The JIBO Kids Corpus: A speech dataset of child-robot interactions in a classroom environment

Natarajan Balaji Shankar,[1] Amber Afshan,[1] Alexander Johnson,[1] Aurosweta Mahapatra,[1] Alejandra Martin,[2] Haolun Ni,[1] Hae Won Park,[3] Marlen Quintero Perez,[2] Gary Yeung,[1] Alison Bailey,[2] Cynthia Breazeal,[3] and Abeer Alwan[1]

[1] *Department of Electrical and Computer Engineering, University of California Los Angeles, Los Angeles, CA, 90095, USA*
[2] *Department of Education, University of California Los Angeles, Los Angeles, CA, 90095, USA*
[3] *MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA*

balaji1312@ucla.edu, *amberafshan@ucla.edu*, *ajohnson49@ucla.edu*, *aurosweta99@ucla.edu*, *alemartin@ucla.edu*, *michaelni12@ucla.edu*, *haewon@mit.edu*, *mquint30@ucla.edu*, *garyyeung@ucla.edu*, *abailey@gseis.ucla.edu*, *cynthiab@media.mit.edu*, *alwan@ee.ucla.edu*

Abstract: This paper describes a novel dataset of children's speech, collected through the use of JIBO, a social robot. The dataset encompasses recordings from 110 children, aged 4 to 7, who participated in a letter and digit identification task, and extended oral discourse tasks requiring explanation skills, totaling 21 hours of session data. Spanning a two-year collection period, this dataset contains a longitudinal component, with a subset of participants returning for repeat recordings. The dataset, with session recordings and transcriptions, is publicly available, providing researchers with a valuable resource to advance investigations into child language development.

[Editor: —] https://doi.org/

Date: 15 October 2024

1. Introduction

The development of language and literacy skills stands as a cornerstone of elementary education. However, empirical findings from the National Assessment of Educational Progress underscore a concerning reality: 37% of fourth-grade students in the United States do not demonstrate reading proficiency aligned with grade-level expectations (Irwin *et al.*, 2022). The foundations of literacy are established in the crucial pre-kindergarten and kindergarten years, where children develop pre-literacy skills such as phonological awareness and letter knowledge (Bus and Van IJzendoorn, 1999). These early developmental stages, therefore, necessitate focused attention and resources to foster language growth.

To enhance the learning experience and capitalize on these advancements, the usage of systems in educational space has become commonplace (Williams *et al.*, 2013), but technological advancements must still address a significant hurdle: the inadequate performance of contemporary automatic speech recognition techniques when tasked with scoring children's responses (Dutta *et al.*, 2022; Yeung and Alwan, 2018). The error-prone nature of automatically generated children's speech transcriptions poses a significant challenge for their integration into educational applications. Research focused on kindergarten-aged children underscores the imperative to specifically tailor ASR systems for this age group, as pre-literacy skills such as phonological and alphabetic knowledge developed at the pre-K and kindergarten level can support the development of literacy skills (Biemiller and Slonim, 2001; Fish and Pinkerman, 2003; Hart *et al.*, 1997; Páez *et al.*, 2007; Snow *et al.*, 2007). However, a notable scarcity of comprehensive children's speech databases persists within the field, particularly with respect to longitudinal datasets.

These longitudinal resources are invaluable for investigating language development and refining child-centered automatic speech recognition and speaker recognition systems (Dutta *et al.*, 2022; Safavi *et al.*, 2012; Yeung and Alwan, 2018). By tracking the same children over time, researchers can map the trajectories of language acquisition. This understanding can guide the development of systems and techniques specifically tailored to the evolving characteristics of children's speech. (Yeung and Alwan, 2019). Longitudinal data also facilitates the development of educational applications specifically tailored to children's voices, through offering insights into how children's speech patterns evolve, supporting applications in areas like personalized learning environments and child-robot interaction.

To effectively collect data from children, researchers must design data collection mechanisms that are engaging and centered on the child's experience. Social robots, with their ability to engage children interactively, hold significant potential as a vehicle for implementing these data-driven insights in clinical and educational settings (Kanero *et al.*, 2018; Kory *et al.*, 2013; Westlund and Breazeal, 2015). Robots can facilitate targeted activities aimed at various objectives, including the evaluation of speech development and phonetic acquisition, and the reinforcement of pronunciation skills.

Leveraging the interactive capabilities of a social robot, JIBO (Spaulding and Chen, 2018), this paper presents a novel children's speech dataset collected over a two-year period. JIBO was employed to administer a series of structured and semi-structured tasks to children in pre-kindergarten, kindergarten, and first grade. These tasks included letter and digit identification, as well as explanation tasks. The dataset's longitudinal component, with a subset of participants returning for follow-up recordings, facilitates the analysis of developmental trajectories in children's speech. As part of a larger human-robot interaction (HRI) study evaluating the effectiveness of social robots in classroom settings in Yeung *et al.* (2019b), Yeung *et al.* (2019a), Tran *et al.* (2020), Johnson *et al.* (2022b), and Johnson *et al.* (2022a), this paper offers a comprehensive discussion of the dataset's collection, encompassing design considerations and recording conditions.

2. Related Work

Several speech corpora exist for studying English-speaking children. For example, the Providence corpus (Demuth *et al.*, 2006) offers longitudinal audio recordings of six English-speaking children, aged 1 to 4, engaged in natural interactions with their mothers. Furthermore, the TBALL dataset (Kazemzadeh *et al.*, 2005) focuses specifically on 256 non-native English speakers, ranging from kindergarten through to fourth grade. The PERCEPT-R (Benway *et al.*, 2022, 2023) corpus comprises data from 281 children, analyzing both typical speech and residual speech sound disorders affecting rhotics. The UltraSuite dataset (Eshky *et al.*, 2018) is a repository of ultrasound and acoustic data, collected from recordings of child speech therapy sessions of 86 children. The SEED dataset (Speights Atkins *et al.*, 2020) comprises 58 children, both with and without speech disorders. Meanwhile, the CID children's speech corpus (Lee *et al.*, 1999) comprises a collection of read speech samples produced by 436 children aged 5 to 17 years. Similarly, the CMU Kids corpus (Eskenazi, 1996) comprises read speech from 76 children, with a narrower age focus of 6 to 11. Additionally, the AusKidTalk corpus (Ahmed *et al.*, 2021) is a large scale corpus of Australian children between the ages of 3 and 12 comprising a collection of single words, utterances, and narrative speech. Likewise, the CSLU OGI Kids' Speech corpus (Shobaki *et al.*, 2000) contains read speech samples from participants encompassing a broad age range from kindergarten to grade ten. Moreover, the CU Kid's Prompted and Read Speech (Cole *et al.*, 2006) corpus comprises read speech data from 663 American English-speaking children aged between 4 and 11 years, while the CU Kid's Read and Summarized Story (Cole and Pellom, 2006) corpus consists of spontaneous speech recordings from 326 children aged between 6 and 11 years. The Birmingham subset of the PF-STAR corpus (Russell, 2006) contains samples from 150 British children between the ages of 4 and 15. The MyST corpus (Ward *et al.*, 2011) contains 499 hours of audio recordings from 1300 children from third to fifth grades interacting with a virtual tutor for science topics. However, to our knowledge, there are no publicly available English language databases that combine both child speech data and a longitudinal component across a large range of speakers, necessitating the creation of the JIBO Kids Corpus.

3. Data Collection

3.1 JIBO

The JIBO Kids Corpus is constructed as a series of structured and semi-structured sessions conducted between a social robot, JIBO, and a child, following a protocol detailed in Bailey and Heritage (2014) and Bailey and Heritage (2018). Initially conceived as a domestic personal assistant robot, JIBO (Spaulding and Chen, 2018) is a social robot capable of 360-degree rotation for expressive body language, including head-tilting, directional gaze shifts, and dance-like movements. To support interaction and assessment delivery, JIBO possesses a small embedded facial screen that primarily displays JIBO's animated eye, synchronized with its physical movements, and also provides a visual interface for presenting text, images, or videos as needed. JIBO's dual cameras are located above the screen, and its microphone array rotates the head to facilitate sound source detection. Two lateral loudspeakers enable audio output for speech and music playback. JIBO's design, including its expressive movements and child-like voice, positions it as the child's peer-like learning companion. This use of social robotics aims to foster a natural conversational setting that encourages spontaneous speech production in young participants.

3.2 Recording Setup

The audio recording setup employed a Logitech C390e webcam, positioned at a 30-45° angle relative to the child and approximately 50 cm away. Recordings took place in an unoccupied office during regular school hours, with some background noise present from the surrounding school environment. The JIBO social robot was placed on a table or desk in front of the child, also at an approximate distance of 50 cm. Adjacent to the child, a researcher assumed the role of an "instructor," engaging in interactive activities with both the child and JIBO. Simultaneously, another researcher, designated as the "operator," oversees the computational aspects and item display through a connected computer interface using a Wizard-of-Oz setup (Dahlbäck *et al.*, 1993). This setup facilitated real-time adjustment of the items displayed by JIBO. In instances where unexpected interactions arose between the child and the social robot, the "instructor" intervened verbally to assist the child in navigating any difficulties encountered during the session. These include instances where the prompt required repetition or children were distracted or reticent. This intervention ensured the smooth progression of the interaction between the child and JIBO.

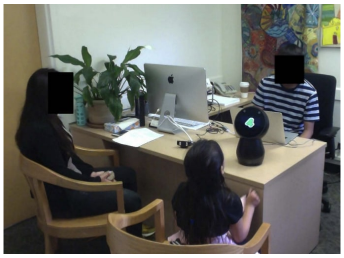

Fig. 1. An example recording session

3.3 Participants

The participants for the study comprised children proficient in English residing in Southern California. Approximately 40% of the participating children reported exposure to additional languages – predominantly Spanish – at home. Nearly one-third of the children were enrolled in a Spanish-English dual language program, and two-thirds were enrolled in an English-only instructional environment. Informed consent was obtained from students and parents for study participation and dissemination of data following school site and institutional review board procedures.

Data collection proceeded over a two-year period. In Year 1, sessions were recorded with a cohort consisting of 38 prekindergarten and 55 kindergarten students. Year 2 of the study involved a cohort of 35 kindergarten and 27 first-grade participants. A subset of children from Year 1 returned the following year, forming a longitudinal cohort. This longitudinal facet enables an exploration of developmental patterns across time, with data available for 22 children progressing from pre-kindergarten to kindergarten and 23 children advancing from kindergarten through to first-grade. A breakdown of speaker grade and gender is present in Tables 1 and 2

To ensure participant privacy, the dataset contains no personally identifiable information, with participants represented solely by anonymized codes. Each child was anonymized in the format TXYY7ZZZ, where: X in {1,2} is the year of the study, YY in {01, 02, 03} is the child's year in school - 01 - Pre-kindergarten (ages 4-5), 02 - Kindergarten (ages 5-6), 03 - grade 1 (ages 6-7). ZZZ is a unique identifier for each child. Boys are odd numbered and girls are even numbered. Thus, T1027235 is child 235, a male kindergartener whose data was collected in year one of the study.

Table 1. Breakdown of participants per year of study. Audio Length indicated is after pre-processing.

Year	Grade	Male	Female	Audio Length (hh:mm)
1	Pre - K	19	19	4:10
	K	23	32	6:57
2	K	20	15	5:42
	1	13	14	4:18

Table 2. Breakdown of Speakers for Longitudinal Study. Audio Length indicated is after pre-processing.

Cohorts	Gender	Participants	Year 1 Audio Length (hh:mm)	Year 2 Audio Length (hh:mm)
Cohort 1	Male	14	1:25	2:09
(Pre-K → K)	Female	8	0:56	1:24
Cohort 2	Male	10	1:16	1:25
(K → 1)	Female	13	1:28	2:19

3.4 Data Pre-processing

To ensure the quality of the data, all audio recordings were subjected to a pre-processing stage. Initial recordings were captured at a uniform sampling rate of 48 kHz, before subsequent downsampling to 16 kHz. All recordings were scrutinized to remove any sessions marred by poor audio quality.

Sessions exhibiting a substantial degree of audio clipping were identified and removed from the dataset. An exception to this criterion was granted for recordings obtained during the ”blocks” task, where pronounced clacking noises are present as participants interact with physical cube toys.

Long periods of silence at the beginning and end of sessions were identified, and the respective sessions were trimmed. Sessions characterized by excessively muted audio levels or instances of inaudible or hushed speech by the participants were also identified and removed. Additionally, sessions contaminated by excessive background noise were eliminated from the dataset. In total, this data cleaning process removed 4:38 hours of noisy audio.

3.5 Transcription

During transcription of all experimental sessions, the speech of child participants were annotated at the word level. Instructor and JIBO speech were excluded from the transcription. All personally identifiable information was redacted from the audio recordings and transcripts. However, sections that exhibited significant crosstalk between the instructor or JIBO, and the child, remain embedded within the accompanying audio recordings.

4. Corpus Composition

JIBO was loaded with educational materials to conduct both a letter and number identification task, along with explanation tasks, utilizing its screen to display visual stimuli. Audio instructions and prompts were recorded by a female researcher with a pitch-shift to emulate a child's voice. Participants were given 2 min of exposure to JIBO's voice through interactive voice prompts prior to the start of the exercises. JIBO intermittently provided positive reinforcement during the session, praising correct answers and fostering engagement. Each session was limited to 30 minutes to maintain consistency and prevent fatigue. The breakdown of average session length, as well as total data collected per task, is presented in Table 3.

4.1 Letter and Digit Identification

In the interviews conducted during the first year of the study, children were asked to identify a sequence of the digits 0-9 and the letters of the English alphabet displayed on JIBO's screen. Accompanying each display was a prompt tailored to the content, such as ”What letter is this?” or ”What number is this?”. Upon the child's identification of the presented letter or number, JIBO transitioned to the next item in the sequence, presenting a new prompt without further intervention or supplementary cues.

For participants in pre-kindergarten and kindergarten in Year 2 of the study, an ”Alphabet Train Game” was introduced to evaluate the child's proficiency in letter identification. Here, JIBO presented a scrolling train on its screen, with each train car adorned with a distinct letter. The child's task involved identifying each letter as it scrolled by. For children in grade 1, a ”Spelling Game” was utilized to gauge their aptitude in basic word decoding. In this task, JIBO displayed a word accompanied by a corresponding image, prompting the child to read the word aloud before subsequently attempting to spell it.

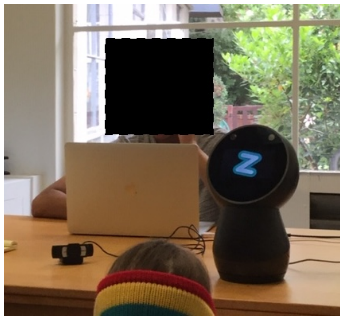

Fig. 2. Sample display of JIBO during the ”Alphabet Train” game

For children in pre-kindergarten and kindergarten in Year 2 of the study, JIBO presented a ”Finger Game”, an image of hands with raised fingers, prompting the child to orally count out the quantity of raised fingers. For children in grade 1, a ”Real World Math Game” tested the math abilities of the children through a series of tasks ranging from simple addition involving candy to more complex scenarios such as determining the age of a child depicted in a birthday celebration scene.

4.2 Explanations

The session recordings for both years of the study featured an interactive component designed to capture spontaneous responses from the participants through extended discourse tasks. In Year 1, children engaged in interviews revolving around two distinct explanation tasks: ”brushing their teeth” (referred to as 'teeth') and ”mixing paint into colors” ('colors'). The children were prompted to articulate their approaches to executing these tasks ('How do you clean your teeth?'), elucidate the rationale behind the task ('Why do you brush your teeth?'), expand on how they would explain the task to a peer ('How would you teach your friend to brush their teeth?'), and justify why their peer should undertake the task in the manner proposed ('Why should they brush their teeth?').

During Year 2 of the study, the teeth-brushing task was revisited ('teeth'), allowing for longitudinal examination of responses. Additionally, participants were presented with a novel task involving an undisclosed number of cubes that could be either joined together or separated, and were tasked with determining the number of cubes provided ('blocks'). Following this task, the same series of questions from the 'teeth' task were posed to elicit insights into participants' approaches, reasoning, and strategies for communicating the counting task to a friend. This semi-structured approach to interactive sessions across both study years offers insights into the developmental trajectories of children's explanation discourse skills, and communicative abilities over time.

5. Conclusion

Table 3. Breakdown of Audio Length by Recorded Session

Task	Year of Study	Grade	Total Audio Length (hh:mm)	Avg. Session Length (mm:ss)
Letter and Digit	1	Pre -K	2:41	4:53
		K	3:59	4:53
	2	K	3:02	5:52
		1	2:19	5:34
Teeth	1	Pre - K	0:41	1:53
		K	1:26	2:06
	2	K	1:01	2:16
		1	0:46	2:18
Blocks	2	K	1:39	3:11
		1	1:12	3:48
Colors	1	Pre-K	0:48	2:10
		K	1:32	2:09
Total			**21:07**	**3:18**

This paper introduces the JIBO Kids Corpus, a unique longitudinal corpus of child-robot interaction speech. This dataset presents a resource for investigating linguistic development in children and advancing automatic speech recognition and speaker verification systems for children. We anticipate that this publicly available dataset will contribute to research on language acquisition and inform the development of educational applications for children.

Acknowledgments
This work was supported in part by the National Science Foundation (NSF).

6. Author Declarations

6.1 Conflict of Interests

The authors have no conflicts of interest to disclose

6.2 Ethics Approval

The research presented in this paper was conducted in accordance with institutional IRB guidelines. All participants, both students and their parents, provided informed consent for the collection and distribution of anonymized speech data. No personally identifiable information was included in the dataset, ensuring participant privacy.

7. Data Availability

The JIBO Kids Corpus described in this paper is publicly available at: https://github.com/balaji1312/Jibo_Kids

References and links

Ahmed, B., Ballard, K. J. *et al.* (**2021**). "AusKidTalk: An Auditory-Visual Corpus of 3- to 12-Year-Old Australian Children's Speech," in *Proc. Interspeech 2021*, pp. 3680–3684, doi: 10.21437/Interspeech.2021-2000.

Bailey, A. L., and Heritage, M. (**2014**). "The role of language learning progressions in improved instruction and assessment of english language learners," Tesol Quarterly **48**(3), 480–506.

Bailey, A. L., and Heritage, M. (**2018**). *Progressing students' language day by day* (Corwin Press).

Benway, N., Preston, J. L. *et al.* (**2022**). "Percept-r: An open-access american english child/clinical speech corpus specialized for the audio classification of /r/," in *Interspeech 2022*, pp. 3648–3652, doi: 10.21437/Interspeech.2022-10785.

Benway, N. R., Preston, J. L. *et al.* (**2023**). "Reproducible speech research with the artificial intelligence–ready percept corpora," Journal of Speech, Language, and Hearing Research **66**(6), 1986–2009.

Biemiller, A., and Slonim, N. (**2001**). "Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition.," Journal of educational psychology **93**(3), 498.

Bus, A. G., and Van IJzendoorn, M. H. (**1999**). "Phonological awareness and early reading: A meta-analysis of experimental training studies.," Journal of educational psychology **91**(3), 403.

Cole, R., Hosom, P., and Pellom, B. (**2006**). “University of colorado prompted and read children’s speech corpus,” Technical Report TR-CSLR-2006-03, Center for Spoken Language Research .

Cole, R., and Pellom, B. (**2006**). “University of colorado read and summarized stories corpus,” Technical Report TR-CSLR-2006-03, Center for Spoken Language Research .

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (**1993**). “Wizard of oz studies: why and how,” in *Proceedings of the 1st international conference on Intelligent user interfaces*, pp. 193–200.

Demuth, K., Culbertson, J., and Alter, J. (**2006**). “Word-minimality, epenthesis and coda licensing in the early acquisition of english,” Language and speech **49**(2), 137–173.

Dutta, S., Tao, S. A. *et al.* (**2022**). “Challenges remain in building asr for spontaneous preschool children speech in naturalistic educational environments,” Proc. Interspeech 2022 .

Eshky, A., Ribeiro, M. S. *et al.* (**2018**). “Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions,” in *Interspeech 2018*, pp. 1888–1892, doi: 10.21437/Interspeech.2018-1736.

Eskenazi, M. S. (**1996**). “Kids: a database of children’s speech,” The Journal of the Acoustical Society of America **100**(4), 2759–2759.

Fish, M., and Pinkerman, B. (**2003**). “Language skills in low-ses rural appalachian children: Normative development and individual differences, infancy to preschool,” Journal of Applied Developmental Psychology **23**(5), 539–565.

Hart, B., Risley, T. R., and Kirby, J. R. (**1997**). “Meaningful differences in the everyday experience of young american children,” Canadian Journal of Education **22**(3), 323.

Irwin, V., De La Rosa, J. *et al.* (**2022**). “Report on the condition of education 2022. nces 2022-144.,” National Center for Education Statistics .

Johnson, A., Fan, R. *et al.* (**2022**a). “Lpc augment: an lpc-based asr data augmentation algorithm for low and zero-resource children’s dialects,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 8577–8581.

Johnson, A., Martin, A. *et al.* (**2022**b). “Can social robots effectively elicit curiosity in stem topics from k-1 students during oral assessments?,” in *2022 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, pp. 1264–1268.

Kanero, J., Geçkin, V. *et al.* (**2018**). “Social robots for early language learning: Current evidence and future directions,” Child Development Perspectives **12**(3), 146–151.

Kazemzadeh, A., You, H. *et al.* (**2005**). “Tball data collection: the making of a young children’s speech corpus,” in *Ninth European Conference on Speech Communication and Technology*.

Kory, J. M., Jeong, S., and Breazeal, C. L. (**2013**). “Robotic learning companions for early language development,” in *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 71–72.

Lee, S., Potamianos, A., and Narayanan, S. (**1999**). “Acoustics of children’s speech: Developmental changes of temporal and spectral parameters,” The Journal of the Acoustical Society of America **105**(3), 1455–1468.

Páez, M. M., Tabors, P. O., and López, L. M. (**2007**). “Dual language and literacy development of spanish-speaking preschool children,” Journal of applied developmental psychology **28**(2), 85–102.

Russell, M. (**2006**). “The pf-star british english childrens speech corpus,” The Speech Ark Limited .

Safavi, S., Najafian, M. *et al.* (**2012**). “Speaker recognition for children’s speech,” in *Proc. Interspeech 2012*, pp. 1836–1839, doi: 10.21437/Interspeech.2012-401.

Shobaki, K., Hosom, J.-P., and Cole, R. (**2000**). “The ogi kids’ speech corpus and recognizers,” Proc. of ICSLP 564–567.

Snow, C. E., Porche, M. V. *et al.* (**2007**). “Is literacy enough? pathways to academic success for adolescents.,” Brookes Publishing Company .

Spaulding, S., and Chen, H. (**2018**). “A social robot system for modeling children’s word pronunciation,” Autonomous Agents and Multi Agent Systems 2018 .

Speights Atkins, M., Bailey, D. J., and Boyce, S. E. (**2020**). “Speech exemplar and evaluation database (seed) for clinical training in articulatory phonetics and speech science,” Clinical linguistics & phonetics **34**(9), 878–886.

Tran, T., Tinkler, M. *et al.* (**2020**). “Analysis of Disfluency in Children’s Speech,” in *Proc. Interspeech 2020*, pp. 4278–4282, doi: 10.21437/Interspeech.2020-3037.

Ward, W., Cole, R. *et al.* (**2011**). “My science tutor: A conversational multimedia virtual tutor for elementary school science,” ACM Transactions on Speech and Language Processing (TSLP) **7**(4), 1–29.

Westlund, J. K., and Breazeal, C. (**2015**). “The interplay of robot language level with children’s language learning during storytelling,” in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction extended abstracts*, pp. 65–66.

Williams, S. M., Nix, D., and Fairweather, P. (**2013**). “Using speech recognition technology to enhance literacy instruction for emerging readers,” in *International Conference of the Learning Sciences*, Psychology Press, pp. 115–120.

Yeung, G., Afshan, A. *et al.* (**2019**a). “Towards the development of personalized learning companion robots for early speech and language assessment,” in *2019 Annual Meeting of the American Educational Research Association (AERA)*.

Yeung, G., and Alwan, A. (**2018**). “On the difficulties of automatic speech recognition for kindergarten-aged children,” Proc. Interspeech 2018 .

Yeung, G., and Alwan, A. (**2019**). “A frequency normalization technique for kindergarten speech recognition inspired by the role of f0 in vowel perception,” Proc. Interspeech 2019 .

Yeung, G., Bailey, A. L. *et al.* (**2019**b). “A robotic interface for the administration of language, literacy, and speech pathology assessments for children,” in *Proc. 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, pp. 41–42.