Social Robot Learning Companions for Personalized Children's Education

1 Introduction & Problem Definition

Learning to read is one of the most important educational tasks of our schools, but in 2015, only 36% of 4th, 34% of 8th grade students tested on the National Assessment of Educational Progress (NAEP) reached proficiency in reading [1]. Literacy skills like phonologic, alphabetic, and vocabulary knowledge, delivered by quality preschool programs support the development of literacy skills in later grades and can help prevent academic failure [2, 3]. Yet, only about 40% of eligible 4 year olds attend preschool (NIEER, 2013). I One of the most important factors for language skill development is sufficient exposure to a rich variety of spoken language and vocabulary [4]. The social context of exposure is also critical to concept development and the learning experience. Simply hearing language is not enough; children need to actively participate and be emotionally and physically engaged to maximize their learning gains [5].

When a child enters Kindergarten, they possess a unique distribution of cognitive, visual, social and linguistic skills. However, in at-risk communities it is almost impossible for a teacher to offer a curriculum that addresses the diverse cognitive and pre-literacy education needs of each child. Young children would clearly benefit from personalized instruction that can measure and adapt to many intersecting domains of skills and abilities during the process of learning to read.

We propose to address this challenge by developing social robot companions that can continuously assess and effectively personalize to meet individual children's diverse needs. Interactions between a child and a robot resemble the speech acts between children and adults or peers, and offer a unique opportunity to personalize social interactions to promote early literacy skills.

2 Innovation Proposal and Relation to State-of-the-Art

Social robot learning companions have great potential to augment the efforts of parents and teachers to promote learning, academic knowledge, and the wellbeing of children. Existing work with older students has shown that Intelligent Tutoring Systems (ITS) that automatically assess and adapt to student skill levels can positively impact student-learning gains [6].

Social robots that use speech interfaces are a compelling vehicle for bringing these benefits to younger users, and while research into personalized robot tutors has gained increased attention [7], the lack of reliable tools for analyzing children's speech is a major impediment to the further development and widespread deployment of truly impactful speech-based social robot tutors.

Through our research, we propose to address multiple high-impact challenges:

2.1 Development of Automatic Speech Recognition and Spoken Language Understanding systems for young children's speech

Compared to adults, children's speech is characterized by more omissions, substitutions, and mispronunciations, diversity in anatomy and developing motor skills, larger variation in spectral and temporal parameters, and larger number of disfluencies [8, 9]. Though some tools achieve single-digit word error rate (WER) [10, 11] for adult speech, there is still no robust technology for child speech with accuracy rate above 50%; this is especially true for children younger than 7 years old.

Figure 1 shows our analysis of the performance of several cloud-based ASR APIs on child speech, tested with a corpus of 5.5 hours of storytelling speech samples from 75 preschool and kindergarten children(ages 4-7 (5.13 ± 0.64)), collected throughout our previous projects.

In preparation for this project, we have started collecting a large database of children's speech while interacting with robots in both Boston and Los Angeles Schools. We will use this data to train Deep Neural Network (DNN) based systems for Automatic Speech Recognition (ASR) for children's speech that are robust to age, gender, classroom noise, and disfluencies. Due to the challenges of unconstrained children's speech recognition [9], we will restrict the domain to ASR for assessment of literacy skills, focusing on the tasks of co-reading and answering dialogic questions.

In addition to our growing corpus of children's speech samples, we will investigate how speech and language cues develop by quantifying variation in segmental and suprasegmental properties (pitch and duration for example) in a variety of learning scenarios. To maintain ASR robustness against speaker variability and developmental changes, we will apply speaker adaptation techniques (e.g., subglottal-based normalization [13]) to reduce spectral mismatch between training and test utterances. In addition, we will apply constraints imposed by speech production theory to model speaker variability. We also focus on noise-robustness of the ASR systems using noise robust features which retain discriminative information while suppressing information which may confuse the recognizer [14]; variable frame rate analysis in which we adaptively sample features according to their discriminative importance [15, 16]; and missing feature theory to locate and compensate for unreliable spectral components via mask estimation [17, 18]. Finally, to ensure robustness to disfluencies and pronunciation variability, we will build acoustic models and modified dictionaries that explicitly incorporate disfluency behavior and pronunciation analysis to accurately model pronunciation variability of children from diverse linguistic backgrounds [19].

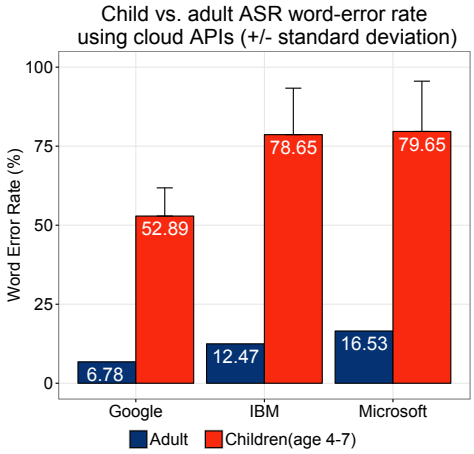

Figure 1: The performance of the state-of-the-art ASR tools is far from functional on child speech.

2.2 Multi-modal assessment and personalization algorithms for Kindergarten age children's spoken language and early reading skills

Robust child speech technologies for literacy skills will enable us to develop and deploy a new class of student models and assessment algorithms based on speech and other multi-modal data. In previous work, we showed that using affective data from facial expression to train models of children's knowledge outperformed traditional Bayesian Knowledge Tracing (BKT) models in assessing single-word reading skills [20]. Given our promising results with this "Affect-BKT" reading assessment algorithm for young children, we propose the following extensions to develop advanced models and algorithms that shall:

Expand beyond tactile based inputs to spoken inputs. Our previous work relied on a child identifying (by tapping on a touchscreen) the written form of a spoken word to assess reading ability. Our development of robust child speech technologies will enable us to incorporate speech-based data, such as spoken responses and pronunciation, enabling new interactions and multimodal models of learning.

Employ active-learning approaches to accelerate algorithm convergence and model training. Generally speaking, assessment models and algorithms improve with additional data. But collecting additional data often requires asking more questions or prompting a child for new demonstrations of a skill. Maintaining child interest over a 20–30 minute educational interaction with an autonomous robot remains a challenge, limiting the amount of useful data from any single session. By employing an active-learning approach, the data we do collect will provide the maximum expected inferential power, allowing our models to more perform better under real-world conditions and practical data constraints.

Personalize to individual students. In part due to data constraints, previous affect-based knowledge models were trained on data from a population of 38 children. By collecting richer, multimodal sources of real-time, real-world data and deploying the active-learning algorithms discussed above, we shall extend our work to develop personalized assessment algorithms, trained on a specific child's unique patterns of affective expression, attention, and prior knowledge.

2.3 Development and evaluation of a fully autonomous, peer-like social robot system in schools and homes with effective educational activities.

In prior work, we assessed children's oral and reading skills offline, after each interaction with a robot, which limited real-time robot behavior adaptation and curriculum personalization. Our enabling innovations of robust child ASR and improved multi-modal assessment algorithms will allow us to develop, deploy, and

Figure 2: 3 activities designed for interactive, educational play between a robot and child to promote literacy skills.

evaluate a completely autonomous social robot system that can adapt to children's unique educational needs over long periods of time at schools and homes.

This robot will interact with children through a suite of educational activity apps integrating automatic speech recognition, automatic assessment and modeling of literacy (reading and linguistic) skills, and personalized content sequencing. The interaction scenario involves the social robot "playing" educational activates with a child like a peer, using a tablet as digital interactive storybook and game surface.

Through interactive storytelling and other educational activities, the social robot system will capture real-time audio, video, touch, and app states on the tablet. This data will be stored in the cloud and used to train novel computational models, assess children's reading and language performance, and adapt the content and robot's behavior in each activity to optimize children's motivation, engagement, and learning.

3 One Year Horizon of project

Building on numerous past system deployments [21, 22, 23] we have developed and evaluated three interactive games and a shared library of grade-appropriate children's content to support multiple-month deployments. First, an interactive storybook, in which the child and the robot alternate telling versions of a given picturebook. Second, a competitive word-pronunciation game, in which the child and robot compete to win points by correctly pronouncing words, and third, a collaborative "I Spy" game, in which the child and the robot work together to identify and pronounce words in a scene that fit a category.

In the first year of the project we are focusing our efforts on designing and evaluating multi-modal assessment algorithms for these games, especially in applying active and transfer learning techniques so models learned from one interaction can be used as a foundation for models in other applications or skills. During gameplay, we will also be recording audio to build up a corpus of high-quality children's speech samples for development of robust children's speech technologies.

As we collect more speech data, we will begin to develop models for children's ASR, with emphasis on: transfer learning to adapt models of adult speech to kids speech, leveraging existing kids' speech databases, exploring model adaptation and also decoding algorithm adaptation for better results, and exploring non-traditional features.

4 Strength of Team to Achieve Proposed Milestones

The UCLA Speech Processing and Auditory Perception Lab has an extensive history developing noise-robust speech recognition and analysis tools for children and adults.

The Personal Robots Group at the MIT Media Lab has developed and procured several robot platforms (see Fig. 3) capable of sustaining engagement and interest in repeated interactions. Over the past several years, they have built extensive cloud-based infrastructure and institutional relationships to support long-term deployments of these systems in schools, hospitals, and homes. Most recently, they completed an 8 week parallel deployment of adaptive storytelling robots across 3 public schools in the Boston area.

The ability to deliver effective, scalable, affordable, early literacy and language interventions to young children is a high impact area for the education system and parents. This proposed work extends and builds upon our teams' established track record of foundational research in the development of social robot learning companions and associated spoken interface technologies for children. By developing robust technologies for analyzing children's speech and combining these technologies with adaptive, personalized assessment algorithms, we propose to significantly advance the state-of-the-art to realize the vision of cloud-connected social robot learning companion systems that deliver effective, socially situated,

and personalized spoken language education experiences for young learners.

Figure 3: Proposed social robot platforms. Tega (left) is a robust research platform designed and built at MIT; Jibo (right) is a commercially-available, consumer-grade social robot platform.

References

[1] National Center for Educational Statistics. (2017, Jan) The condition of education 2015 (nces 2015-144). National Center for Education Statistics. Washington, DC. [Online]. Available: https://nces.ed.gov/pubs2015/2015144.pdf

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

[3] M. M. Páez, P. O. Tabors, and L. M. López, "Dual language and literacy development of spanish-speaking preschool children," *Journal of applied developmental psychology*, vol. 28, no. 2, pp. 85–102, 2007.

[4] S. S. Asaridou, Ö. E. Demir-Lira, S. Goldin-Meadow, and S. L. Small, "The pace of vocabulary growth during preschool predicts cortical structure at school age," *Neuropsychologia*, 2016.

[5] G. Wells, "Dialogic inquiry in education," *Vygotskian perspectives on literacy research*, pp. 51–85, 2000.

[6] M. C. Desmarais and R. S. Baker, "A review of recent advances in learner and skill modeling in intelligent learning environments," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1-2, pp. 9–38, 2012.

[7] D. Leyzberg, S. Spaulding, and B. Scassellati, "Personalizing robot tutors to individuals' learning differences," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*.

[8] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Child speech recognition in human-robot interaction: Evaluations and recommendations," in *Proceedings of the 2017 ACM/IEEE Human-Robot Interaction Conference*, 2017.

[9] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving children's speech recognition through out-of-domain data augmentation," *Interspeech 2016*, pp. 1598–1602, 2016.

[10] G. Saon, H.-K. J. Kuo, S. Rennie, and M. Picheny, "The ibm 2015 english conversational telephone speech recognition system," *arXiv preprint arXiv:1505.05899*, 2015.

[11] P. Price, J. Tepperman, M. Iseli, T. Duong, M. Black, S. Wang, C. K. Boscardin, M. Heritage, P. D. Pearson, S. Narayanan, and A. Alwan, "Assessment of emerging reading skills in young native speakers and language learners," *Speech Communication*, vol. 51, no. 10, pp. 968–984, 2009.

[12] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[13] H. Arsikere, G. K. Leung, S. M. Lulich, and A. Alwan, "Automatic estimation of the first two subglottal resonances in children's speech with application to speaker normalization in limited-data conditions." in *INTERSPEECH*.

[14] B. Strope and A. Alwan, "Robust word recognition using threaded spectral peaks," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 625–628.

[15] Q. Zhu and A. Alwan, "Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise," *Computer speech & language*, vol. 17, no. 4, pp. 381–402, 2003.

[16] H. You, Q. Zhu, and A. Alwan, "Entropy-based variable frame rate analysis of speech signals and its application to ASR," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–549.

[17] B. J. Borgstrom and A. Alwan, "Missing feature imputation of log-spectral data for noise robust ASR," in *Workshop on DSP in Mobile and Vehicular Systems*, 2009, p. 2009.

[18] L. N. Tan and A. Alwan, "Feature enhancement using sparse reference and estimated soft-mask exemplar-pairs for noisy speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1710–1714.

[19] J. Tepperman, J. F. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan, "Pronunciation verification of children's speech for automatic literacy assessment." in *INTERSPEECH*, 2006.

[20] S. Spaulding, G. Gordon, and C. Breazeal, *Affect-Aware Student Models for Robot Tutors*, ser. AAMAS '16. International Foundation for Autonomous Agents and Multiagent Systems, 2016, p. 864–872.

[21] H. W. Park, M. Gelsomini, J. J. Lee, and C. Breazeal, "Telling stories to robots: The effect of backchanneling on a child's storytelling," in *Proceedings of the 12th ACM/IEEE international conference on Human robot interaction*. ACM, 2017.

[22] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective personalization of a social robot tutor for children's second language skills," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 3951–3957.

[23] K. Westlund, M. Jacqueline, S. Jeong, H. W. Park, S. Ronfard, A. Adhikari, P. L. Harris, D. DeSteno, and C. L. Breazeal, "Flat vs. expressive storytelling: Young children?s learning and retention of a social robot?s narrative," *Frontiers in Human Neuroscience*, vol. 11, p. 295, 2017.