Activity focused Speech Recognition of Preschool Children in Early Childhood Classrooms

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

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

Abstract

A supportive environment is vital for overall cognitive development in children. Challenges with direct observation and limitations of access to data driven approaches often hinder teachers or practitioners in early childhood research to modify or enhance classroom structures. Deploying sensor based tools in naturalistic preschool classrooms will thereby help teachers/practitioners to make informed decisions and better support student learning needs. In this study, two elements of eco-behavioral assessment: conversational speech and real-time location are fused together. While various challenges remain in developing Automatic Speech Recognition systems for spontaneous preschool children speech, efforts are made to develop a hybrid ASR engine reporting an effective Word-Error-Rate of 40%. The ASR engine further supports recognition of spoken words, WH-words, and verbs in various activity learning zones in a naturalistic preschool classroom scenario. Activity areas represent various locations within the physical ecology of an early childhood setting, each of which is suited for knowledge and skill enhancement in young children. Capturing children's communication engagement in such areas could help teachers/practitioners fine-tune their daily activities, without the need for direct observation. This investigation provides evidence of the use of speech technology in educational settings to better support such early childhood intervention.

1 Introduction

The preschool classroom is a viable space for capturing young children's interactions with teachers and peers. The quality and number of interactions children experience is a key factor in child language development (Hart and Risley, 1995). However, for supporting teachers working with young children with or without developmental delays, the use of direct observations or manual video recording and coding is not a scalable endeavor (Tapp et al., 1995). Sensor-based monitoring tools in classrooms can assist teachers in creating and maintaining a rich learning environment for all children. Feedback from these tools could allow teachers to better identify children who could benefit from further support. Rich and frequently available data can not only help in creating better classroom structure, but also create opportunities to maximize children's communication and interaction (Diamond et al., 2013).

Eco-behavioral observational assessment has often been used to measure moment-to-moment effects with multiple environmental events on specific behaviors and interactions that occur in an early childhood inclusive classroom (Greenwood et al., 1994; Watson et al., 2011). These assessment samples are centered around teacher and child behavior, and overall classroom learning context (e.g., the interactions between them) by adding situational or contextual factors. Specifically for inclusive classrooms, a child's daily interaction can influence their development and by using an eco-behavioral assessment, conclusions can be drawn between environmental contexts and the interactions that occur within them (Brown et al., 1999). These findings can inform practitioners how to arrange their inclusive environments to best support language development of all children. The variety of spontaneous language in an inclusive preschool classroom comes from a variety of speakers and includes both adults and children. Although the Language Environment Analysis (LENA) framework is used extensively by the early childhood research community for a digital measurement system that is automatic (Soderstrom and Wittebolle, 2013; Dykstra et al., 2013; Burgess et al., 2013;

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 92 - 100

Irvin et al., 2017; Greenwood et al., 2018), LENA does not possess an Automatic Speech Recognition (ASR) engine to convert the child speech-to-text, nor does it capture location in the classroom. Apart from conversational speech, children's coordinated movement and location within classrooms also act as an acquisition context driver for critically important skills including language, cognition, and social communication (Eliot, 2000; Council et al., 2000; Piek et al., 2008). Therefore, automatic location tracking within the classroom can provide the ability to monitor interventions while maximizing learning opportunities (Irvin et al., 2018).

Our multi-disciplinary educational research project focuses on quantifying "learning" based on social engagement for use in classroom settings by teachers - and thus we are building a tool that captures the granularity of eco-behavioral observational assessment. It is based on spontaneous interactions between multiple teachers and preschool children (3 to 5 years) with and without developmental delays within naturalistic noisy preschool classroom environments. In this study, we present a translational framework to automatically track conversational speech of preschool children in various activity areas supported by speech technology based on ASR which is fine-tuned specifically for preschool children taking into account their developing nature and developmental delays.

2 Speech and language development in young children

Right from their first babbles, children start developing various speech sounds (Shriberg, 1993) until mid-elementary school. Typically-developing children are expected to progressively acquire various speech sounds throughout early childhood (birth to 8 years). These development occurs primarily in three stages: (i) 'Early' stage from 1 to 3 years, (ii) 'Middle' stage from 3 to $6\frac{1}{2}$, and (iii) 'Late' stage from 5 to $7\frac{1}{2}$. In the 'Early' stage, speech sounds like M (nasal; "mama"), B (stop; "baby"), Y (semivowel; "you"), N (nasal; "no"), W (semivowel; "we"), D (stop; "Daddy"), P (stop; "Pop"), HH (aspirate; "hi") are expected to be developed. While sounds like T (stop; "two"), NG (nasal; "running"), K (stop; "cup"), G (stop; "go"), F (fricative; "fish"), V (fricative, "van"), CH (affricate, "chew"), and JH (affricate, "jump") are expected to be acquired in the 'Middle' stage. Finally, in the 'late' stage, children develop slightly harder sounds like SH (fricative; "sheep"), S (fricative; "see"), TH (fricative; "think,that", R (liquid; "red"), Z (fricative; "zoo"), L (liquid; "like") and ZH (fricative; "measure"). Children may omit, substitute or have inconsistency in production of speech sounds while they are learning. Apart from speech, language planning is also evolving, so word selection and grammar may have issues. Not all children acquire these skills at a similar pace, especially those with developmental delays.

3 Challenges of developing Automatic Speech Recognition systems for young children

Various developmental factors like articulation/pronunciation, motor skills, vocabulary, etc., makes the task of developing ASR systems for children challenging than that for adults (Gerosa et al., 2007). Also, children in early childhood (birth to 8 years) have significantly different speech and language skills as compared to their older peers. Prior research from the Speech Technology community on Children ASR (Stemmer et al., 2003; Shivakumar et al., 2014; Tong et al., 2017; Wu et al., 2019; Shivakumar and Georgiou, 2020; Yeung et al., 2021; Rumberg et al., 2021; Gretter et al., 2021) is captivating. But these focused on: (i) older children, including kindergarten (6-15 yrs), (ii) data collected using head-mounted microphones or close-proximity handheld smartphones in clean/controlled settings under adult supervision, and (iii) with just one speaker using prompts or read stimuli, and limited spontaneous (not scripted) speech. Limited focus and data is available for processing of adult-child interactions in naturalistic preschool settings (3-5 yrs) while they are involved in various activities throughout the day. There is lack of publicly available young child speech corpora (primarily due to privacy/regulations). However, a recent study (Yeung and Alwan, 2018) described various challenges in developing ASR systems for single-word utterances read aloud by kindergarten (5-6 yrs) children achieving a Word Error Rate (WER) of 25%. Therefore, all these factors make the task of developing ASR systems for spontaneous preschool children speech in naturalistic educational settings extremely challenging.

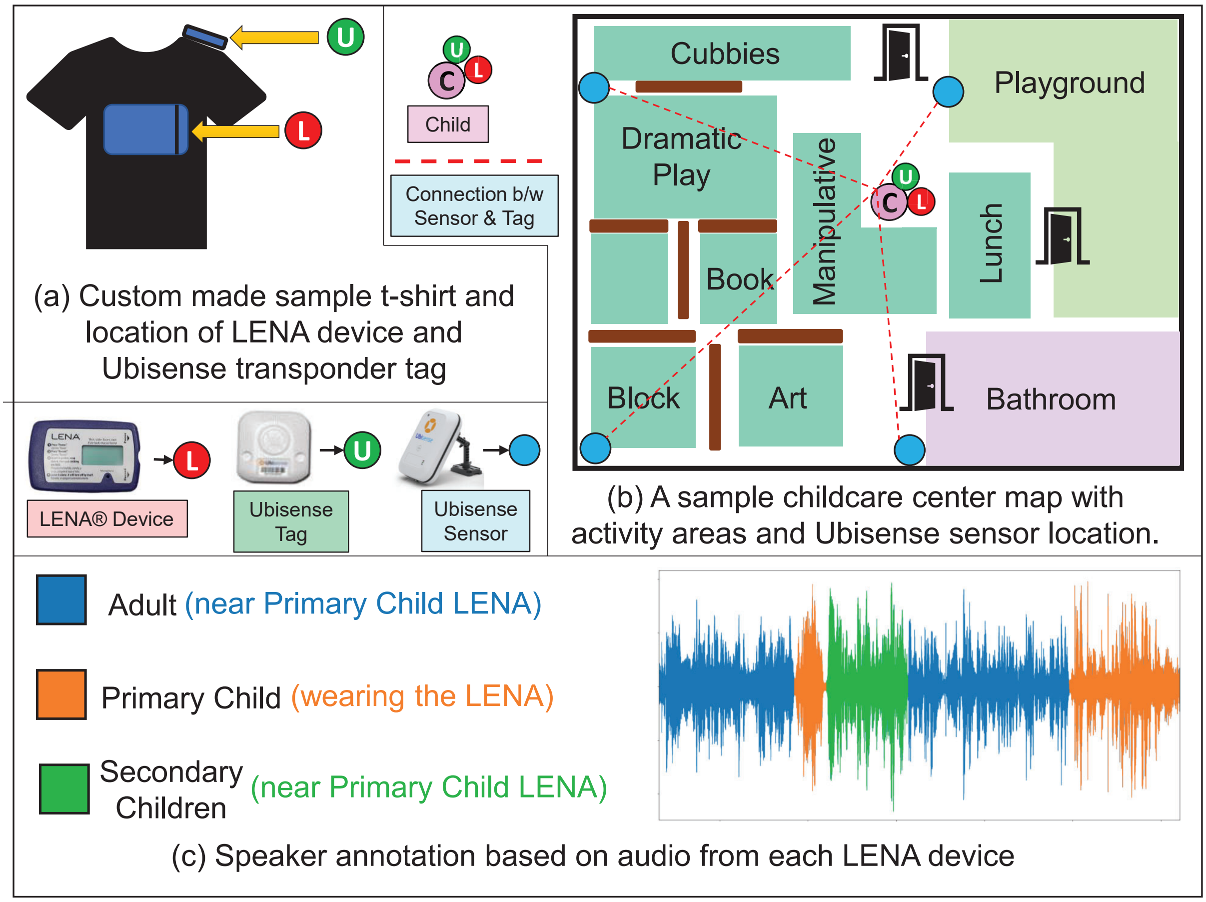

Figure 1: Speech and location data collection in preschool.

4 Data and Collection

4.1 Participants & Procedure

A total of 33 children aged 3 to 5 years with and without developmental delays, and 8 adults teachers participated in this study. The data was collected in preschool classrooms (refer Fig.1(b)) in a large urban community in a Southern state in US, and in multiple sessions over several days in different classrooms with different groups of participants. Data from each participant was linked to an anonymous id for privacy. All participants consented to the use of de-identified data for analysis. This study was approved by the Institutional Review Board of both KU and UTD for analysis.

4.2 Conversational Speech

Conversational speech was collected using a light weight compact digital audio recorder (LENA[1]) attached to participants (refer Fig.1(a)). The LENA Language ENvironmental Analysis system consists of an audio recorder and audio processing software (Ford et al., 2008). The recorder uses an omnidirectional microphone and the final audio is obtained by a computer or laptop running the software to which the recorder is plugged in. The final audio has a sampling frequency of 16 kHz, with a recording unit having a capacity of 16 hours. Although LENA provides adult word counts, conversational turns, and child and peer vocalizations; it does not provide the speech-to-text output. The LENA unit data can be considered as individual audio stream and was tagged into three speaker (Fig.1(c)) categories: Primary child (speech initiated by child wearing that LENA unit), Secondary child (speech originated by any other children within close proximity of primary child), and Adult (speech originated by any adult in close proximity). It is noted that for each LENA audio stream, there is only 1 Primary child and multiple Secondary Children and Adults (e.g., each LENA stream is associated with anonymous child id).

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

4.3 Real-time Location

Ubisense [2], a Real-Time Location Tracking System (RTLS) based on Ultra-Wideband (UWB), is deployed in this study. Ubisense is capable of providing 3D location for every second simultaneously for up to 100 individuals both indoors and outdoors. The RTLS data can also provide information on movement patterns and direction apart from location. This system is established by the combination of receiver sensors and wearable light-weight transponder tags (refer Fig.1(a,b)), both of which broadcast live location co-ordinates to a laptop or computer in network running the Ubisense Location Engine software. With proper calibration, the accuracy of Ubisense is ±15 cm under ideal measurement conditions, and ±30 cm in challenging measurement conditions. Ubisense has been previously deployed in various clinical research studies for individuals at risk for dementia (Kearns et al., 2008; Vuong et al., 2014). Sensors are placed in four corners of the space to ensure maximum coverage and connected to a laptop computer via cords. Then the dimensions of the classroom are established based on the Ubisense measurements, followed by calibrating the real-time location system sensors to their 3-D (x, y, z) locations. Measures to minimize electronic interference caused by other devices (i.e., Wi-Fi routers) was considered. Real-time location was not recorded when the children went outside of the classroom dimension set by Ubisense sensors (like playground).

4.4 Mapping activity area with real-time location information and speech

Activity areas represent information about the location (permanent or temporary) of the child within the physical ecology of an early childhood setting. For this study, various individual literacy areas in the classroom were outlined in consultation with the preschool teachers. These areas are outlined in Table 1. This is followed by setting up boundaries around the individual literacy areas in the classroom using the Geometry feature of Ubisense. This subsequently helped to identify when children wearing a transponder tag were in these areas (refer Fig.1(b)). Ubisense scanning rate was set to 1 Hz. Human-transcriptions of conversational speech, the actual start time of the Ubisense location tracking system, and the actual recording start time of every individual LENA unit (worn by different children) were used for the mapping between the activity areas and spoken text.

Table 1: *Activity Area Codes and their significance.*

Area Code	Significance
Art	Area for painting, drawing, coloring, writing, or sculpting
Snack	Area for snack/food breaks
Block	Areas with large building or construction materials, on floor
Cozy/Book	Areas with books for reading alone or in groups
Computer	Areas for computer activity
Dramatic play	Areas for dress up clothes, kitchen utensils, dollhouse, etc. or that support activities with other children that contain make-believe roles or themes like fireperson, doctor, etc.
Manipulative	Areas for small motor movements of the hand, fingers, wrists, and hand-eye coordination
Story	Areas for reading, listening and telling stories

5 Developing Preschool Children Automatic Speech Recognition System

5.1 Acoustic and Language Modelling

Acoustic model training and decoding experiments were performed using Kaldi (Povey et al., 2011), N-gram language models were trained using SRILM toolkit (Stolcke, 2002) and the RNN-based using PyTorch. Care was taken to avoid overlap of the same group of children between train/test. Ground-truth was based on human transcriptions and only the segments spoken by both primary and secondary children were considered for ASR assessment. The GMM-HMM systems were trained to provide frame-to-phone alignments for the DNN based systems. For the GMM-HMM systems, Mel-frequency cepstral coefficients (MFCCs) (Young, 1996) were extracted for every 25 ms window and 10 ms overlap. 13 MFCCs along with their Δ and $\Delta\Delta$ features were used as front-end features. The input features to the DNN-HMM models included a 40-D high resolution MFCCs of current and neighbouring frames and a 100-D i-vector(Hansen and Hasan, 2015) of the current frame. The i-vectors were calculated by generating speed-perturbed training data with 3 (0.9,1.0,1.1) speed factors. In

[2] https://ubisense.com/

Table 2: *Child ASR Performance.*

#	Features ♣	Acoustic Model Training Data♠	Acoustic Model	Language Model Training Data♠	Language Model	WER (%) of Preschool Test
1	MΔ	PS	GMM-Tri3	LibriSpeech	3-gram	90.28
2	MΔ + I3	PS	TDNN-F(11)	LibriSpeech	3-gram	63.66
3	MΔ + I3	PS	TDNN-F(11)	PS	3-gram	49.02
4	E + I3	PS	TDNN-F(17)	PS	3-gram	47.02
5	E_S + I3	PS	CNN(6) + TDNN-F(9)	PS	3-gram	43.03
6	E_S + I3	PS	CNN(6) + TDNN-F(9) + Attn(1)	PS	3-gram	42.00
7	E_S + I3	PS	CNN(6) + TDNN-F(9) + Attn(1)	PS	LSTM	**40.67**
8	E_S + I3	PS + CMU + OGI	CNN(6) + TDNN-F(9) + Attn(1)	PS + CMU + OGI	3-gram	43.57

♣ MΔ $\rightarrow$ MFCC & Δ & $\Delta\Delta$, E/E_S $\rightarrow$ Filter-Bank Energy (/with SpecAugment), I3 $\rightarrow$ 3* Speed pert. i-vector
♠ PS $\rightarrow$ Preschool, CMU $\rightarrow$ CMU Kids Corpora, OS $\rightarrow$ OGI Kids Corpora

addition, these high-resolution MFCCs were also replaced with 40-dimensional Mel-frequency Filter Banks Energies (MFBE) (Paliwal, 1999) by Inverse Discrete Cosine Transform. Factorized time-delay neural networks (TDNN-F)(Povey et al., 2018a), originally proposed as a data-efficient alternative to TDNN for enhancing ASR performance of low-resource languages with less than 100 hours of data, were primarily used as hidden layers for the hybrid DNN-HMM acoustic models. Apart from TDNN-F layers, CNN layers were deployed. A time-restricted self-attention (Vaswani et al., 2017; Povey et al., 2018b) mechanism (with multiple heads) was also deployed. Another data augmentation approach called SpecAugment(Park et al., 2019) was applied directly to MFBEs. For the RNN-based LMs, we used 2-layer LSTMs of 650 embedding size and 650 hidden dimension. Dropout was considered to overcome overfitting. Lattice rescoring(Li et al., 2021) was used to decode the RNN-based LM. CMU Pronouncing Dictionary[3] was used. Various non-linguistic markers included: laugh, cough, scream, gasp, breath, babble, cry, loud music, crowd and play noise, and other noise. Data-augmentation using publicly available corpora like OGI Kids corpus (Shobaki et al., 2000) ($\approx$ 60 hours; Kindergarten to Grade 10) and CMU Kids corpus (Eskenazi et al., 1997) ($\approx$ 9 hours; Grade 3 to 5) was also considered.

5.2 ASR Model Performance & Discussions

Child ASR performance results are summarized in Table 2. A triphone GMM-HMM Acoustic model trained on Preschool speech generate a very high WER of 90.28% (#1) for pre-trained 3-gram LibriSpeech LM. As shown in #2, using an 11-layer TDNN-F based Acoustic model, 40 MFCC features and speed-perturbed i-vector (of factor 3), a much lower WER of 63.66% was achieved using the same language model. Now in #3, we notice a significant drop of WER to 49.02% by training the language model using our Preschool data. Using a language model trained on in-domain shows much benefit in our study than using pre-trained LibriSpeech language model, as compared to previous studies (Wu et al., 2019; Yeung et al., 2021) for older children speech where Librispeech just worked fine. This signifies that young children do not follow the same language patterns in spoken English or that of adults. In #4, #5, and #6, shows various acoustic model enhancements based on TDNN-F, CNN, and Attention layers with #6 reporting a WER of 42.00%. Finally, in #7 by replacing the 3-gram language model with an RNN-based one, with LSTM layers (see Section 5.1) we reach a WER of 40.67%. As shown in #8, data augmentation does not enhance the performance of the ASR model.

6 Activity-area based Child Speech Recognition and Discussions

All experiment results for this section are summarized in Table 3. The results here are shown for 3 preschool children who were typically developing (without delays) and were present in the same classroom. From a child ASR perspective, these 3 children belong to the test split of the Preschool data and were tagged as primary children (speakers wearing the LENA units). The ASR model used here is the best model as reported in Section 5.2. The results are primarily subdivided into three categories: (i) all words spoken, (ii) WH-words (who, what, where, etc.), and (iii) Verbs; followed by the child IDs: Primary Child #1, #2 and #3. Average WER (irrespective of activity areas) for Primary Child #1, #2 and #3 are 28.49%, 36.13%, and 47.59% respectively. Number of words in sen-

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

Table 3: *Activity-area based child Speech Recognition results.*

Table 3(A)

	Primary Child #1			**Primary Child #2**			**Primary Child #3**		
Activity Area	Time (min)	WER (%)	Words spoken	Time (min)	WER (%)	Words spoken	Time (min)	WER (%)	Words spoken
Art/Snack	18.6	17.39	307	32.3	53.11	270	21.8	56.03	112
Block	<1	13.79	29	1.8	36.36	44	14.7	46.39	217
Computer	4.3	37.5	83	3.3	38.18	55	3.7	23.33	30
Cozy/Book	2.1	NA	0	4	47.61	20	1.9	NA	0
Dramatic Play	4.1	27.1	96	12.4	24.93	384	25.2	43.03	851
Manipulative	<1	12.5	7	9.8	26.62	342	2.1	32.25	31
Story	<1	25	13	1	58.33	12	<1	50	6

Table 3(B)

	Primary Child #1			**Primary Child #2**			**Primary Child #3**		
Activity Area	Time (min)	WH-words (%)	Verbs (%)	Time (min)	WH-words (%)	Verbs (%)	Time (min)	WH-words (%)	Verbs (%)
Art/Snack	18.6	83.33	83.33	32.3	66.67	72.72	21.8	50	50
Block	<1	100	100	1.8	NA	100	14.7	50	71.48
Computer	4.3	100	57.14	3.3	100	60	3.7	NA	50
Cozy/Book	2.1	NA	NA	4	NA	50	1.9	NA	NA
Dramatic Play	4.1	100	66.67	12.4	83.33	84.61	25.2	66.67	68.22
Manipulative	<1	NA	100	9.8	100	82.22	2.1	0	100
Story	<1	100	50	1	NA	66.67	<1	NA	NA

Time (min) = Total time spent by each child in that specific activity area
WER (%) = Word error rate of the ASR model for all words spoken in that specific activity area
Words spoken = Total number of words spoken by each child in that specific activity area
WH-words (%) = Total % of WH-words correctly predicted by the ASR model spoken in that specific activity area
Verbs (%) = Total % of Verbs correctly predicted by the ASR model spoken in that specific activity area
NA = Not applicable; primarily due to no words spoken

tences, WH-words and verbs are a few of the prominent language learning milestones established by the American Speech–Language–Hearing Association [4], outlined by the American Academy of Pediatrics (Gerber et al., 2010; Zubler et al., 2022), and adopted as CDC's (Centers for Disease Control and Prevention) Developmental Milestones[5] program *"Learn the Signs. Act Early."* Table 3(A) shows the time spent by each child in an activity area, followed by WER and all words count spoken in that area. Table 3(B) shows the time spent by each child in an activity area, followed by % of total WH-words and verbs spoken those were predicted correctly in that area by the ASR engine. The "Time Spent" factor is important to better normalize the results across multiple subjects. Primary Child #1 spends the most quality time in 'Art/Snack' area (WER: 17.39%), followed by close to 5 mins in 'Computer'(WER: 37.5%) and 'Dramatic Play'(WER: 27.1%) areas. The amount of spoken words is relatively much higher in 'Art/Snack' area. Child #1 spends less than a minute in 'Block', 'Manipulative', and 'Story' areas, which is also reflected in the word spoken count. Overall across all activity areas, Primary Child #1 spends much less time and spoke less as compared to Child #2 and #3. Primary Child #2 and #3 spent more time in the classroom boundary, and therefore word counts spoken were much higher. Primary Child #2 spends quality time in 'Art/Snack' (WER: 53.11%), 'Dramatic play' (24.93%), 'Manipulative' (26.62%), and close to 4 mins in 'Computer'(WER: 38.18%) and 'Cozy/Book'(WER: 47.61) areas. Primary Child #3 spends quality time in 'Art/Snack' (WER: 56.03%), 'Block' (46.39%), 'Dramatic Play' (43.03%), and close to 4 mins in 'Computer'(WER: 23.33%) areas. Irrespective of the child, performance of the ASR engine in detecting WH-words and verbs across all activity areas is quite good, given the naturalistic noisy dynamic learning environment. While areas like 'Cozy/Book' are more personal learning spaces. Areas like 'Dramatic Play', 'Manipulative', 'Block', 'Art/Snack' alternatively encourage group activity. 'Computer' and 'Story' areas are more focused on listening or seeing. Some observations here can be: (i) Primary Child #1 did not engage much in areas of group activity - signifying difficulty to engage in groups, (ii) Primary Child #1 and #3 produced higher WH-word

[4]https://www.asha.org/public/speech/development/chart
[5]https://www.cdc.gov/ncbddd/actearly/milestones

counts (not shown in the Table) in 'Computer' and 'Dramatic Play' areas - signifying more curiosity. Longitudinal data of the same group of children over a significant time period should help in better informed decisions. However, amendments to classroom structure and plan will be at the discretion of teachers. Performance of the ASR engine can help to monitor/track such elements in a naturalistic preschool classrooms.

7 Towards Data-Based Inclusion Planning in Classrooms

Non-segregated or inclusive educational settings possess a design best suited to prepare young children with disabilities for kindergarten (US Dept. Health, 2015; Barton and Smith, 2015). Careful considerations regarding environmental factors are imperative for meaningful interactions between children in inclusive classrooms (Ganz, 2007). High-quality inclusive classrooms can also foster and support friendship development between children with and without disabilities (Buysse et al., 2008). Through communication skills and social interactions, individuals can begin to form meaningful social relationships and friendships, which could promote positive psychological states (e.g., happiness and self-efficacy; Umberson and Karas Montez, 2010). Teachers and peers as communicators can play important roles for inclusive classrooms to support communication skills of children with disabilities and facilitating social interactions between one another. The quantity and quality of interactions significantly influence the language environment and communication opportunities for young children with disabilities (W Vernon et al., 2018). Also, it may be more important for a child with Autism Spectrum Disorder (ASD) to spend quality time in activity areas that promote language and social engagament because of the social-communication and play limitations that accompany ASD. Using audio recorded by LENA and real-time location using Ubisense supported by advanced speech processing algorithms could provide teachers with information about "what" and "where" child interactions are taking place so that they may be better able to discern when to provide additional support.

8 Conclusion

This study has provided evidence and lays the foundation of deploying sensor-based monitoring tools to acquire and interpret eco-behavioral data (speech and location) in naturalistic early childhood settings to better support teachers and child learning. This work tends to address a major challenge faced by early childhood educators in supporting children (with and without developmental delays) due to a lack of real-time data to inform daily practices and that lead to longer-term school readiness outcomes. Another component in this study has addressed the development of ASR systems for preschool children, which is a very low-resource scenario. Both collection and transcription of such data is a major challenge, especially due to both noisy data and speech intelligibility of young children. Future work will focus on analyzing more children with and without developmental delays, and also collection of such naturalistic data. Future work will also consider speaker group classification (adult vs. child) using speaker-group diarization as compared to human transcriptions.

Acknowledgements

This study was supported by the National Science Foundation Grant #1918032 award to Hansen. The authors would like to thank all the families for participating in this study and the reviewers for their fruitful comments and suggestions.

References

Erin E Barton and Barbara J Smith. 2015. Advancing high-quality preschool inclusion: A discussion and recommendations for the field. *Topics in Early Childhood Special Education*, 35(2):69–78.

William H Brown, Samuel L Odom, Shouming Li, and Craig Zercher. 1999. Ecobehavioral assessment in early childhood programs: A portrait of preschool inclusion. *The Journal of Special Education*, 33(3):138–153.

Sloane Burgess, Lisa Audet, and Sanna Harjusola-Webb. 2013. Quantitative and qualitative characteristics of the school and home language environments of preschool-aged children with asd. *Journal of Communication Disorders*, 46(5-6):428–439.

Virginia Buysse, Barbara Davis Goldman, Tracey West, and Heidi Hollingsworth. 2008. Friendships in early childhood: Implications for early education and intervention.

National Research Council et al. 2000. From neurons to neighborhoods: The science of early childhood development.

Karen E Diamond, Laura M Justice, Robert S Siegler, and Patricia A Snyder. 2013. Synthesis of ies research on early intervention and early childhood education. ncser 2013-3001. *National Center for Special Education Research*.

Jessica R Dykstra, Maura G Sabatos-DeVito, Dwight W Irvin, Brian A Boyd, Kara A Hume, and Sam L Odom. 2013. Using the language environment analysis (lena) system in preschool classrooms with children with autism spectrum disorders. *Autism*, 17(5):582–594.

Lise Eliot. 2000. *What's going on in there?: how the brain and mind develop in the first five years of life*. Bantam.

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

Michael Ford, Charles T Baer, Dongxin Xu, Umit Yapanel, and Sharmi Gray. 2008. The lenatm language environment analysis system.

Jennifer B Ganz. 2007. Classroom structuring methods and strategies for children and youth with autism spectrum disorders. *Exceptionality*, 15(4):249–260.

R Jason Gerber, Timothy Wilks, and Christine Erdie-Lalena. 2010. Developmental milestones: motor development. *Pediatrics in review*, 31(7):267–277.

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

Charles R Greenwood, Carmen Arreaga-Mayer, and Judith J Carta. 1994. Identification and translation of effective teacher-developed instructional procedures for general practice. *Remedial and Special Education*, 15(3):140–151.

Charles R Greenwood, Alana G Schnitz, Dwight Irvin, Shu Fe Tsai, and Judith J Carta. 2018. Automated language environment analysis: A research synthesis. *American Journal of Speech-Language Pathology*, 27(2):853–867.

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

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

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

Dwight W Irvin, Stephen A Crutchfield, Charles R Greenwood, William D Kearns, and Jay Buzhardt. 2018. An automated approach to measuring child movement and location in the early childhood classroom. *Behavior research methods*, 50(3):890–901.

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

William D Kearns, Donna Algase, D Helen Moore, and Sadia Ahmed. 2008. Ultra wideband radio: A novel method for measuring wandering in persons with dementia. *Gerontechnology*, 7(1):48.

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

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

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

Jan P Piek, Lisa Dawson, Leigh M Smith, and Natalie Gasson. 2008. The role of early fine and gross motor development on later motor and cognitive ability. *Human movement science*, 27(5):668–681.

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

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

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

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

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

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

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

Lawrence D Shriberg. 1993. Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders. *Journal of Speech, Language, and Hearing Research*, 36(1):105–140.

Melanie Soderstrom and Kelsey Wittebolle. 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.

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

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

Jon Tapp, Joseph Wehby, and David Ellis. 1995. A multiple option observation system for experimental studies: Mooses. *Behavior Research Methods, Instruments, & Computers*, 27(1):25–31.

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

Debra Umberson and Jennifer Karas Montez. 2010. Social relationships and health: A flashpoint for health policy. *Journal of health and social behavior*, 51(1_suppl):S54–S66.

Education US Dept. Health, Human Services. 2015. Policy statement on inclusion of children with disabilities in early childhood programs. *Infants & Young Children*, 29(1):3–24.

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

Nhu Khue Vuong, Syin Chan, and Chiew Tong Lau. 2014. Automated detection of wandering patterns in people with dementia. *Gerontechnology*, 12(3):127–147.

Ty W Vernon, Amber R Miller, Jordan A Ko, Amy C Barrett, and Elizabeth S McGarry. 2018. A randomized controlled trial of the social tools and rules for teens (start) program: an immersive socialization intervention for adolescents with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 48(3):892–904.

Silvana MR Watson, Robert A Gable, and Charles R Greenwood. 2011. Combining ecobehavioral assessment, functional assessment, and response to intervention to promote more effective classroom instruction. *Remedial and Special Education*, 32(4):334–344.

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

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

Gary Yeung, Ruchao Fan, and Abeer Alwan. 2021. Fundamental frequency feature normalization and data augmentation for child speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6993–6997. IEEE.

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

Jennifer M Zubler, Lisa D Wiggins, Michelle M Macias, Toni M Whitaker, Judith S Shaw, Jane K Squires, Julie A Pajek, Rebecca B Wolf, Karnesha S Slaughter, Amber S Broughton, et al. 2022. Evidence-informed milestones for developmental surveillance tools. *Pediatrics*, 149(3).