Speech Communication 140 (2022) 98–108

Contents lists available at ScienceDirect

Speech Communication

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

Assessing child communication engagement and statistical speech patterns for American English via speech recognition in naturalistic active learning spaces

Rasa Lileikyte [a], Dwight Irvin [b], John H.L. Hansen [a,*]

[a] Center for Robust Speech Systems, University of Texas at Dallas, 2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, USA
[b] Juniper Gardens Children's Project, University of Kansas, Kansas City, KS, USA

ARTICLE INFO

Keywords:
Automatic speech recognition
Speech engagement
Data augmentation
Child naturalistic speech
Classroom setting

ABSTRACT

Assessing child growth in terms of speech and language development is a critical indicator of long term learning ability and life-long development progress. The earlier a child who is at-risk is identified, the earlier support can be provided to reduce the social impact of the speech or language issue. The preschool classroom provides an opportunity for monitoring growth in young children's interactions. To date, limited research has been possible for young child based speech recognition in classroom settings due to speech data access, as well as limitations on speech recognition performance for naturalistic child communication. This study addresses American English speech recognition for children's speech in a naturalistic noisy early childhood setting, where child age varies from 3 to 5 years. This study investigates the effectiveness of data augmentation techniques to improve both language and acoustic models, since this is relatively under explored for young child speech. We consider alternate text augmentation approaches using adult data, Web data, and text generated by recurrent neural networks. We also compare several acoustic augmentation techniques including: speed perturbation, tempo perturbation, and adult data. In the study, we also comment on child word count rates to assess child speech development. Finally, insights are provided into the statistical patterns of naturalistic child speech such as word complexity, stop words, part of speech, etc., which are intended to serve as a representative of high quality language engagement in adult–child learning environments.

1. Introduction

A supportive language environment generally includes gains in communication and social skills of typically developing and children at risk (e.g., speech or language delayed) (Burchinal et al., 2008). The amount of teacher and peer language children are exposed to in preschool childhood settings contribute to essential language developmental outcomes in early childhood. Adult and peer language has been shown to be even more beneficial for children at risk (Brown et al., 1999; Perry et al., 2018), since it plays a role in improving social communication (Warren and Yoder, 2004). Therefore, identifying which children are at-risk for low language engagement is of high importance so that these children could, in turn, receive more teacher support in classroom settings. Limited research has been performed employing speech technologies to measure child speech recognition when speech is recorded in naturalistic learning environments. Advancements based on LENA Foundation[1] have shown that measuring child speech interaction is possible, but research has shown that in early education spaces (Dykstra et al., 2012; Gilkerson et al., 2017; Irvin et al., 2017), where children and adults are mobile with diverse noise/acoustic conditions, and conversational interactions are often spontaneous and unscripted, this and other prior technologies run into significant challenges in diarizating speech.

It is known that automatic speech recognition (ASR) for child conversational speech is more challenging than for adults, specifically because of the developing language planning, physiology, and motor skills of young speakers. Basically, children do not necessary have the mature language/grammar skills to consistently form coherent logical sentences, or accurately articulate each phoneme when proper word selection is possible. Moreover, there is a lack of available child speech corpora to advance such speech technology. The diversity of child speech production physiology causes issues as well, since speaking traits can vary significantly from child to child who are typically developing, as well as those that might be at-risk. Children's speech structure within the age range of 3–5 years differs significantly from 6–18 year old speakers. Most prior child speech recognition efforts have

* Corresponding author.
E-mail addresses: rasa.lileikyte@utdallas.edu (R. Lileikyte), john.hansen@utdallas.edu (J.H.L. Hansen).

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

https://doi.org/10.1016/j.specom.2022.01.006
Received 28 June 2021; Received in revised form 10 November 2021; Accepted 17 January 2022
Available online 12 March 2022

focused on the older child group. Children in the age range of 3–5 have reduced vocal system physiologies, they are still developing their speech motor skills, pronunciation, and vocabulary. Young children do not necessarily follow adult grammar rules and proper linguistic structure. In the studies by Justice et al. (2013) and Hart and Risley (1995), the language interaction traits such as child word count rate was shown to be important in the early stages of language development. Not as much work as considered quantifying the language development knowledge/experience in Wernicke's area. Also, not all ASR recognition errors are equally important. Thus, there is a need to analyze statistical patterns of naturalistic child speech, that could serve as a sample of high quality language engagement in adult–child learning spaces.

While research has considered child ASR in the past, most studies focus on:

- older child speech (6–18 age group) (Gerosa et al., 2009; Fainberg et al., 2016);
- child read speech, and structured human–computer interaction speech;

Only a few studies have explored preschool child speech recognition, using words, phrases, and structured human–computer interaction scenario (Smith et al., 2017; Kothalkar et al., 2018a,b; Hagen et al., 2003). In our study, the scenario is based on American English naturalistic conversational interaction between child–adult and child–child in the preschool classroom, while wearing LENA recorders. We investigate young child ASR, where age varies from 3 to 5 years. The extensive work from the LENA Foundation has investigated naturalistic speech of preschoolers, but always one-on-one scenarios where one child is with one adult, normally in fairly quiet spaces. This prior effort has not considered ASR, since their language assessment strategy only estimates word count based on phoneme change sequence counting (Gilkerson et al., 2017; Dykstra et al., 2013; Soderstrom and Wittebolle, 2013; Ota and Austin, 2013).

The corpus used in our study is comprised of 15 h of transcribed children audio for training. This task is very challenging, where past studies suggest it is common to experience high word error rates (WERs) in such ASR conditions. For example, on large 2000 h corpus, adult conversational speech recognition yields 11% WER (Hadian et al., 2018). Meanwhile, systems with 3 h and 40 h training sets, achieve 52% and 42% WER, respectively (Lileikytė et al., 2018, 2017; Hartmann et al., 2017). Therefore, training data size and speaker diversity significantly impact system performance.

The motivational aims of this study are as follows:

- Investigate young child (age from 3 to 5 years) naturalistic American English speech recognition, when speech was recorded in active learning spaces. Meanwhile most of the past studies explore older child (6–18 years) read speech/directed computer interactive systems for child reading. Naturalistic speech offer significantly different acoustic conditions as well as context based voice exchanges;
- Assess the effect of data augmentation techniques for child speech under low resource conditions, when only 15 h of children transcribed speech is used;
- Explore if word count rates can provide insight to help separate at-risk and typically developing children.
- Analyze statistical patterns of naturalistic child speech. Our aim is to explore the most common 100 words spoken by children: what has been said, and what is distribution of the word types/categories.

It is specifically stated that the goal is not to achieve child ASR WER's similar to adult WER's, since that may not be feasibly possible given a child's language/grammar and pronunciation abilities.

This study is the extension of our previous work (Lileikyte et al., 2020). New research results are reported, providing long short-term memory (LSTM) results with augmented data, WER analysis for each child, and a unique first analysis of statistical patterns of naturalistic child speech.

In this study: *(i) we explore data augmentation techniques for young child naturalistic speech recognition.* Data augmentation has been shown to consistently improve performance of adult ASR systems. However, it has not been extensively studied for child naturalistic speech. To cope with a limited amount of child training data, previous studies have explored the use of adult speech (e.g., in Fainberg et al. (2016), Smith et al. (2017) and Serizel and Giuliani (2014)). Children's speech between 3 to 5 years differs significantly from adult speech. As such, migrating adult based speech technologies towards this young child population is significantly more challenging. In our work, we explore (1) language model augmentation via text generated using recurrent neural networks (RNNs) (Mikolov and Zweig, 2012), and (2) acoustic model augmentation using speed and tempo perturbation (Ko et al., 2015). These techniques have been used for adult speech augmentation (Mikolov and Zweig, 2012; Ko et al., 2015). However, to the best of our knowledge, this work is perhaps one of the first to study these approaches for child speech. Our earlier study (Lileikyte et al., 2020) did explore initial approaches for data augmentation with promising results. Here we compare these techniques and the use of adult data as well.

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

(iii) Finally, statistical patterns of naturalistic child speech are explored. Specifically, we consider which words carry more meaningful information for adult–child interaction, and therefore are more important for automatic recognition. The analyses of statistical patterns of naturalistic child speech are intended to serve as a sample of high quality language engagement in adult–child learning environment.

2. Challenges in child speech recognition for naturalistic scenario

Speech recognition for preschool children (3–5 yrs) within a naturalistic classroom setting is more challenging versus older children speech recognition, where most prior child speech recognition efforts have focused on (e.g., 6–18 yrs). Children in the 3–5 year range are still developing pronunciation and grammar/language understanding for proper sentence formulation. They have significantly different acoustic speech production/model experience (see Table 1). These differences are attributed mainly to anatomical and morphological differences in the vocal tract geometry, less precise control of the articulators, and the inability to control features such as prosody. They are also associated with less formed knowledge and experience in grammar and vocabulary associated with Wernicke's area. A child's vocal system is smaller than an adult's. The physiological development, both present and as they grow, contribute to higher variability of child speakers, resulting in gender (male/female) assessment as a difficult task. Children voices when compared to those of adults include higher fundamental frequencies, shifted formant frequencies due to reduced vocal tract lengths, and greater spectral variability. Children display higher variability in speaking rate, vocal effort, and degree of spontaneity. Children's spontaneous speech may be ill-formed, or include incomplete sentences. Example phrases illustrating some common phenomena found in casual child speech are given in Table 2. Children are also more likely to seamlessly change conversational topics rapidly. Disfluencies such as filled pauses, repetitions, repairs, and false starts are frequent in conversational child speech. Paralinguistic events are more common for children speech, such as non-speech vocalizations including laughter, crying, shouting, yawning, coughing, or sneezing. Due to a lack of communication

Fig. 1. A typical high quality childcare learning center.

Table 1
Child speech development in different age groups.

Child age	Lexicon	Language model	Acoustic model
3–5 years	Still developing	Still developing	Still developing
6–18 years	Developed	Developed	Developed

Table 2
Examples of conversational child speech phrases.

Event	Example
Hesitations	*hmm* but the prize box is happy okay pretend *hmm* the blocks can walk
Filler words	*yeah* does not match because we got pink *oh yeah* that i thought you said sock socks they use socks
Paranormal	*[gasp]* i am going to give *[gasp]* if i say your name kitty meow *[cough]* meow come on kitties lets go
Word fragments	can i have some jelly on my *bisc-* biscuit that *trea-* treasure box is right here
Word repetitions	well *if you if you* are a vampire you cannot play with us *do do* you want to play something with me

experience, speech overlap is also frequent in child speech. While an adult may follow socially accepted traditional conversational turn-taking protocol, children typically speak with little regard to expected sentence or turn-taking boundaries. The classroom setting is also noisy, containing large crowd noise, babble noise, as well as competing speech (see Fig. 1).

3. Related work

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

Most corpora containing children's speech focus on the 6–18 age group, and are either limited in scope or are task directed. Corpora such as the American English CID children corpus (Lee et al., 1999), KIDS (Eskenazi et al., 1997), CU Kids' Audio Speech corpus (Hagen et al., 2003), and PF-STAR (British English, Italian, German and Swedish) (Batliner et al., 2005) are all read speech corpora. As for spontaneous speech, corpora include children–machine interaction scenarios (e.g., computer based education or speech training systems) rather than child–adult naturalistic scenarios. The NICE database consists of spoken dialogue interaction between children and animated characters in a game setting (Bell et al., 2005). In Batliner et al. (2004), a child–robot interaction corpus was presented, where children are spontaneously communicating with the AIBO robot. In Narayanan and Potamianos (2002), a database was collected in a Wizard-of-Oz scenario, where children play a computer game and verbally interact with animated characters. The corpus (Xiao et al., 2002) contains speech recorded during child–machine interaction via a multimodal voice and pen interface. ChIMP is a database of child–machine spoken dialog interaction also in a game setting (Narayanan and Potamianos, 2002). The CHILDES corpus (MacWhinney, 2014) is comprised of child–human conversational speech.

For young children (up to 6 yrs) there are only a few data sets which are not easily shared. The speech of preschoolers appears in subsets of CU kids' (Hagen et al., 2003) and PF-STAR (Batliner et al., 2005) databases (4–6 years). These recordings contain isolated words and read speech. One of the largest databases in existence for child speech is from LENA Natural Language (Gilkerson and Richards, 2008), comprised very young children (1 to 4 years) spontaneous naturalistic speech. It is based on child–adult speech interaction in naturalistic home environments. The CHILDES corpus (MacWhinney, 2014) also contains naturalistic speech of young children ranging in age from 1 to 6 years. It should also be noted here that most child speech corpora, unlike adult data (from LDC), are generally not made freely available for the research community.

Some automatic speech recognition studies have been performed for young children (3–6 yrs), however most ASR systems have investigated older children. Isolated word and phrase recognition for 3 to 6 yr children with speech disorders was analyzed in Smith et al. (2017), Kothalkar et al. (2018a) and Hagen et al. (2003). The LENA naturalistic speech database has been mostly used for word count estimation based on phoneme change sequence, but not directly considering ASR (Gilkerson et al., 2017; Dykstra et al., 2013; Sangwan et al., 2015; Soderstrom and Wittebolle, 2013; Ota and Austin, 2013). When the 6–18 yr age group is explored, besides isolated word and sentence recognition, studies have also included continuous child speech (Fainberg et al., 2016; Shahnawazuddin et al., 2017; Potamianos and Narayanan, 2003; Wilpon and Jacobsen, 1996; Hassanali et al., 2015; Tong et al., 2017; Serizel and Giuliani, 2014; Matassoni et al., 2018; Sirithunge et al., 2018; Hagen et al., 2007; Shivakumar et al., 2014; Sheng et al., 2019; Yeung and Alwan, 2019; Gale et al., 2019; Li and Qian, 2019; Nagano et al., 2019).

Most previous studies have explored older children's speech recorded in relatively quiet environments, where spontaneous speech

Fig. 2. LENA recording device, which children wear in the front pocket of a specially designed vest.

is based on children-machine dialogues. Our study here focuses on very young children's speech (3 to 5 years), where spontaneous recordings are obtained from naturalistic conversations between child–adult and child–child during daily activities in noisy daycare spaces.

4. Data

All experiments reported here use American English child spontaneous conversations captured in a high quality childcare learning center. All participating children were enrolled in a center-based program in the United States.

Data was collected using LENA recorders from 33 children from two pre-kindergarden classrooms of age 3 to 5 years, and from 4 adults/teachers (3 females and 1 male). Eight of the children are at-risk (e.g., speech or language delayed) and were receiving speech-language services. A total of 80 h of speech and non-speech child and adult data was transcribed by the UTDallas CRSS transcription team (e.g., speech/language science experts). The child training corpus contains 15 h of manually transcribed audio, with transcripts containing a total of 120 K word tokens. Adult data consists of 23 h of manually transcribed audio, with 300 K words in the transcripts. The remaining 42 h (out of 80 h) are silence segments. In addition to data gathered from the childcare learning center, an out-of-domain conversational-like Web text corpus (Rousseau et al., 2014) was also used, consisting of 2.6 million word tokens. All child based speech results are reported using a held-out 3 h test set of child speech. For development, a 1.5 h data set was used. No speaker appeared simultaneously in either training, development, and test sets.

The LENA system consists of an audio recording device and speech analysis software (Xu et al., 2009). In this study, only LENA digital recorder was used (i.e., no LENA Foundation analysis software was used to analyze the audio). The LENA recorder was used with a specially designed vest/shirt to minimize sound friction and optimize microphone placement (see Fig. 2). The LENA vest/shirt plus LENA recorder was worn by child participants during classroom routines such as breakfast, classroom learning activity areas, and circle time. Recorders are lightweight and compact, cause minimal self-awareness for speakers, and allow voice capture during naturalistic conversations. The corpora include speech initiated by the speaker wearing the recording unit and speech originated by other speakers within his/her close proximity (around 6 ft). Children typically would wear the vest/shirt for a half or complete school day, and teachers were instructed to continue their regular classroom routines. No other specialized directions were provided to teachers or children. As illustrated in Fig. 3, recordings (e.g., art, blocks) would contain speech in various environments such as dramatic play, block center, manipulatives, science, art, books, music, dining space, indoor and outdoor playground. All learning spaces are open and can be noisy, providing free movement for children between areas, resulting in distractions such as crowd/babble noise, and competing speech content/sections.

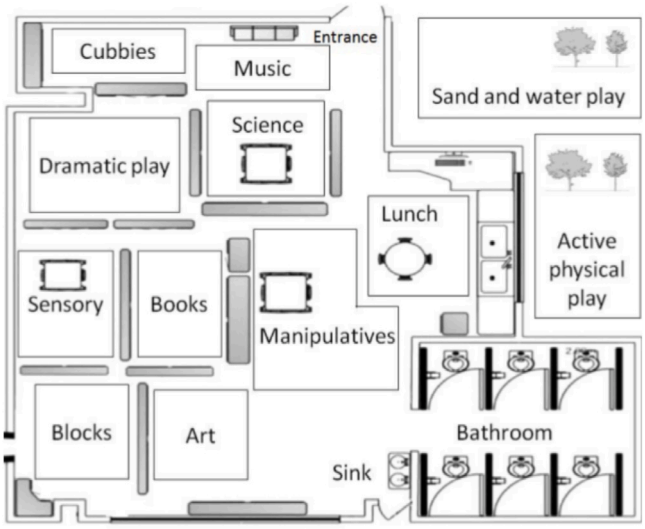

Fig. 3. A typical high quality childcare learning center.

5. Baseline recognition system

In our experiments, an ASR system is constructed using an initial 15 h of transcribed conversational child speech within an age range of 3–5 years, as described in Section 4. Acoustic models are tied-state, left-to-right 3-state HMMs with Gaussian mixture observation densities. Also, triphone-based models are word position-dependent. The acoustic models are trained on 39-dimensional Mel-Frequency Cepstral Coefficients (MFCC). The features are 9 frame spliced and projected into 40 dimensions using linear discriminant analysis (LDA) along with maximum likelihood linear transform (MLLT) application. Next, speaker adaptive training (SAT) is performed using a single feature-space maximum likelihood linear regression (fMLLR).

A 3-gram back-off language model is built using manual transcriptions from the child corpus (more details in Section 4). The lexicon used is from Rousseau et al. (2014), which consists of the most frequent 150 K words found in the web based speech/text content. To build the system, we use the Kaldi speech recognition toolkit (Povey et al., 2011). ASR performance is measured with WER.

This baseline model was used to decode the core open test data set, resulting in a child based WER of 69.4% (Table 3). A relatively high WER is naturally expected, given the spontaneous young-child multi-speaker conversational language environment.

6. DNN and LSTM system

In addition to the baseline HMM system, an advanced network solution is also considered. A deep-neural network (DNN) system is trained to estimate the HMM state likelihoods (Dahl et al., 2012). The DNN uses the same features as our SAT GMM-HMM system described in Section 5: features are spliced using a context of 9 frames, followed by LDA+MLLT+fMLLR. Alignments are produced by the SAT GMM-HMM system. In experiments with the original child training data set, we use the following DNN topology: 2 hidden layers, 2048 neurons per layer, and an output layer based on the softmax function. Sequence-discriminative training is applied with an sMBR objective function (Veselỳ et al., 2013). The learning rate is set to 1e–5, and the number of epochs is 5. In this study, we perform acoustic model augmentation experiments using DNN systems, with training data ranging from 15 to 158 h. For all experiments, the same DNN topology is used, but a different number of hidden layers is employed. When the audio data set is augmented from 38 to 75 h, 4 hidden layers are used. Furthermore,

Table 3
Results for GMM-HMM contrastive language model training conditions: manual children transcriptions (trs), manual adult transcriptions (adult), Web-texts (web), RNN generated texts based on child transcripts (rnn). In all experiments, acoustic models are based on children transcribed audio.

Language model	N-gram	#Tokens	Ppx	% WER
trs	1-gram	120K	358	84.3
trs	2-gram	120K	90	71.7
trs (baseline)	3-gram	120K	75	69.4
trs + adult	3-gram	420K	66	69.6
trs + web	3-gram	2.6M	62	69.4
trs + rnn	3-gram	30M	74	69.5
trs + adult + web + rnn	3-gram	33M	63	69.6

when increasing the quantity of training data, we expand the DNN structure to 6 hidden layers.

The long short-term memory (LSTM) system (Hochreiter and Schmidhuber, 1997) is trained using concatenated 40 dimensional MFCC features and 100 dimensional i-vectors to perform speaker adaptation (Saon et al., 2013). The i-vector extraction is based on a GMM Universal Background Model (UBM), trained using LDA+MLLT transformed MFCCs which are spliced across ±4 frames of context. The GMM-UBM consists of 512 mixture components, and the LSTM architecture has 3 hidden layers, followed by a softmax output layer. Each hidden layer contain 1024 neurons, with both recurrent and non-recurrent projection dimensions set to 256. The network is optimized using a stochastic gradient descent with an initial learning rate of 3e–4 and a momentum of 0.9. A cross entropy criterion is applied for training, and the training process is repeated for 6 epochs. Only a data set of 158 h is used for LSTM evaluation.

7. Data augmentation

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

7.1. Language model augmentation

To improve the language model, three alternate data augmentation techniques are investigated: (i) adding adult data, (ii) incorporating web text content, and (iii) producing additional text sequences via RNNs (Mikolov and Zweig, 2012). The language model is estimated using supplemental text resources and interpolated with the original baseline language model. The expectation maximization (EM) algorithm is used for interpolation to minimize overall perplexity of the development set.

(i) Adult data usage. The use of manually annotated adult transcriptions is investigated for data augmentation. All conversational based adult data was recorded in the same childcare center, as described in Section 4.

(ii) Web data usage. Recovery of conversational-like Web text data similar to that in the child learning space was explored to improve the language model (see Section 4).

(iii) RNN based text generation. We also investigate additional text generation using an RNN as proposed in Mikolov and Zweig (2012). The RNN consists of 2 hidden layers and 512 units per layer. We randomly shuffled the training transcripts and split this into five non-overlapping subsets. For each split, the RNN was trained using four sets and the fifth set used for validation. The RNN finds long contextual regularities, produces quite meaningful sentences, and maintains a consistent vocabulary content to the original vocabulary lexicon.

To assess the improvement derived from the use of supplemental text resources, contrastive experiments are performed with alternate language models. The use of 1-gram and 2-gram models leads to ASR performance degradation, as shown in Table 3. Augmentation techniques are explored with 3-gram language models. From Table 3, it is observed that each LM augmentation technique helps improve word perplexity compared to the baseline. The highest perplexity reduction of 13 points (75 vs. 62) is achieved with the language model incorporating Web texts, resulting in texts with 2.6M word tokens. A very similar word perplexity of 63 is obtained using adult training transcripts, Web texts, and RNN generated texts, resulting in texts with 33M word tokens (bottom entry). However, while word perplexity is reduced, there is no corresponding meaningful WER improvement achieved over the baseline using alternate LM augmentation techniques. The outcome here suggests that the children are producing ill-formed sentence structures at some level, so augmenting with either adult text data, Web text data, or RNN text generation is not producing the type of sentence structures that typically developing or children at risk are employing. So using more well formed text does not help in LM advancements. Alternate strategies to create more child scenario sentence structures are needed.

7.2. Acoustic model augmentation

Acoustic data augmentation is assessed via three alternate approaches: speed and tempo perturbation combinations as described in Ko et al. (2015), where an adult data set is used. We investigate the impact of different perturbation coefficients and alternate number of copies of the original child data set (15 h).

Speed perturbation. Speed perturbation emulates both pitch and tempo variations in the speech signal. Speed modification is achieved by resampling the signal. We use the *speed* command of the *sox*[2] tool (e.g., time-domain pitch synchronous overlap and add (TD-PSOLA)) to modify the speed of the signal. We explore augmentation of the training data set by changing the speed (e.g., overall duration) of the audio signal, resulting in four versions of the original child training data with speed scaling factors of 0.8, 0.9, 1.1, and 1.2 (10% and 20% increase and decrease in the original rate of speed).

Tempo perturbation. The tempo of the signal is modified, while the pitch and spectral envelope of the signal is not changed. To perform tempo perturbation, we used the *sox* with *tempo* command. It uses an overlap-add technique based on waveform similarity (WSOLA) implementation (Upperman, 2012). The training data set was enlarged by creating four additional copies of the original child training data by modifying the tempo with scaling factors to 0.8, 0.9, 1.1, and 1.2 (again, a 10% and 20% increase or decrease in tempo).

Adult data usage. We joined both child and adult training data audio sets. The adult data set is comprised of 23 h of transcribed audio, where most speakers are females. All data was recorded in the same childcare center (see Section 4).

Acoustic model augmentation results are provided in Table 4. In the experiments, we use a language model from manually annotated child transcriptions.

Table 4 shows that for GMM-HMM system there is no WER improvement obtained by incorporating various acoustic data sets. The lowest 69.4% WER is achieved using only the original child training or also incorporating all augmented acoustic sets.

The performance of the DNN-HMM systems are also summarized in Table 4. The top line shows that with the original children transcribed audio set, there is an improvement of +4.3% absolute using DNN-HMM training over GMM-HMM. Comparing DNN performance with different acoustic model sets, it can be observed that an absolute WER reduction of 2.6% is achieved using the 45 h data set which incorporates augmented speech audio data based on speed perturbed with 0.9, 1.1 factors (65.1% vs. 62.5%). Other perturbation combinations are also shown to be beneficial compared to the original child training audio

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

Table 4
Results for GMM-HMM, DNN-HMM, LSTM contrastive acoustic model training conditions: manually transcribed children audio (trs), copies of children training set with different speed perturbation factors (speed), with different tempo perturbation factors (tempo), adult transcribed audio (adult). Different amount of hours for training is used (#Hrs).

Acoustic model	Perturb. factors	#Hrs	% WER		
			GMM	DNN	LSTM
trs	–	15	69.4	65.1	–
trs + adult	–	38	74.1	65.2	–
trs + speed	0.9, 1.1	45	69.5	62.5	–
trs + tempo	0.9, 1.1	45	70.3	64.1	–
trs + speed	0.8, 0.9, 1.1, 1.2	75	70.8	63.2	–
trs + tempo	0.8, 0.9, 1.1, 1.2	75	70.0	64.7	–
trs + speed + tempo	0.8, 0.9, 1.1, 1.2	135	70.3	64.1	–
trs + speed + tempo + adult	0.8, 0.9, 1.1, 1.2	158	69.4	62.2	57.8

Table 5
Results for the best GMM-HMM, DNN-HMM, LSTM in terms of substitution, deletion, and insertion.

System	% WER	Sub	Del	Ins
GMM	69.4	39.4	25.4	4.6
DNN	62.2	35.4	22.3	4.6
LSTM	57.8	35.3	16.1	6.4

Table 6
Results for the best GMM-HMM, DNN-HMM, LSTM for each child. In all experiments, language models are based on children transcriptions, and all augmented audio data sets are incorporated for acoustic models.

#ID child	% WER		
	GMM	DNN	LSTM
1 (at-risk)	74.6	69.0	63.3
2 (at-risk)	70.7	64.1	62.9
3 (at-risk)	82.9	76.4	72.6
4 (at-risk)	67.4	60.5	55.5
5 (at-risk)	66.9	60.5	54.8
6	79.6	72.8	68.8
7	71.2	61.0	57.6
8	62.8	56.6	47.3
9	50.9	45.6	39.9
10	85.1	76.6	74.6
11	60.2	52.8	48.1
12	77.4	69.6	67.5

set, but not better than using two copies of the speed perturbation. Also, there is only a tiny improvement obtained using all incorporated data in comparison with two copies of the speed perturbation. We investigate the expanded 158 h data set that includes additional transcribed adult data. In this case, however, an improvement of +7.2% is achieved over baseline (69.4% vs. 62.2%).

Finally, we explore LSTM system with the expanded 158 h data set. In this case, the greatest WER improvement of +11.6% is observed over the baseline (69.4% vs. 57.8% in Table 4).

Table 5 shows the results of our best GMM-HMM (69.4% WER), DNN-HMM (62.2% WER), and LSTM (57.8% WER) systems in terms of substitution, deletion, and insertion. For all systems, the majority of errors are in fact substitution errors versus insertions and deletions.

The results of our best GMM-HMM, DNN-HMM, and LSTM systems (69.4%, 62.2%, and 57.8% WER) for each child are summarized in Table 6 to explore per child performance variability. Considering first the LSTM system, it is seen that WER of children at risk ranges from 54.8% to 72.6%; meanwhile, WER of typically developing children range from 39.9% to 74.6%. The highest recognition results are obtained for typically developing children child#9, child#11, child#8, followed by child at risk #5 and child#4. The WERs of GMM-HMM, DNN-HMM system are higher, but the best recognition results are arranged in the same order as LSTM. In general, average WERs for typically developing children were lower (69.6%, 62.1%, 57.7%) versus at risk children (72.5%, 66.1%, 61.8%) for (GMM, DNN, LSTM) systems.

Table 7
Word counts of each child in test set: the number of words in references (#N ref), in hypothesis (#N hyp) obtained using LSTM system.

#ID child	#N ref	#N hyp
1 (at-risk)	562	479
2 (at-risk)	699	655
3 (at-risk)	726	605
4 (at-risk)	1865	1700
5 (at-risk)	2780	2446
6	1197	1096
7	1515	1431
8	2310	2030
9	3077	2840
10	3634	3445
11	3715	3322
12	3794	3333

At this point, it is clear that benefits can be achieved using various data augmentation concepts for child speech in naturalistic settings. While WERs do improve, their values are fundamentally different from adults. This is expected, since the fundamental structure of child language/grammar as well as word pronunciation may not follow a traditional assumed set of rules experienced for adult speech. So, at this point, given the available knowledge estimated and extracted from the child portion of child–adult/child–child naturalistic communication, we now turn to a systematic investigation of child word and language structure in order to explore ways to understand language engagement.

8. Child language engagement

A rich, supportive language environment within the early childhood classroom is essential for all children. It is important to recognize that both environment and the role of adult-to-child and peer-to-child communication, serve as key components for young children's language development — particularly children at-risk or with disabilities (Brown et al., 1999; Mahoney and Wheeden, 1999; Sontag, 1997; Warren and Yoder, 2004; Perry et al., 2018). There is a need to assess language/communication engagement of children at-risk for or with disabilities to determine if these children should receive greater teacher and peer support during learning activities.

In this section, we assess children's speech development using word count rates. Word counts are estimated based on hypothesis output transcription from our best ASR system (LSTM with 57.8% WER, Table 6). It should be noted that in this context, even somewhat greater WERs are acceptable as long as most errors are substitution errors. Our test set consists of children's speech as described in Section 4. For the purpose of word count, while WER is 57.8%, it should be noted that the majority of errors are substitution errors vs. insertions and deletions (see Table 5). The results of word count estimation for children's speech are provided in Table 7. From word count references (column #N ref), it can be observed that child#1, child#2, and child#3 are at-risk and have the lowest vocal interaction level. Child#4 and child#5 are at-risk, but have relatively high vocal communication interaction. Comparing word counts from references (Table 7, column #N ref) with counts in the hypothesis (column #N hyp), it can be seen that even with ASR system errors, it is still possible to establish which children have low conversational interaction and are at-risk (e.g., child#1, child#2, and child#3). The system reorders only child#3 based on word counts in this hypothesis. We note that after reordering, child#3 is still included as the at-risk category with the lowest speech engagement. Due to challenges in this naturalistic child–child and adult–child learning space, word counts are not always accurate, however they are consistent, and we are still able to quantify which children have low conversational interaction.

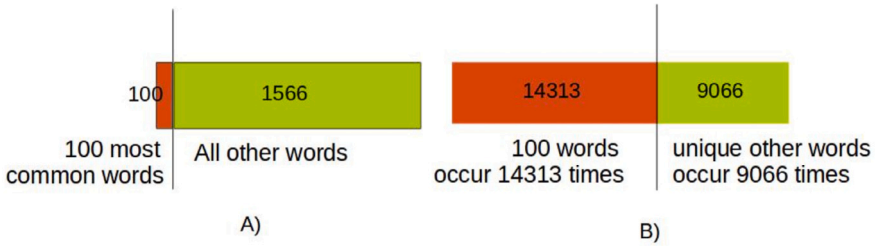

Fig. 4. Child word count in ASR system hypothesis: (A) 100 most frequently occurring words, and word count of unique other words. (B) Total word count of 100 most frequently occurring words, and word count of all other words.

9. Statistical patterns of naturalistic child speech for American English

Child speech from naturalistic real world conditions rarely conform grammatically or in word choice to an orderly structure of formal adult speech. In spontaneous child speech, speaking rates and styles vary across children, word selection may be non-traditional, and grammar rules are typically not strictly followed.

Typically adults are able to decode child speech more successfully than ASR systems. ASR systems focus on identification of individual elements, such as phones, words or sentences. Humans have the capability to focus on pivot words around which the surrounding lexical items assume their shape. ASR systems assume that all recognition errors are equally important in the decoding process, however not all errors are equal. For example, it is not as important to recognize some common occurring *stop* words versus less frequently occurring words which have multiple syllables, or more information bearing content. Stop words are typically associated with single/short syllable words, which occur often with limited information context: a, the, it, and, or, of, etc. A list of common words can be used in conjunction with other knowledge sources to interpret the speech stream. It is important to effectively recognize multi-syllable words as this class provides more meaningful information regarding child sentence/language perplexity. In contrast, stop words do not carry meaningful information.

In this section we analyze statistical patterns of spontaneous child speech for American English. The analyses presented here are intended to serve as a representative sample to assess high quality language engagement between adult–child diads in learning spaces.

9.1. Word frequency and category

It is known that words differ greatly in terms of their frequency of occurrence in language. Common words occur more frequently than the least common, and may reflect different levels of information content. Recognizing information rich words are more important than low information words.

The most frequent 100 words of our best ASR system hypothesis (containing only child speech as described in Section 4) are summarized in Table 8. The most common words occur by at least several orders of magnitude more frequently than the least common. From column #N indicating word count, it can be observed that of the 100 most frequently occurring words, the most frequent word "I" occurs 20 times more frequently than the least occurring word "too".

Total word counts of ASR system hypothesis are shown in Fig. 4. It can be observed that the 100 most frequent words make up 63% of the spoken words: the 100 most frequent words occurred 14 313 times, while the remaining 1566 words occurred only 9066 times.

A summary of the most frequent 100 words in the references which contain children conversational speech can be found in Appendix. We note even with ASR system errors, there is a 91% correct ASR word match for this top 100 word list.

Although a list of common words does not provide sufficient data to interpret speech, it can be used in conjunction with other knowledge sources. We characterize these most common words in terms of categories and parts of speech in order to analyze word and language structure diversity and perplexity (column category in Table 8): *quantity, animal, question, social, time, stop words, other*. Fig. 5 shows that most words (55%) come from the *stop words* category as expected, however this category comprises low information bearing content (e.g. a, and, the). The category of *quantity* accounts for 10% of the individual tokens. The other categories contain varying low percentages. The parts of speech are also summarized in Fig. 5. The most frequent parts of speech are verbs (27%), adverbs (14%), and pronouns (14%), that typically do not bear as much meaningful information. In contrast, nouns and adjectives bring more meaningful information, but occur only 13% and 8%, respectively.

9.2. Syllable structure

In this section, we characterize the most frequently occurring words in terms of syllable structure. Single-syllable words may carry some or only limited information (e.g., *stop words* such as of, the, and, or, it, etc.). However, other single-syllable words carry more meaningful content (e.g., cat, bat, jump, run, help, stop, stand, walk, etc.).

In contrast, multi-syllable words consistently bear meaningful information content (e.g., computer, hippopotamus, running, histogram, Washington, etc.). Table 8 (column syllable #N) lists the 100 most frequent lexical items of naturalistic child speech of which only 7% are two-syllable (going, kitty, kitties, meow, mister, because, little), while others are single-syllable. Both stop words and single-syllable words carry very limited meaningful information, and do not reflect high quality language engagement between adult–child in learning spaces. Meanwhile, for adults, it is multi-syllable words that carry important information and dominate conversation (as reported on the SwitchBoard corpus by Greenberg (1997)). This statistical skew towards short syllabic forms provides another interpretative constraint in decoding speech streams. Knowing the number of syllables in a word also provides some degree of grammatical information. This is a consequence of the tendency for multi-syllabic words to be either a noun, verb, or adjective. In general, verbs are rarely longer than two syllables in length (Greenberg, 1997).

It is difficult to accomplish accurate automatic syllabification in real time. Here, we analyze the heterogeneity of syllabic forms of naturalistic child speech in an offline manner. Table 8, column *syllable structure*, provides the composition of syllabic forms from consonants [C] and vowels [V], where the syllable type is based on the most frequent variation. In the study of Switchboard corpus (Greenberg, 1997) it was illustrated that syllabic composition of spontaneous adult English might assume a wide range of patterns (e.g., "strengths" is represented as CCCVCCC). However, syllable structure of English adult speech is more homogeneous, with the four most common syllable structures consisting of: consonant + vowel [CV], consonant + vowel + consonant [CVC], vowel + consonant [VC] or vowel [V].

The syllable structure of conversational child speech is also homogeneous, as shown in Fig. 6. Here, 70% of the adult–child corpus syllables are made up of CVC (36%), CV (18%), and VC (16%) classes. Syllables with complex patterns do not occur frequently (e.g., words

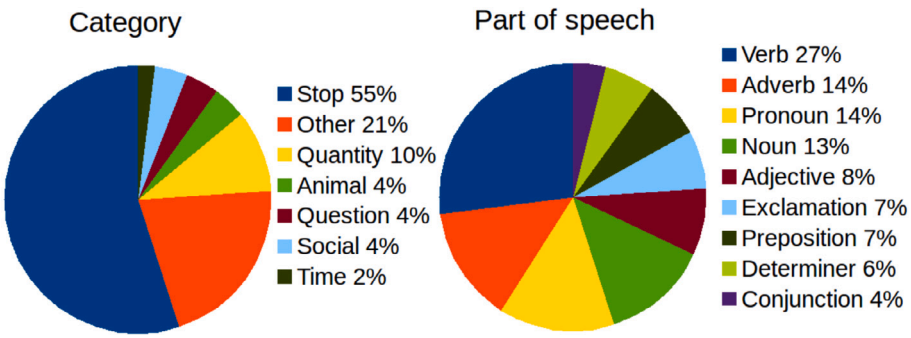

Fig. 5. Frequency of categories and parts of speech for the most common 100 child words in ASR system hypothesis.

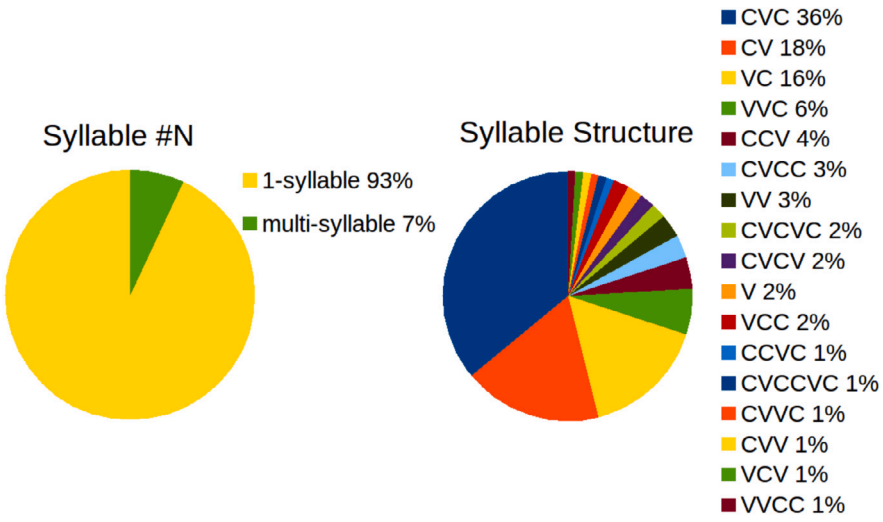

Fig. 6. Syllable counts and frequency of occurrence for various forms of syllable structure (consonant C and vowel V) for the most common 100 child words in ASR system hypothesis.

CVCVC structure occur 2% of the time: "kitties", "because"). The results from Table 8, Figs. 5 and 6 illustrate that stop words make up 55% of conversational words. As such, raw word counts while useful, do not convey the full information bearing quality of child speech in adult–child conversational interaction. Identifying question words (4%), animal (4%), nouns (13%), and verbs (27%) provide greater insight as to the quality of child conversational interaction.

10. Conclusions

This study has investigated the use of speech technology to assess word and word structure in child language engagement (American English) in naturalistic learning spaces. We investigated the benefits of applying data augmentation techniques for young child (age from 3 to 5 years) communications in assessing child naturalistic engagement in adult–child learning spaces through speech recognition. We explored several data augmentation techniques to advance language and acoustic models to improve child speech recognition, and showed which technology improvements provided gains in ASR performance. We also explored assessment of child language development and engagement via word count rates. The results showed that even low performing ASR systems can contribute to effective conversation engagement assessment. Finally, we analyzed the statistical patterns of naturalistic child speech. It was demonstrated that the most common word type in conversational adult–child speech are stop words which carry little meaningful information, followed by single syllable words which have less information bearing content. Multi-syllable words occur only 7% of the time, but carry significantly more value for engagement assessment.

In addition to ASR advancement, alternate text augmentation approaches were investigated to increase the limited amount of original transcribed conversational child speech using: (i) adult text data, (ii) Web text data, and (iii) texts generated by RNN. Interpolating these texts collectively leads to a perplexity improvement of 13 points, but unfortunately did not result in a corresponding improvement in ASR WER over the original baseline.

Next, acoustic data augmentation techniques for child speech were explored based on: (i) speed perturbation, (ii) tempo perturbation, and adding (iii) adult data. The experiments were performed with training data varying from 15 to 158 h. Both speed and tempo perturbation were shown to improve WER, with speed perturbation factors of 0.9, 1.1 to be the most beneficial. The greatest WER reduction of 11.6% absolute was achieved over the baseline after incorporating all augmented audio data sets, and using our LSTM system.

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

Finally, we analyzed statistical patterns of conversational preschoolers speech, since it was hypothesized that not all recognition errors are equally important. In naturalistic child speech, the most common

Table 8
Statistical properties for the most common 100 words from CRSS-UTDallas Adult–Child corpus, results from ASR system hypothesis containing child speech: the word count (#N), syllable type composed of consonant (C) and vowel (V), syllable count (syllable #N), category, and part of speech.

	Word	#N	Syllable struc.	Syllable #N	Category	Part of speech
1	I	830	V	1-syllable	stop	pronoun
2	to	714	CV	1-syllable	stop	preposition
3	you	684	VV	1-syllable	stop	pronoun
4	the	643	CV	1-syllable	stop	determiner
5	a	420	V	1-syllable	stop	determiner
6	and	387	VCC	1-syllable	stop	conjunction
7	it	320	VC	1-syllable	stop	pronoun
8	have	292	CVC	1-syllable	stop	verb
9	we	291	CV	1-syllable	stop	pronoun
10	this	286	CVC	1-syllable	stop	pronoun
11	that	270	CVC	1-syllable	stop	pronoun
12	on	266	VC	1-syllable	stop	preposition
13	get	256	CVC	1-syllable	other	verb
14	my	250	CV	1-syllable	stop	pronoun
15	me	243	CV	1-syllable	stop	pronoun
16	is	233	VC	1-syllable	stop	verb
17	can	232	CVC	1-syllable	stop	verb
18	one	226	VVC	1-syllable	quantity	adjective
19	going	214	CVVC	2-syllable	other	verb
20	do	203	CV	1-syllable	stop	verb
21	no	197	CV	1-syllable	stop	exclamation
22	in	194	VC	1-syllable	stop	preposition
23	want	191	VVCC	1-syllable	social	verb
24	go	180	CV	1-syllable	other	verb
25	what	179	CVC	1-syllable	question	pronoun
26	not	149	VCV	1-syllable	stop	adverb
27	meow	149	CVV	2-syllable	animal	noun
28	got	137	CVC	1-syllable	other	verb
29	for	137	CVC	1-syllable	stop	preposition
30	your	133	VVC	1-syllable	stop	pronoun
31	be	130	CV	1-syllable	stop	verb
32	look	124	CVC	1-syllable	other	verb
33	mess	123	CVC	1-syllable	other	noun
34	but	119	CVC	1-syllable	stop	conjunction
35	are	119	VC	1-syllable	stop	verb
36	like	118	CVC	1-syllable	social	verb
37	he	117	CV	1-syllable	stop	pronoun
38	oh	114	VC	1-syllable	stop	exclamation
39	don't	114	CVCC	1-syllable	stop	verb
40	here	111	CVC	1-syllable	stop	adverb
41	just	110	CVCC	1-syllable	stop	adjective
42	hey	110	CV	1-syllable	stop	exclamation
43	was	100	VVC	1-syllable	stop	verb
44	up	98	VC	1-syllable	stop	adverb
45	need	98	CVC	1-syllable	social	verb
46	kitty	98	CVCV	2-syllable	animal	noun
47	out	95	VC	1-syllable	stop	adverb
48	all	95	VC	1-syllable	quantity	determiner
49	well	94	VVC	1-syllable	stop	adverb
50	its	94	VCC	1-syllable	stop	determiner
51	uh	92	VC	1-syllable	stop	exclamation
52	put	89	CVC	1-syllable	other	verb
53	of	89	VC	1-syllable	stop	preposition
54	see	87	CV	1-syllable	other	verb
55	there	86	CVC	1-syllable	stop	adverb
56	guys	85	CVC	1-syllable	other	noun
57	yeah	84	VV	1-syllable	stop	exclamation
58	make	84	CVC	1-syllable	other	verb
59	when	83	CVC	1-syllable	question	adverb
60	two	83	CV	1-syllable	quantity	noun
61	some	83	CVC	1-syllable	quantity	determiner
62	three	79	CCV	1-syllable	quantity	noun
63	so	79	CV	1-syllable	stop	adverb
64	ah	79	VC	1-syllable	stop	exclamation
65	play	78	CCV	1-syllable	other	verb
66	at	73	VC	1-syllable	stop	preposition
67	more	72	CVC	1-syllable	quantity	determiner
68	then	71	CVC	1-syllable	stop	adverb
69	how	71	CV	1-syllable	question	adverb
70	these	68	CVC	1-syllable	stop	pronoun
71	am	68	VC	1-syllable	stop	verb
72	know	67	CVC	1-syllable	other	verb
73	with	66	CVC	1-syllable	stop	preposition
74	right	66	CVC	1-syllable	other	adjective
75	now	63	CV	1-syllable	time	adverb
76	kitties	63	CVCVC	2-syllable	animal	noun
77	time	59	CVC	1-syllable	time	noun
78	them	59	CVC	1-syllable	stop	pronoun
79	because	58	CVCVC	2-syllable	stop	conjunction
80	why	57	CCV	1-syllable	question	adverb
81	can't	57	CVCC	1-syllable	stop	verb
82	off	55	VC	1-syllable	stop	adverb
83	mister	55	CVCCVC	2-syllable	other	noun
84	dog	54	CVC	1-syllable	animal	noun
85	woof	53	VVC	1-syllable	stop	noun
86	little	53	CVCV	2-syllable	quantity	adjective
87	blue	52	CCV	1-syllable	other	adjective
88	big	51	CVC	1-syllable	quantity	adjective
89	whoa	50	CV	1-syllable	stop	exclamation
90	watch	50	VVC	1-syllable	other	verb
91	back	50	CVC	1-syllable	other	verb
92	come	49	CVC	1-syllable	other	verb
93	if	49	VC	1-syllable	other	conjunction
94	said	47	CVC	1-syllable	other	verb
95	good	45	CVC	1-syllable	social	adjective
96	green	44	CCVC	1-syllable	other	adjective
97	our	43	VV	1-syllable	stop	pronoun
98	four	43	CVC	1-syllable	quantity	noun
99	five	43	CVC	1-syllable	quantity	noun
100	too	42	CV	1-syllable	stop	adverb

(*continued on next page*)

used words are stop words and single-syllable words that carry limited meaningful information. In contrast, for adults the multi-syllable words which carry more content based structure generally dominate the conversation. Finally, a systematic study of the top 100 most frequently occurring words, as well as their syllable structure revealed much insight into the distribution of word type, as well as the word frequency typical in adult–child conversational analysis in preschool learning spaces. This in theory, would offer opportunities to strengthen the quality of teacher adult–child and peer-child interactions for *all* children in preschool learning spaces.

CRediT authorship contribution statement

Rasa Lileikyte: Conceptualization, Methodology, Software, Validation, Investigation, Writing. **Dwight Irvin:** Corpus collection, Protocol for adult-child learning space data collection, Investigation. **John H.L. Hansen:** Supervision, Conceptualization, Resources, Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This work was supported by a project from NSF Grant #1918032 (UTDallas CRSS), which is noted in Sec. 11 Acknowledgments section. All authors have no financial interests which would suggest a Conflict of Interest for the work contained in this study.

Acknowledgment

This study was supported in part by National Science Foundation (NSF), United States Grant #1918032 (UTDallas CRSS).

Appendix. The most common 100 child words in ASR system references (#N is the word count)

	Word	#N		Word	#N		Word	#N
1	I	908	35	your	125	69	uh	68
2	the	725	36	here	125	70	three	68
3	to	712	37	its	121	71	there	68
4	you	694	38	be	121	72	make	67
5	a	490	39	look	120	73	woof	66
6	and	481	40	but	119	74	well	66
7	it	392	41	just	118	75	kitties	66
8	can	310	42	ho	117	76	down	66
9	me	307	43	don't	117	77	these	65
10	this	295	44	when	116	78	mister	65
11	have	282	45	watch	108	79	will	64
12	do	271	46	hey	100	80	right	64
13	we	266	47	kitty	97	81	because	64
14	is	266	48	play	94	82	them	62
15	in	263	49	oh	93	83	off	62
16	my	248	50	with	91	84	too	59
17	meow	232	51	need	88	85	put	59
18	on	231	52	for	88	86	four	56
19	going	221	53	two	87	87	they	55
20	get	216	54	was	85	88	then	55
21	one	215	55	some	85	89	guys	54
22	what	196	56	now	85	90	did	53
23	that	185	57	all	83	91	dog	52
24	want	182	58	more	82	92	why	50
25	no	174	59	so	81	93	little	48
26	go	173	60	at	80	94	blue	48
27	he	154	61	how	78	95	our	47
28	are	144	62	know	77	96	green	46
29	not	143	63	out	76	97	eat	45
30	like	142	64	ah	75	98	come	45
31	up	141	65	see	73	99	back	45
32	of	140	66	can't	70	100	time	44
33	got	138	67	if	69			
34	mess	128	68	yeah	68			

References

Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., Wong, M., 2005. The PF_STAR children's speech corpus. In: Ninth European Conference on Speech Communication and Technology.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M.J., Wong, M., 2004. "You Stupid Tin Box"-children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In: Lrec.

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: Interspeech 2005-Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005. ISCA, pp. 2765–2768.

Brown, W.H., Odom, S.L., Li, S., Zercher, C., 1999. Ecobehavioral assessment in early childhood programs: A portrait of preschool inclusion. J. Spec. Educ. 33 (3), 138–153.

Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., Barbarin, O., 2008. Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher–child interactions and instruction. Appl. Dev. Sci. 12 (3), 140–153.

Dahl, G.E., Yu, D., Deng, L., Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20 (1), 30–42.

Dykstra, J.R., Boyd, B.A., Watson, L.R., Crais, E.R., Baranek, G.T., 2012. The impact of the advancing social-communication and play (ASAP) intervention on preschoolers with autism spectrum disorder. Autism 16 (1), 27–44.

Dykstra, J.R., Sabatos-DeVito, M.G., Irvin, D.W., Boyd, B.A., Hume, K.A., Odom, S.L., 2013. Using the language environment analysis (LENA) system in preschool classrooms with children with autism spectrum disorders. Autism 17 (5), 582–594.

Eskenazi, M., Mostow, J., Graff, D., 1997. The CMU kids corpus LDC97S63. Linguist. Data Consortium Database.

Fainberg, J., Bell, P., Lincoln, M., Renals, S., 2016. Improving children's speech recognition through out-of-domain data augmentation.. In: Interspeech. pp. 1598–1602.

Gale, R., Chen, L., Dolata, J., Van Santen, J., Asgari, M., 2019. Improving asr systems for children with autism and language impairment using domain-focused dnn transfer techniques. In: Interspeech. 2019, NIH Public Access, p. 11.

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. ACM, p. 7.

Gilkerson, J., Richards, J.A., 2008. The LENA Natural Language Study. LENA Foundation, Boulder, CO, Retrieved March 3, 2009.

Gilkerson, J., Richards, J.A., Warren, S.F., Montgomery, J.K., Greenwood, C.R., Oller, D.K., Hansen, J.H.L., Paul, T.D., 2017. Mapping the early language environment using all-day recordings and automated analysis. Am. J. Speech-Lang. Pathol. 26 (2), 248–265.

Greenberg, S., 1997. On the origins of speech intelligibility in the real world. In: Robust Speech Recognition for Unknown Communication Channels.

Hadian, H., Sameti, H., Povey, D., Khudanpur, S., 2018. End-to-end speech recognition using lattice-free MMI. In: Interspeech. pp. 12–16.

Hagen, A., Pellom, B., Cole, R., 2003. Children's speech recognition with application to interactive books and tutors. In: IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, St. Thomas, USA. pp. 186–191.

Hagen, A., Pellom, B., Cole, R., 2007. Highly accurate children's speech recognition for interactive reading tutors using subword units. Speech Commun. 49 (12), 861–873.

Hart, B., Risley, T.R., 1995. Meaningful Differences in the Everyday Experience of Young American Children. Paul H Brookes Publishing.

Hartmann, W., Hsiao, R., Tsakalidis, S., 2017. Alternative networks for monolingual bottleneck features. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5290–5294.

Hassanali, K.-n., Yoon, S.-Y., Chen, L., 2015. Automatic scoring of non-native children's spoken language proficiency. In: SLaTE. pp. 13–18.

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

Irvin, D.W., Crutchfield, S.A., Greenwood, C.R., Simpson, R.L., Sangwan, A., Hansen, J.H.L., 2017. Exploring classroom behavioral imaging: Moving closer to effective and data-based early childhood inclusion planning. Adv. Neurodev. Disorders 1 (2), 95–104.

Justice, L.M., McGinty, A.S., Zucker, T., Cabell, S.Q., Piasta, S.B., 2013. Bi-directional dynamics underlie the complexity of talk in teacher–child play-based conversations in classrooms serving at-risk pupils. Early Child. Res. Q. 28 (3), 496–508.

Ko, T., Peddinti, V., Povey, D., Khudanpur, S., 2015. Audio augmentation for speech recognition. In: Sixteenth Annual Conference of the International Speech Communication Association.

Kothalkar, P., Rudolph, J., Dollaghan, C., McGlothlin, J., Campbell, T., Hansen, J.H.L., 2018a. Fusing text-dependent word-level i-vector models to screen'at risk'child speech. In: Interspeech. pp. 1681–1685.

Kothalkar, P.V., Rudolph, J., Dollaghan, C., McGlothlin, J., Campbell, T.F., Hansen, J.H.L., 2018b. Automatic screening to detect'at risk'child speech samples using a clinical group verification framework. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 4909–4913.

Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. J. Acoust. Soc. Amer. 105 (3), 1455–1468.

Li, C., Qian, Y., 2019. Prosody usage optimization for children speech recognition with zero resource children speech. In: Interspeech. pp. 3446–3450.

Lileikytė, R., Fraga-Silva, T., Lamel, L., Gauvain, J.-L., Laurent, A., Huang, G., 2017. Effective keyword search for low-resourced conversational speech. In: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, pp. 5785–5789.

Lileikyte, R., Irvin, D., Hansen, J.H.L., 2020. Assessing child communication engagement via speech recognition in naturalistic active learning spaces. In: Proc. Odyssey 2020 The Speaker and Language Recognition Workshop. pp. 396–401.

Lileikytė, R., Lamel, L., Gauvain, J.-L., Gorin, A., 2018. Conversational telephone speech recognition for Lithuanian. Comput. Speech Lang. 49, 71–82.

MacWhinney, B., 2014. The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database. Psychology Press.

Mahoney, G., Wheeden, C.A., 1999. The effect of teacher style on interactive engagement of preschool-aged children with special learning needs. Early Child. Res. Q. 14 (1), 51–68.

Matassoni, M., Gretter, R., Falavigna, D., Giuliani, D., 2018. Non-native children speech recognition through transfer learning. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6229–6233.

Mikolov, T., Zweig, G., 2012. Context dependent recurrent neural network language model. SLT 12, 234–239.

Nagano, T., Fukuda, T., Suzuki, M., Kurata, G., 2019. Data augmentation based on vowel stretch for improving children's speech recognition. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, pp. 502–508.

Narayanan, S., Potamianos, A., 2002. Creating conversational interfaces for children. IEEE Trans. Speech Audio Process. 10 (2), 65–78.
Ota, C.L., Austin, A.M.B., 2013. Training and mentoring: Family child care providers' use of linguistic inputs in conversations with children. Early Child. Res. Q. 28 (4), 972–983.
Perry, L., Prince, E., Valtierra, A., Rivero-Fernandez, C., Ullery, M., Katz, L., et al., 2018. A year in words: the dynamics and consequences of language experiences in an intervention classroom. PLoS One 13 (7), e0199893.
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: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. No. EPFL-CONF-192584, IEEE Signal Processing Society.
Rousseau, A., Deléglise, P., Estève, Y., 2014. Enhancing the TED-LIUM corpus with selected data for language modeling and more ted talks. In: LREC. pp. 3935–3939.
Sangwan, A., Hansen, J.H.L., Irvin, D., Crutchfield, S., Greenwood, C., 2015. Studying the relationship between physical and language environments of children: Who's speaking to whom and where? In: 2015 IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE). IEEE, pp. 49–54.
Saon, G., Soltau, H., Nahamoo, D., Picheny, M., 2013. Speaker adaptation of neural network acoustic models using i-vectors. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, pp. 55–59.
Serizel, R., Giuliani, D., 2014. Deep neural network adaptation for children's and adults' speech recognition. In: Italian Computational Linguistics Conference (CLiC-it).
Shahnawazuddin, S., Deepak, K., Pradhan, G., Sinha, R., 2017. Enhancing noise and pitch robustness of children's ASR. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5225–5229.
Sheng, P., Yang, Z., Qian, Y., 2019. Gans for children: A generative data augmentation strategy for children speech recognition. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, pp. 129–135.
Shivakumar, P.G., Potamianos, A., Lee, S., Narayanan, S., 2014. Improving speech recognition for children using acoustic adaptation and pronunciation modeling. In: WOCCI. pp. 15–19.
Sirithunge, H.C., Muthugala, M.V.J., Jayasekara, A.B.P., Chandima, D., 2018. A wizard of oz study of human interest towards robot initiated human-robot interaction. In: 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, pp. 515–521.
Smith, D., Sneddon, A., Ward, L., Duenser, A., Freyne, J., Silvera-Tawil, D., Morgan, A., 2017. Improving child speech disorder assessment by incorporating out-of-domain adult speech. In: Interspeech. pp. 2690–2694.
Soderstrom, M., Wittebolle, K., 2013. When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. PLoS One 8 (11), e80646.
Sontag, J.C., 1997. Contextual factors influencing the sociability of preschool children with disabilities in integrated and segregated classrooms. Except. Child. 63 (3), 389–404.
Tong, R., Chen, N.F., Ma, B., 2017. Multi-task learning for mispronunciation detection on singapore children's mandarin speech. In: Interspeech. pp. 2193–2197.
Upperman, G., 2012. Changing pitch with PSOLA for voice conversion. retrieved from on Nov 12, 1.
Veselỳ, K., Ghoshal, A., Burget, L., Povey, D., 2013. Sequence-discriminative training of deep neural networks. In: Interspeech. pp. 2345–2349.
Warren, S.F., Yoder, P.J., 2004. Early intervention for young children with language impairments. In: Classification of Developmental Language Disorders: Theoretical Issues and Clinical Implications. Lawrence Erlbaum Mahwah, NJ, pp. 367–381.
Wilpon, J.G., Jacobsen, C.N., 1996. A study of speech recognition for children and the elderly. In: 1996 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 1. IEEE, pp. 349–352.
Xiao, B., Girand, C., Oviatt, S., 2002. Multimodal integration patterns in children. In: Seventh International Conference on Spoken Language Processing.
Xu, D., Yapanel, U., Gray, S., 2009. Reliability of the LENATM Language Environment Analysis System in Young Children's Natural Language Home Environment. LENA Foundation, Boulder, CO, Retrieved from http://www.lenafoundation.org/TechReport. aspx Find this author on.
Yeung, G., Alwan, A., 2019. A frequency normalization technique for kindergarten speech recognition inspired by the role of f0 in vowel perception. In: Interspeech 2019.