. Acoust. Soc. Am. 117 (1), 338–350.
Ward, N., 1996. Using prosodic clues to decide when to produce back-channel utterances. In: Proceedings of the 4th International Conference on Spoken Language, 3. IEEE, pp. 1728–1731.
Ward, N., Tsukahara, W., 2000. Prosodic features which cue back-channel responses in english and japanese. J. Pragmat. 32 (8), 1177–1207.
Williams, I., Kannan, A., Aleksic, P.S., Rybach, D., Sainath, T.N., 2018. Contextual speech recognition in end-to-end neural network systems using beam search. In: Proceedings of Interspeech, 19th Annual Conference of the International Speech Communication Association, pp. 2227–2231.
Witten, I.H., Bell, T.C., 1991. The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. IEEE Trans. Inf. Theory 37 (4), 1085–1094.
Yan, R., Song, Y., Wu, H., 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM, pp. 55–64.
Yang, Z., Metallinou, A., Narayanan, S., 2014. Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues. IEEE Trans. Multimed. 16 (6), 1766–1778.
Yu, K., Li, H., Oguz, B., 2018. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In: Proceedings of the 3rd Workshop on Representation Learning for NLP, pp. 175–179.