

# How Effective is LENA in Detecting Speech Vocalizations and Language Produced by Children and Adolescents with ASD in Different Contexts?

Rebecca M. Jones , Daniela Plesa Skwerer, Rahul Pawar, Amarelle Hamo, Caroline Carberry, Eliana L. Ajodan, Desmond Caulley, Melanie R. Silverman, Shannon McAdoo, Steven Meyer, Anne Yoder, Mark Clements, Catherine Lord, and Helen Tager-Flusberg

The LENA system was designed and validated to provide information about the language environment in children 0 to 4 years of age and its use has been expanded to populations with a number of communication profiles. Its utility in children 5 years of age and older is not yet known. The present study used acoustic data from two samples of children with autism spectrum disorders (ASD) to evaluate the reliability of LENA automated analyses for detecting speech utterances in older, school age children, and adolescents with ASD, in clinic and home environments. Participants between 5 and 18 years old who were minimally verbal (study 1) or had a range of verbal abilities (study 2) completed standardized assessments in the clinic (study 1 and 2) and in the home (study 2) while speech was recorded from a LENA device. We compared LENA segment labels with manual ground truth coding by human transcribers using two different methods. We found that the automated LENA algorithms were not successful (<50% reliable) in detecting vocalizations from older children and adolescents with ASD, and that the proportion of speaker misclassifications by the automated system increased significantly with the target-child's age. The findings in children and adolescents with ASD suggest possibly misleading results when expanding the use of LENA beyond the age ranges for which it was developed and highlight the need to develop novel automated methods that are more appropriate for older children. *Autism Research* 2019, 12: 628–635. © 2019 International Society for Autism Research, Wiley Periodicals, Inc.

**Lay Summary:** Current commercially available speech detection algorithms (LENA system) were previously validated in toddlers and children up to 48 months of age, and it is not known whether they are reliable in older children and adolescents. Our data suggest that LENA does not adequately capture speech in school age children and adolescents with autism and highlights the need to develop new automated methods for older children.

**Keywords:** autism spectrum disorder; automated language detection; LENA; communication; acoustic recordings

## Introduction

Language difficulties are a hallmark of neurodevelopmental disorders with the majority of behavioral interventions targeting spoken language [Abbeduto, McDuffie, Thurman, & Kover, 2016]. Changes in social communication are commonly assessed through caregiver or clinician report [Budimirovic et al., 2017; McConachie et al., 2015], yet these measurements can be biased by expectations [Guastella et al., 2015; Jones, Carberry, Hamo, & Lord, 2017; Jones et al., 2015; King et al., 2013; Jones, Tarpey, Hamo, Carberry, & Lord, 2018], warranting the need for objective, direct measures from the child. Speech from the child is typically assessed through human, manual

transcriptions [Berry-Kravis et al., 2013; Sheinkopf, Mundy, Oller, & Steffens, 2000]; however, these procedures are costly and time intensive. In recent years, a portable microphone, LENA digital language processor (DLP), with accompanying automated software, LENA Pro, has facilitated language collection and automated acoustic extraction approaches both in and out of clinic environments [Gray, Baer, Xu, & Yapanel, 2007]. Designed and validated for use in children 0 to 4 years of age, LENA's utility in children 5 years of age and older is not known even for children of developmental language levels below 5 years or above. The present study used acoustic data from two samples of individuals with autism to examine how well LENA automated software performed for detecting child and

From the Weill Cornell Medicine, Center for Autism and the Developing Brain, White Plains, New York (R.M.J., A.H., C.C., E.L.A., M.R.S., C.L.); Department of Psychological and Brain Sciences, Boston University, Boston, Massachusetts (D.P.S., S.M., S.M., A.Y., H.T.-F.); School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia (M.C.); University of California Los Angeles, Semel Institute for Neuroscience and Behavior, Los Angeles, California (C.L.)

R. M. Jones and D. Plesa Skwerer contributed equally to this work.

Received September 10, 2018; accepted for publication December 16, 2018

Address for correspondence and reprints: Rebecca Jones, Weill Cornell Medicine, Center for Autism and the Developing Brain, 21 Bloomingdale Road, White Plains, NY 10065. E-mail: rej2004@med.cornell.edu

Published online 14 January 2019 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/aur.2071

© 2019 International Society for Autism Research, Wiley Periodicals, Inc.

adolescent speech utterances across the clinic and home environments.

The LENA DLP and LENA Pro software automatically detects typically developing infant and young child speech which has enabled large quantities of data to be recorded and analyzed in home settings [e.g., Ramirez-Esparza, Garcia-Sierra, & Kuhl, 2014]. Its hardware and software have been instrumental in providing data to show the importance of adult-child conversations for language acquisition [Zimmerman et al., 2009]. Tools available through LENA replicated seminal work by Hart and Risley [1995] demonstrating that the amount of language to which young children are exposed varies by socioeconomic status [Gilkerson et al., 2017].

Automated LENA tools have also been valuable for evaluating vocalizations and speech in young children with autism. Oller et al. [2010] developed automated algorithms, now commercially available through the LENA research foundation, that categorize differences in infant and toddler speech (10–48 months) in children who are typically developing or have autism or language delay. Work with these algorithms has demonstrated that a full day of data was sufficient for accurately categorizing the vocal age of young children with autism and typically developing children [Yoder, Oller, Richards, Gray, & Gilkerson, 2013]. In addition, children with autism can be further differentiated from typically developing children by the characteristics of conversations [Warren et al., 2010]. Further work has suggested that vocal development in very young children with autism predicts future spoken language [Woynaroski et al., 2017] and is related to standardized language and cognitive assessments [Dykstra et al., 2013]. These studies were carried out on children 48 months and younger; however, the utility of the automated LENA algorithms in children 5 years of age and older is less understood.

The present study addressed whether the current LENA software algorithm would accurately detect vocalizations in children and adolescents with autism beyond preschool age. We focused on 5–18 years of age because these children are often the target age for clinical trials [Berry-Kravis et al., 2012; Scahill et al., 2015; Veenstra-VanderWeele et al., 2017] and there is a significant need for direct measurements of language from this age cohort. We leveraged expressive language samples collected from children with autism from Boston University (study 1) and Weill Cornell Medicine (study 2), who had varying levels of language, across two environmental settings: the lab/clinic (study 1 and 2) and the home (study 2).

## Study 1: Boston University

### Methods

**Participants.** Participants were 24 children and adolescents with ASD between 5 and 18 years enrolled in a

phenotyping study conducted at Boston University (19 males). Based on parent report, all the participants enrolled had little to no functional language and were considered minimally verbal (MV-ASD). They were recruited from a variety of resources in the community such as news and social media, including schools and clinics. Informed consent was obtained from the parents, and study procedures were approved by the Boston University Institutional Review Board.

All participants met criteria for ASD on the Autism Diagnostic Interview-Revised [Rutter, Le Couteur, & Lord, 2003] and the Autism Diagnostic Observation Schedule-2 [ADOS-2; Lord et al., 2012]; diagnoses were confirmed by expert clinical judgment. As part of the research protocol, participants received the ADOS module 1 if aged between 5 and 12 years, or the Adapted ADOS module 1 [Hus et al., 2011; Lord et al., 2012] for adolescents, aged 12 years or older. The Adapted ADOS module 1 is appropriate for assessing ASD symptomatology in older children and adolescents who have limited spoken language (i.e., who do not consistently use phrase speech, comparable to the preverbal/single words level for which ADOS module 1 was designed). It involves activities and materials modified to be more interesting and engaging for adolescents [Bal, Katz, Bishop, & Krasileva, 2016]. Calibrated symptom severity scores (CSS) using the Module 1 algorithm were calculated for Social Affect (SA) and Restricted and Repetitive Behaviors (RRBs).

Cognitive functioning (nonverbal IQ) was measured with the Leiter-3 [Roid, Miller, Pomplun & Koch, 2013], and receptive vocabulary was assessed with the Peabody Picture Vocabulary Test-4 [Dunn & Dunn, 2007]. Table 1 presents the demographic and behavioral characteristics of this sample. Although none of the individuals included in the sample used phrase speech functionally and consistently, the participants showed a range of verbal abilities, from no speech-like vocalizations to fully intelligible simple word combinations. For the purposes of this study, we will refer to the participant's vocal output as vocalizations of the "target-child" even though some of the participants were adolescents. The other speakers present in the room were usually an adult examiner and a parent.

**Data Collection.** The LENA DLP was used to collect vocal output from the child and adults (e.g., examiner and parent) during the standard administration of the ADOS as well as other assessments. When participants first arrived at the Center, every effort was made to have them wear a shirt or vest with a pocket containing the LENA recorder. The LENA recorder was on at all times during the research visits. Some of the participants refused to wear the recorder in the shirt or vest even after several attempts; in such cases, the LENA DLP was placed on the assessment table as close as possible to the participant. The LENA recorder was placed on the table for nine

**Table 1. Demographic Characteristics of Participants in Study 1**

	MV-ASD participants	
	(N = 24)	
	M (SD)	Range
Age	11.18 (4.6)	5.4–18.4
Nonverbal reasoning (IQ) <sup>1</sup>	62.04 (17.8)	30–94
Verbal IQ <sup>2</sup>	34.88 (15.1)	20–66
ADOS severity scores		
Overall CSS	7.58 (2.3)	1–10
SA-CSS	7.27 (2.3)	1–10
RRB-CSS	7.67 (2)	1–10
Boys/girls (N)	19/5	
Race (N)		
African-American	3	
Asian	3	
White	14	
Hispanic	2	
More than one race	2	

<sup>1</sup> Standard scores derived from the Leiter-3 assessment.

<sup>2</sup> Standard scores derived from the PPVT-4 assessment.

of the 24 participants tested. ADOS sessions were also videotaped. The LENA audio-recordings of the ADOS session were transcribed verbatim by a research assistant using the *Systematic Analysis of Language Transcriptions* [SALT; Miller & Iglesias, 2015] system. Then, a second research assistant checked all 24 SALT transcripts against the videotapes of the sessions and edited the transcripts to include any vocalization information that may have been missed when listening to the audio recording.

**Table 2. Demographic Characteristics of Participants in Study 2**

	ASD participants	
	(N = 36)	
	M (SD)	Range
Age	8.03 (3.3)	5.0–17.0
Nonverbal reasoning (IQ) <sup>1</sup>	98.5 (27.2)	29–152
Verbal IQ <sup>2</sup>	93.47 (29.1)	22–162
ADOS severity scores		
Overall CSS	8.11 (1.6)	4–10
SA-CSS	8.03 (1.6)	4–10
RRB-CSS	7.56 (2.2)	1–10
Boys/girls (N)	30/6	
Race (N)		
African-American	1	
Asian	2	
White	25	
Hispanic	1	
Native Hawaiian	2	
Other	1	
More than one Race	4	

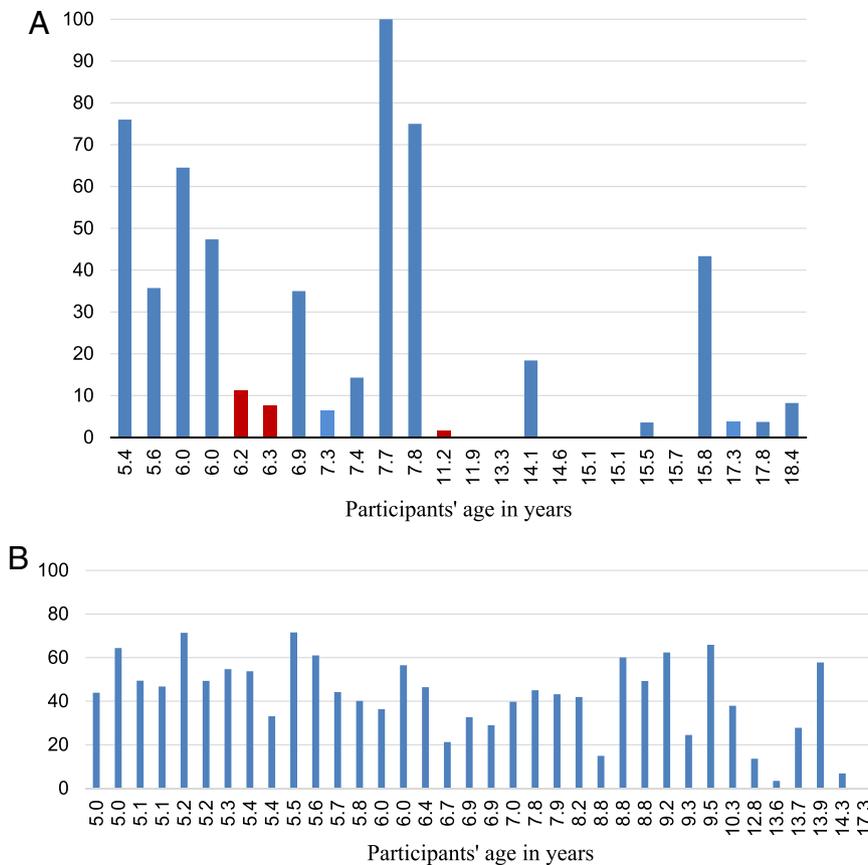
<sup>1</sup> Standard scores or ratio scores derived from the DAS early years, DAS school age, WAIS, Mullen, or WPPSI-IV.

<sup>2</sup> Standard scores or ratio scores derived from the DAS early years, DAS school age, WAIS, Mullen, WPPSI-IV, or PPVT-4 assessment.

### **Manual Transcriptions Timestamped versus LENA**

**Output.** For each participant, the human transcriber selected three time samples of 5 min each from the ADOS evaluation. The transcriber started from the beginning of the recording and listened for the child's first utterance, then marked the transcript as the first sample from that moment until the end of 5 min (time segment 1 or T1). Then the transcriber moved ahead 10 min in the recording, listened for the child's next utterance, and began transcribing from that moment until the end of another 5 min (T2). The transcriber repeated this process (T3) until three samples of 5 min were selected from each child, for a total of 15 min of audio recording per participant. The mean percent correct identifications of target-child utterances by the LENA algorithms did not differ significantly by the time segment analyzed. The next step was to align the time of each child's utterance with the speaker label provided by the LENA output for the same timestamp or as close as possible within 100 ms. Each child utterance was manually timestamped for its beginning and end on the transcript, based on the audio recording "clock." We used the LENA ADEX program to identify the acoustic samples corresponding to the 15 min selected for transcription from the ADOS session. Each time a child vocalization was detected in the human transcription, it was aligned and compared to the corresponding time in the LENA exported file (LENA Interpreted Time Segments or ITS file), which provides codes for the speaker/audio segment targeted (Xu, Yapanel, Gray, & Baer, 2008). Using the timestamped start and end of each target-child utterance as marked by the human transcriber, a "match" was noted when the transcribed child utterance lined up with a LENA target-child speaker code (i.e., CHN). Every instance that the LENA output diverged from labeling target-child speech as "CHN" was recorded manually as a mismatch. (e.g., *Child speaker* temporally lined up with *Male adult*, *Female adult*, *Other child*, *TV/electronics audio*, and *Undetermined noise*). Each type of mismatch between the target-child vocalization as identified by the human transcriber and the LENA label was tallied for further analyses.

The accuracy of the LENA system in identifying target-child utterances was calculated based on the proportion of matches with the human transcriber on speaker identification at the time stamps aligned with the transcription clock. Instances of overlapping speech were subtracted from the tally of speaker identification labels in both the human transcript and in the LENA output file. This method enabled a comparison of "clean" target-child utterances with the speaker code provided by the LENA output. The method of time-aligning at the level of the individual utterances transcribed for each target-child vocalization usually yielded a one-to-one correspondence between the child utterance and one of the labels generated by LENA (i.e., Target-Child, Other Child, Adult Female, Adult Male, Overlap, Noise, Uncertain, Electronic Media, and Silence) for the same timestamp ( $\pm 100$  ms).



**Figure 1.** (a) Percent agreement in study 1 between human transcriber and LENA speaker labels for target-child vocalizations for each participant (in red participants recorded with the LENA DLP on the table, four of the six participants with 0% agreement were recorded with the LENA DLP on the table.). (b) Percent agreement in study 2 between human annotations and LENA speaker labels for target-child vocalizations for a sample of participants ordered by age.

However, sometimes there were several LENA audio segment labels listed within the duration of one target-child utterance; in such cases we marked each occurrence of a different LENA label than CHN during the duration of a single target-child utterance as a disagreement, as noted by the human transcriber.

*Results*

The percent agreement between the LENA output and the manually labeled target-child vocalizations ranged from 0 to 100%, across all participants, with a mean accuracy of 23.15% (*SD* = 29.63) (Fig. 1). The target-child utterances as noted by the human transcriber averaged 1.7 sec in duration, whereas the LENA average duration for target-child utterances was 1.4 sec.

To identify factors that might contribute to the discrepancies between the human transcriber and the LENA output, we first examined whether percent agreement for target-child labeling differed based on the location of the recording device (DLP). In Figure 1, we present the percent agreement for individual participants who did and who did not wear the device during the ADOS session, respectively.

A comparison of these two groups showed that percent agreement on target-child speaker was significantly lower when the LENA was placed on the table  $t(22) = 2.91$ ,  $P = 0.008$ , (Mean percent = 35.01, *SD* = 32.11 for the group wearing the DLP in their clothing and Mean percent = 3.40, *SD* = 4.12 for the participants recorded with the device on the table, respectively). The percent agreement of speaker identifications by LENA also varied as a function of the age of the target-child: Correlations between LENA correct target-child speaker detection and participant age were significant ( $r(15) = -0.677$ ,  $P = 0.006$ ) for those who wore the device and ( $r(24) = -0.563$ ,  $P = 0.004$ ) for the entire sample.

Given the significant difference in the detection of target-child speech based on the placement of the LENA DLP, we further examined the types of disagreement between the manually labeled child utterances and the LENA automated output only for the 15 participants who had the recording device correctly placed in their clothing during the research visit. Of these 15 participants, eight were between 5 and 8 years of age, and seven were over 14 years: LENA classified correctly 56% of utterances as the target-child for the 5 to 8 year olds, whereas percent agreement was only 11% for

the older children, a difference that was statistically significant,  $t(13) = 3.77, P = 0.002$ . Moreover, the type of misclassifications produced by LENA differed by age group: for the younger children, the LENA mislabeled 36.74% of the target-child utterances as “other child” (non-target child) and 37.28% as an adult speaker. In contrast, for the adolescents, only 4.92% of their utterances were misclassified as “other child,” and 46.92% were misclassified as an adult speaker. For both age groups, the other vocalizations that were misclassified were labeled as (FUZ code in the (ITS) Interpreted Time Segments, 18.6% in the younger group and 20.2% in the older group, respectively), with the rest labeled as “overlap.” The LENA code “overlap” refers to an audio segment in which 2 or more speakers vocalize at the same time. Interestingly, no target-child vocalizations were misclassified by the LENA algorithms as TV/electronic media in any age group. Percent agreement between the manually labeled child utterances and the LENA automated output did not differ significantly by the gender of the examiner,  $t(13) = 0.91, P = 0.38$  or the target-child,  $t(13) = 1.98, P = 0.068$ . There was a trend toward higher agreement for male compared to female target-speakers with ASD (mean percent 35.42% for boys and 11.65% for girls), but the small number of girls providing data ( $N = 5$  in the sample of 15 children included in this analysis) prevents us from drawing strong conclusions about this finding.

## Study 2: Weill Cornell Medicine.

### Methods

**Participants.** Thirty-seven families were recruited through the Center for Autism and the Developing Brain (CADB) in White Plains, NY to participate in a study examining novel outcome measures. Participants (target-child) were 5-17 years old (31 boys), see Table 2 for participant demographics. The language level of the target-child varied from two to three word phrases to fluent speech. Weill Cornell Medicine’s IRB approved the study. Caregivers gave written consent; when possible, children 7 years and above assented.

A diagnosis of ASD was confirmed prior to participation by a licensed clinician at CADB using the Autism Diagnostic Observation Schedule (ADOS-2, Modules 1–3) [Lord et al., 2012] or the Adapted ADOS Module 1. CSS for SA and RRBs were calculated. IQ scores were calculated from developmentally appropriate cognitive testing [see Jones et al., 2017].

**Data Collection.** The target-child and their caregiver completed either a 1-week or an 8-week study that involved coming to the CADB clinic on one occasion (1-week protocol) or three separate occasions (8-week protocol) and completing study procedures in their home. Detailed study procedures are summarized in Jones et al. (2017). Briefly,

during the first clinic visit, caregivers were trained on operating the LENA DLP. The target-child wore a t-shirt that contained a pocket for the LENA DLP located on the chest during data collection. All participants completed recordings with the LENA device placed in the t-shirt.

In the clinic, the target-child completed a series of standardized assessments while wearing the LENA device for ~50 min. The assessments included a modified version of the Brief Observation of Social Communication Change [Grzadzinski et al., 2016], as well as the Purdue Pegboard task for 10 min, playing a puzzle game on an iPad for 10 min and watching a series of Pixar short movies on the iPad for 10 min.

In the home, caregivers were instructed to record their child’s speech for 3 days a week up to 1.5 hr per day during week 1 (for both the 1-week and 8-week protocol), as well as weeks 4 and 8 (8-week protocol only). They were encouraged to record their child’s speech during times when the child would likely to be talking with them (e.g., dinner time).

**Manual Transcriptions versus LENA Output.** One target-child was missing human transcriber data and was excluded from analyses. We used the LENA ADEX program to identify segments (LENA Pro analysis software (version V3.4.0)). Two research assistants, (i.e., human transcribers), manually annotated at least 1,500 segments of data (1,540–4,424, mean = 3,070,  $SD = 576$ ) recorded in the clinic and the home for each target-child, ~1 hr of audio data. A graphical user interface (GUI) based labeling toolkit was developed to complete the annotations. The labels that could be assigned to each segment were: target-child vocalization, adult vocalization, silence, environmental noise, multiple speakers, or overlap. There were also affect labels (laughing and crying). This study focused on the labels of target-child and adult vocalizations, as these labels are most relevant to our primary question and parallel to the analyses carried out in study 1. A subset of data were checked for inter-rater reliability between the two research assistants, with agreement between the annotators of 88.06% and Cohen’s  $\kappa$  of 0.82.

To determine percent agreement between the LENA automated output with the human transcriber, we calculated “percent agreement” in the same way as in study 1. We calculated percent agreement separately for data collected in the clinic and data collected in the home. There were no significant differences across the two contexts, so we collapsed the data for all reported results. We calculated a total percent agreement across all participants by summing the correctly identified LENA target-child utterance segments by the number of target-child utterances identified through human transcribers.

We performed a secondary analysis that divided the number of target-child utterances that were *correctly* identified by LENA by the total number of target-child

utterances identified by LENA. This calculation is known in engineering terms as *precision*. In addition, we calculated percent agreement for adult voice as defined above. These values were not available for study 1.

## Results

The percent agreement between the LENA automated algorithms and manually labeled target-child vocalizations ranged from 0 to 72%, with a mean accuracy of 41.66% ( $SD = 18.7$ ) (Fig. 1b). The data suggests that more than half of all child utterances were missed by LENA. The percent agreement of speaker identifications by LENA varied as a function of the age of the target-child: The correlation between LENA correct child speaker detection and participant age was significant  $r(36) = -0.596, P < 0.001$ .

Of the child utterances identified by LENA, human transcribers confirmed that 74.73% were the target-child. However, LENA missed 53.84% of the target-child utterances classified by human transcribers. The percent agreement between LENA automated algorithms and human transcribers coding for adult utterances was on average 55.67%. Approximately 28% of child utterances were misclassified as an adult utterance, further highlighting the difficulty that LENA has in automatically detecting an older child's voice. Lastly, precision for child utterances was 74.73% and 46.60% for adult utterances.

## Discussion

Across two samples of children and adolescents with autism, the automated LENA algorithms did not adequately detect child utterances. This finding is consistent with LENA's recommendations about the target age range for their software. In the two samples, LENA detected the child's voice <50% of the time in children 5 years of age and older. These findings suggest that researchers and clinicians should not assume that the current LENA algorithms are reliable at detecting vocalizations for all child ages.

The language abilities of the participants and the contexts in which language was recorded varied across the two studies, but the results were similar despite these differences. In study 1, the children and adolescents were all minimally verbal and the language samples were collected as part of a standardized diagnostic assessment in the clinic. In study 2, the children and adolescents had varying language abilities, with many individuals with fluent speech, and the samples were collected in both the clinic and the home. While the procedures varied across the studies, all the participants were 5 years of age and older; thus, the data suggest that it was the age of the participants that was the challenge for the LENA algorithms.

The methods used to assess LENA complemented one another across the two studies. In study 1 we used manual transcription of the target-child vocalizations as the starting point, marked the timestamp of each transcribed child utterance and then searched in the LENA segmented output for the speaker label provided at the same timestamp. In study 2, the automated segmentation that LENA relies upon was used as the starting point for the manual transcriptions. Despite these differences, both studies came to similar conclusions. The agreement between the manual labeled data (ground truth) and the LENA automated output varied substantially, as shown in Figure 1a,b. Despite the limited nature of the expressive language produced by the participants in study 1, these children's vocalizations were often mislabeled as adult speech, with this category of misclassifications increasing with the child's age. The high variability in the LENA algorithms' ability to detect the vocalizations of the MV-ASD subjects raises questions about the reliability of the system for automatically detecting targeted speech in this particular population, even though their language is developmentally similar to younger children.

While previous studies suggested that the LENA technology could be successfully used as a screener for preverbal children with ASD [Xu, Gilkerson, Richards, Yapanel, & Gray, 2009; Oller et al., 2010], our findings show that the LENA output does not capture reliably the speech/language of older, school age children, and adolescents with ASD in different contexts. It is likely that the vocal atypicalities of young preverbal children with ASD may be different from those of older children with ASD, even those who remain minimally verbal after school age [Sheinkopf et al., 2000]. The LENA automated analysis was not able to differentiate vocalizations of target children with ASD from other speakers, including adults, although confusions with other types of audio segments (e.g., Noise, Electronic Media, and Silence/SIL) were infrequent. It is possible that the abnormal prosody and speech style of school age children with ASD [Grossman, Edelson, & Tager-Flusberg, 2013; Fusarolli, Lambrechts, Bang, Bowler, & Gaigg, 2017; McCann & Peppe, 2003] may have contributed to the challenges that the LENA automated analyses had detecting child speech.

Several factors emerged as limiting the reliability of the LENA system in detecting the appropriate speaker in the contexts sampled. From study 1, it was clear that the proximity of the DLP device to the target speaker is critical for the LENA algorithms to identify the source of the audio segment correctly, as recommended by LENA procedures, despite the excellent quality of the overall recording provided by the device. Although we made every effort to train participants to wear the DLP, a few would not do so. On the contrary, in study 1, the human transcribers using the audio-recordings provided by the LENA DLP were able to identify the speakers easily even when the device was

placed on the testing table, and the follow-up checking of the transcripts against the video-recordings resulted in minimal revisions. Yet, when the LENA recorder was placed on the table, the automated detection of the target speaker vocalizations was very poor, averaging 3.4% agreement with the human transcriber. Thus, if a participant will not wear the DLP, LENA should not be used for automated analyses regardless of other factors. It is of note that in study 1, none of the participants were verbally fluent. It is possible that the overall lack of intelligible speech contributed to the challenges for LENA's classifications.

Perhaps unsurprisingly, another factor that impacted the percent agreement between LENA speaker classifications and the human transcriber was the age of the target-child. Both correlational results and the analysis of the individual distributions of percent agreement between LENA and the human transcribers showed that LENA speaker misclassifications surged for older children and adolescents, compared to the younger 5 to 8 year olds (where misclassifications were still quite frequent), at least among the MV-ASD participants. Given the changes in voice quality that occur across childhood and adolescence, this finding is not surprising, but may pose a real problem for extending the use of LENA technology to automate coding of natural language samples for adolescents with ASD.

In sum, the findings of these two studies, which included school age, minimally verbal and verbally fluent children and adolescents with ASD, draw attention to possibly misleading results when expanding the use of LENA beyond the population for which it was developed. While the LENA system provides valuable information for the population and recording conditions for which it was designed, there remains a need to develop additional automated methods for coding natural language samples that can be extended to the analysis of vocal output of older and/or minimally verbal individuals with communication impairments. In addition, ideally, such a system would require smaller amounts of live recorded data to train the system than the day-long recording prescribed for LENA [Gilkerson & Richards, 2008]. Further, LENA output is limited by the detection of utterances, not actual words, grammatical morphemes, or sentences. Such language outputs are important for better understanding the challenges and heterogeneity in children and adolescents with ASD [Wittke, Mastergeorge, Ozonoff, Rogers, & Naigles, 2017]. Automatic speech recognition technology is advancing rapidly, but, as suggested by the findings reported here, any new system made available on the market for use in research or in clinical practice should be thoroughly tested on a variety of populations and contexts, to ensure that the tools reliably capture the communication profiles and language environments of the targeted individuals.

## Acknowledgments

Funding sources: SFARI 336363 (PI:CL); SFARI 391635 (PI:CL), R01 HD081199 (PI:CL), P50 DC 13027 (PI:HTF); SFARI 513775 (PI:MAC). Catherine Lord receives royalties from the ADOS and ADI-R and all proceeds related to this project were donated to charity.

## References

- Abbeduto, L., McDuffie, A., Thurman, A. J., & Kover, S. T. (2016). Language development in individuals with intellectual and developmental disabilities: From phenotypes to treatments. *International Review of Research in Developmental Disabilities*, 50, 71–118.
- Bal, V. H., Katz, T., Bishop, S. L., & Krasileva, K. (2016). Understanding definitions of minimally verbal across instruments: Evidence for subgroups within minimally verbal children and adolescents with autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, 57(12), 1424–1433.
- Berry-Kravis, E., Doll, E., Sterling, A., Kover, S. T., Schroeder, S. M., Mathur, S., & Abbeduto, L. (2013). Development of an expressive language sampling procedure in fragile X syndrome: A pilot study. *Journal of Developmental and Behavioral Pediatrics*, 34(4), 245–251.
- Berry-Kravis, E. M., Hessel, D., Rathmell, B., Zarevics, P., Cherubini, M., Walton-Bowen, K., ... Hagerman, R. J. (2012). Effects of STX209 (arbaclofen) on neurobehavioral function in children and adults with fragile X syndrome: A randomized, controlled, phase 2 trial. *Science Translational Medicine*, 4(152), 152ra127.
- Budimirovic, D. B., Berry-Kravis, E., Erickson, C. A., Hall, S. S., Hessel, D., Reiss, A. L., ... Kaufmann, W. E. (2017). Updated report on tools to measure outcomes of clinical trials in fragile X syndrome. *Journal of Neurodevelopmental Disorders*, 9, 14.
- Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test, fourth edition*. Minneapolis, MN: Pearson Education.
- 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.
- Fusarolli, R., Lambrechts, A., Bang, D., Bowler, D. M., & Gaigg, S. B. (2017). Is voice a marker for autism spectrum disorder? A systematic review and meta-analysis. *Autism Research*, 10, 384–407.
- Gilkerson, J., & Richards, J. A. (2008). *The LENA natural language study* (Tech. Rep. LTR-02-2). Boulder, CO: LENA. Retrieved from [www.lenafoundation.org/wp-content/uploads/2014/10/LTR-02-2\\_Natural\\_Language\\_Study.pdf](http://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-02-2_Natural_Language_Study.pdf)
- Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., ... Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2), 248–265.
- Gray, S. S., et al. (2007). The LENA Language Environment Analysis System: The Infoture Time Segment (ITS) File.
- Grossman, R. B., Edelson, L. R., & Tager-Flusberg, H. (2013). Emotional facial and vocal expressions during story retelling

- by children and adolescents with high-functioning autism. *Journal of Speech, Language, and Hearing Research*, 56, 1035–1044.
- Grzadzinski, R., Carr, T., Colombi, C., McGuire, K., Dufek, S., Pickles, A., & Lord, C. (2016). Measuring changes in social communication behaviors: Preliminary development of the brief observation of social communication change (BOSCC). *Journal of Autism and Developmental Disorders*, 46, 2464–2479.
- Guastella, A. J., Gray, K. M., Rinehart, N. J., Alvares, G. A., Tonge, B. J., Hickie, I. B., ... Einfeld, S. L. (2015). The effects of a course of intranasal oxytocin on social behaviors in youth diagnosed with autism spectrum disorders: a randomized controlled trial. *Journal of Child Psychology and Psychiatry*, 56(4), 44–452.
- Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H Brookes Publishing.
- Hus, V., Maye, M., Harvey, L., Guthrie, W., Liang, J., & Lord, C. (2011). *The adapted ADOS preliminary findings using a modified version of the ADOS for adults who are nonverbal or have limited language*. In Poster presented at the International Meeting for Autism Research, San Diego, CA.
- Jones, R. M., Carberry, C., Hamo, A., & Lord, C. (2017). Placebo-like response in absence of treatment in children with Autism. *Autism Research*, 10(9), 1567–1572.
- Jones, R. M., Risi, S., Wexler, D., Anderson, D., Corsello, C., Pickles, A., & Lord, C. (2015). How interview questions are placed in time influences caregiver description of social communication symptoms on the ADI-R. *Journal of Child Psychology and Psychiatry*, 56(5), 577–585.
- Jones, R. M., Tarpey, T., Hamo, A., Carberry, C., & Lord, C. (2018). Smartphone measures of day-to-day behavior changes in children with autism. *npj Digital Medicine*, 34.
- King, B. H., Dukes, K., Donnelly, C. L., Sikich, L., McCracken, J. T., Scahill, L., ... Hirtz, D. (2013). Baseline factors predicting placebo response to treatment in children and adolescents with autism spectrum disorders: A multisite randomized clinical trial. *JAMA Pediatrics*, 167(11), 1045–1052.
- Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism Diagnostic Observation Schedule: ADOS-2*. Los Angeles, CA: Western Psychological Services.
- McCann, J., & Peppe, S. (2003). Prosody in autism spectrum disorders: A critical review. *International Journal of Language and Communication Disorders*, 38(4), 325–350.
- McConachie, H., Parr, J. R., Glod, M., Hanratty, J., Livingstone, N., Oono, I. P., ... Williams, K. (2015). Systematic review of tools to measure outcomes for young children with autism spectrum disorder. *Health Technology Assessment*, 19(41), 1–506.
- Miller, J., & Iglesias, A. (2015). *Systematic Analysis of Language Transcripts (SALT), Version 16 [Computer software]*. Middleton, WI: SALT Software, LLC.
- Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., ... Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences of the United States of America*, 107(30), 13354–13359.
- Ramirez-Esparza, N., Garcia-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, 17(6), 880–891.
- Roid, G. H., Miller, L. J., Pomplun, M., & Koch, C. (2013). *Leiter International Performance Scale, (Leiter 3)*. Torrance CA: Western Psychological Services.
- Rutter, M., Le Couteur, A., & Lord, C. (2003). *Autism Diagnostic Interview – Revised*. Los Angeles, CA: Western Psychological Services.
- Scahill, L., McCracken, J. T., King, B. H., Rockhill, C., Shah, B., Politte, L., ... Research Units on Pediatric Psychopharmacology Autism, N. (2015). Extended-release guanfacine for hyperactivity in children with autism spectrum disorder. *The American Journal of Psychiatry*, 172(12), 1197–1206.
- Sheinkopf, S. J., Mundy, P., Oller, D. K., & Steffens, M. (2000). Vocal atypicalities of preverbal autistic children. *Journal of Autism and Developmental Disorders*, 30(4), 345–354.
- Veenstra-VanderWeele, J., Cook, E. H., King, B. H., Zarevics, P., Cherubini, M., Walton-Bowen, K., ... Carpenter, R. L. (2017). Arbaclofen in children and adolescents with autism spectrum disorder: A randomized, controlled, phase 2 trial. *Neuropsychopharmacology*, 42(7), 1390–1398.
- Warren, S. F., Gilkerson, J., Richards, J. A., Oller, D. K., Xu, D., Yapanel, U., & Gray, S. (2010). What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism. *Journal of Autism and Developmental Disorders*, 40(5), 555–569.
- Wittke, K., Mastergeorge, A. M., Ozonoff, S., Rogers, S. J., & Naigles, L. R. (2017). Grammatical language impairment in autism spectrum disorder: Exploring language phenotypes beyond standardized testing. *Frontiers in Psychology*, 8, 532.
- Woynaroski, T., Oller, D. K., Keceli-Kaysili, B., Xu, D., Richards, J. A., Gilkerson, J., ... Yoder, P. (2017). The stability and validity of automated vocal analysis in preverbal preschoolers with autism spectrum disorder. *Autism Research*, 10(3), 508–519.
- Xu, D., Yapanel, U., & Gray, S. (2009). *Reliability of the LENA™ Language Environment Analysis System in Young Children's Natural Home Environment*. Tech. Rep. LTR-05-2. Boulder, CO: LENA Foundation. Retrieved from LENA Foundation: <http://www.lenafoundation.org/TechReport.aspx/Reliability?LTR-05-02>.
- Xu, D., Yapanel, U., Gray, S. S., & Baer, C. T. (2008). *The LENA Language Environment Analysis System: The Interpreted Time Segment (ITS)*. LENA™. Tech. Rep. LTR-04-2, Boulder, CO: The LENA Foundation. Retrieved from [www.lena.org/wp-content/uploads/2016/07/LTR-04-2\\_ITS\\_File.pdf](http://www.lena.org/wp-content/uploads/2016/07/LTR-04-2_ITS_File.pdf).
- Yoder, P. J., Oller, D. K., Richards, J. A., Gray, S., & Gilkerson, J. (2013). Stability and validity of an automated measure of vocal development from day-long samples in children with and without autism spectrum disorder. *Autism Research*, 6(2), 103–107.
- Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., & Yapanel, U. (2009). Teaching by listening: the importance of adult-child conversations to language development. *Pediatrics*, 124(1), 342–349.