Behavior Research Methods (2021) 53:113–138
https://doi.org/10.3758/s13428-020-01419-y

Circumspection in using automated measures: Talker gender and addressee affect error rates for adult speech detection in the Language ENvironment Analysis (LENA) system

Matthew Lehet[1] · Meisam K. Arjmandi[1] · Derek Houston[2] · Laura Dilley[1]

Published online: 24 June 2020

Abstract
Automatic speech processing devices have become popular for quantifying amounts of ambient language input to children in their home environments. We assessed error rates for language input estimates for the Language ENvironment Analysis (LENA) audio processing system, asking whether error rates differed as a function of adult talkers' gender and whether they were speaking to children or adults. Audio was sampled from within LENA recordings from 23 families with children aged 4–34 months. Human coders identified vocalizations by adults and children, counted intelligible words, and determined whether adults' speech was addressed to children or adults. LENA's classification accuracy was assessed by parceling audio into 100-ms frames and comparing, for each frame, human and LENA classifications. LENA correctly classified adult speech 67% of the time across families (average false negative rate: 33%). LENA's adult word count showed a mean +47% error relative to human counts. Classification and Adult Word Count error rates were significantly affected by talkers' gender and whether speech was addressed to a child or an adult. The largest systematic errors occurred when adult females addressed children. Results show LENA's classifications and Adult Word Count entailed random – and sometimes large – errors across recordings, as well as systematic errors as a function of talker gender and addressee. Due to systematic and sometimes high error in estimates of amount of adult language input, relying on this metric alone may lead to invalid clinical and/or research conclusions. Further validation studies and circumspect usage of LENA are warranted.

Keywords LENA · Speech · Language · Automatic processing · Validation · Error

Introduction

Language is a quintessential human behavior which children must learn through exposure to competent talkers. It is well-established that rates of speech and language skill attainment depend on the quantity of speech experienced by young children (Greenwood, Thiemann-Bourque, Walker, Buzhardt, & Gilkerson, 2011; Hart & Risley, 1995; Hoff & Naigles, 2002; Montag, Jones, & Smith, 2018; Romeo et al., 2018; Rowe, 2012; Weisleder & Fernald, 2013; Weizman & Snow, 2001). Specifically, developmental language attainment appears best predicted by the quantity of language directed to young children themselves – i.e., the amount of so-called *infant-directed speech* – rather than the amount of overheard or *adult-directed speech* (Romeo et al., 2018; Weisleder & Fernald, 2013). Quantification of the amount of language in children's natural home environments – and ideally the amount directed to children themselves – is therefore an essential method for both basic behavioral research as well as clinical purposes. In research, accurate quantification of the amount of language in a child's home is critical for evaluating theoretical questions about the nature of language development (e.g., Montag et al., 2018; Shneidman, Arroyo, Levine, & Goldin-Meadow, 2013; Weisleder & Fernald, 2013). Further, in clinical practice, child early interventionists use information about amounts of caregiver communication to assess the effectiveness of caregiver-centered interventions for enhancing frequency and quality of child-directed communications (Roberts & Kaiser, 2011; Vigil, Hodges, & Klee, 2005).

Matthew Lehet and Meisam K. Arjmandi contributed equally to this work.

✉ Laura Dilley
ldilley@msu.edu

[1] Department of Communicative Sciences and Disorders, Michigan State University, 104 Oyer Centerp, East Lansing, MI 48824, USA

[2] Department of Otolaryngology – Head and Neck Surgery, The Ohio State University, Columbus, OH, USA

Technological devices for automatic speech processing have become increasingly popular methods of quantifying amounts of ambient language in children's environments. One device widely employed by researchers and clinicians is the Language ENvironment Analysis (LENA™; LENA Research Foundation, Boulder, CO) system (Christakis et al., 2009; Ford, Baer, Xu, Yapanel, & Gray, 2008; Gilkerson, Coulter, & Richards, 2008; Gilkerson & Richards, 2008; Greenwood et al., 2011; Xu, Yapanel, & Gray, 2009c; Zimmerman et al., 2009). This system consists of an audio recorder capable of holding up to 16 hours of audio within a vest worn by a child. LENA uses offline software to generate an automated Adult Word Count and other measures that have now been widely used in numerous basic scientific and applied clinical studies and settings (Burgess, Audet, & Harjusola-Webb, 2013; Caskey, Stephens, Tucker, & Vohr, 2011, 2014; Caskey & Vohr, 2013; Johnson, Caskey, Rand, Tucker, & Vohr, 2014; Oller et al., 2010; Pae et al., 2016; Sacks et al., 2014; Soderstrom & Wittebolle, 2013; Suskind, Leffel, et al., 2016b; Thiemann-Bourque, Warren, Brady, Gilkerson, & Richards, 2014; Wang et al., 2017; Warlaumont, Richards, Gilkerson, & Oller, 2014; Warren et al., 2010; Weisleder & Fernald, 2013; Zhang et al., 2015). This includes several large-scale (e.g., city-sponsored) programs deploying LENA devices for providing data used for decision-making relative to provision of clinical and social services and evaluation of the effectiveness thereof (e.g., the city of Providence LENA effort) (Wong, Boben, & Thomas, 2018).

Yet, questions remain about LENA's strengths – and weaknesses – as a tool for language quantification (Cristia et al., in press). *The goal of this paper is to provide rigorous evaluation of potential sources of error in LENA's automated language measures, thereby providing researchers and interventionists with high-quality information to assist judicious and circumspect deployment of LENA's time-saving automated technology*. We focused on addressing unanswered questions about the accuracy of LENA's measure of ambient language by adult speakers, a measure known as Adult Word Count. We focus on this measure due to its widespread adoption as a means of quantifying ambient language in a child's environment. Although LENA generates other automatic measures of language, such as conversational turn counts and/or vocalizations by the child wearing the vest, the focus of the present project was on accuracy in quantification of adult language input. Given this priority for developing the human coding system used in the present project, the methods used in this paper do not permit full evaluation of these other metrics, which is beyond the scope of present work.

To date, it remains unanswered whether the accuracy of LENA's language quantification is high and consistent across family home environments, and whether accuracy is affected by incidental variables such as the genders of talkers, and/or whether they are addressing a child (or an adult). Achieving a high, consistent level of accuracy across these incidental variables is essential for drawing appropriate, correct interpretations about the language input provided to children within individual family environments. It would be concerning, for instance, if LENA's technology incorrectly and systematically attributed a mother's utterances to a child – particularly a child who was at risk for developmental disorder or delay. If these sorts of errors were attested, then invalid clinical or research inferences could readily be drawn from automatic language measures, specifically that: (a) the caregiver was vocalizing *less* than she actually was, and that (b) the child was vocalizing *more* than he actually was. Such a pattern of errors could, for instance, lead clinicians to believe that the caregiver, rather than the child, should be the focus of intervention efforts – for example, through use of caregiver-oriented therapies aimed at increasing the quantity of speech input she provides to the child. Errors in measurement of the actual amount of speech content in the environment, if large enough, could potentially obscure profoundly the true picture of behavior. Large errors of undercounting would make families who were providing a lot of language input appear as if they had provided little language input. Large errors of overcounting would make families who were providing very little language input appear as if they had provided a lot of language input. We therefore pose the following rhetorical and practical question: What is an acceptable degree (and type) of error in quantification of amount of language input in a child's environment? If this error is too great, it may suggest a need to develop alternative measures of automated language quantification and/or encourage a return to greater reliance on hand-coding by human listeners.

A number of studies have provided validation data of LENA's automatic language estimates by narrowly examining the extent to which its final Adult Word Count estimate correlates with counts generated by humans of numbers of words spoken by adults across samples of recordings (Table 1). Other studies (Ambrose, Walker, Unflat-Berry, Oleson, & Moeller, 2015; Burgess et al., 2013; Ramírez-Esparza, García-Sierra, & Kuhl, 2014) have transcribed audio from LENA recordings to calculate word counts from human transcription but did not directly compare these word counts to LENA's Adult Word Count. These studies collectively indicate that, across large numbers of recordings, the LENA Adult Word Count tends to follow the general trend of human-counted numbers of words in recordings. However, the proportion of unexplained variance ($1 - r^2$) in this relationship has been shown to range up to 40% (cf. Soderstrom & Wittebolle, 2013). Thus, for any *given* recording, there is little basis for predicting whether LENA's Adult Word Count will be under- or overcounting actual numbers of adult spoken words, or how big the actual discrepancy will be.

Table 1 Correlation (Pearson's r) coefficients between LENA- and human-quantified adult word counts from prior studies

Authors	Language	Pearson's r	Sample
Xu et al. (2009c)*	English	.92	One-hour samples with high vocal activity from $N = 70$ families recorded at home (a total of 4200 minutes)
McCauley et al. (2011)*	English	.81	Five-minute segments from $N = 30$ preschool recordings (a total of 150 minutes)
Caskey et al. (2014)	English	.93	$N = 5$ 5-minute recordings from a neonatal intensive care unit (a total of 25 minutes)
Gilkerson et al. (2018)	English	.95	A total of 5000 minutes from $N = 94$ families (including the 70 families from Xu et al. 2009)
Soderstrom & Wittebolle (2013)	English	.76	One hundred eighty-three 5-minute intervals from $N = 11$ children recorded at home and at daycare (a total of 915 minutes)
Schwarz et al. (2017)*	Swedish	.67	Forty-eight 5-minute intervals selected from $N = 4$ 12-hour recordings (a total of 240 minutes)
Weisleder & Fernald (2013)	Spanish	.80	Sixty minutes constructed from non-contiguous 5-minute intervals from 10 at-home recordings
Oetting et al. (2009)*	English	.71 and .85	Seventeen 30-minute samples of prerecorded mothers and their children (a total of 510 minutes)
Gilkerson et al. (2015)	Chinese	.73	Three 5-minute samples from daylong at-home recordings of $N = 22$ families (a total of 330 minutes)
Busch et al. (2017)	Dutch	.87	Forty-eight 5-minute samples from 8 recordings from 6 children (a total of 240 minutes)
Canault et al. (2016)	French	.64	Three hundred twenty-four 10-minute samples from home recordings of $N = 18$ children recorded at 3 time points (a total of 3240 minutes)
Pae et al. (2016)	Korean	.72	Twenty-seven 10-minute samples from home recordings and 36 10-minute samples from an experimental reading task (a total of 630 minutes)

Note. Citations with asterisks (*) did not appear in peer-reviewed journals

However, correlation coefficients are a poor means of assessing accuracy, not to mention variability in accuracy (Busch et al., 2017). Estimates which were "fully" accurate would entail a function $y = x$ (with slope 1 and intercept 0), indicating that for each "true" value of x, no adjustment to the slope or intercept offset is needed to obtain a value y. Correlations instead indicate the degree of scatter of values around an *arbitrary* line of best fit and thus do not reveal degree of measurement bias. Measurement bias might be proportional (i.e., a difference in slopes of best-fit lines from 1) or fixed (i.e., a non-zero intercept; Busch et al., 2017; Ludbrook, 1997). Therefore, evaluations of whether one method (e.g., actual human word counts) can be replaced with another (LENA's Adult Word Count estimates) should not be based solely on correlations. Further, it is not clear that using ordinary least squares to regress results from one method (e.g., LENA's Adult Word Count) on another (e.g., human word counts) is a valid way of comparing methods given that under such methods, a low least-squares error could be obtained while still demonstrating proportional or fixed biases (Bland & Altman, 1986; Busch et al., 2017; Ludbrook, 1997). Therefore, additional nuanced examination of accuracy is warranted.

By design, accuracy of LENA's Adult Word Count depends on its accuracy with respect to a *prior* step, namely accuracy in classifying audio as human communicative vocalizations. However, the literature is lacking in rigorous, peer-reviewed, independent studies examining the relationship between LENA's degree of accuracy in classifying audio as human communicative vocalizations and accuracy in its Adult Word Counts. Indeed, LENA's Adult Word Count is the end result of multiple, hierarchically dependent signal-processing steps for classifying audio sound sources, and errors at any stage could potentially be compounded to affect Adult Word Count accuracy. The initial steps of LENA's algorithms involve classifying (i.e., labeling) stretches of audio of variable length as female adult speech (labeled as FAN in LENA's ADEX software), male adult speech (MAN), key child (CHN), other child (CXN), overlapping vocalization (OLN), TV/electronic media (TVN), noise (NON), silence (SIL), or uncertain (FUZ). Next, the seven categories other than silence are divided into "near-field" or "far-field" sounds based on the energy in the acoustic signal. Next, short stretches of audio categorized as (near-field) speech or speech-like vocalizations by an adult or child that are temporally close to one another are grouped together into units called "conversational blocks", while remaining contiguous stretches of audio classified as "far-field" (or "faint") are reclassified as "Pause" units (on the rationale that speech in such audio may likely be unintelligible or hard to hear) (Xu, Yapanel, Gray, & Baer, 2008a; Xu, Yapanel, Gray, Gilkerson, et al., 2008b). Finally, stretches of audio classified as (near-field) male or female adult speech (MAN or FAN) are used to derive LENA's Adult Word Count values.

The LENA Foundation has provided data on the relationship between Adult Word Count accuracy and segment classification accuracy. In a well-cited but unpublished study, Xu et al. (2009) reported an overall Pearson's *r* of 0.92 between human transcription and LENA's Adult Word Count within 1-hour samples from 70 recordings. A number of important details of their analysis are not reported, such as how temporal mismatch between human-identified vocalization events and LENA's machine classifications was dealt with in the agreement analysis. This is a critical detail for a rigorous analysis of agreement which is necessary to be able to replicate a study, since otherwise it is not clear under what conditions a human transcription should count as having agreed with an automatic classification or not. Xu et al. further reported a difference in word count estimates (human – LENA) for two separate 12-hour recordings, one in a quiet environment and one in a noisy environment; the difference was roughly −0.4% for the former but −27.3% for the latter. This tantalizing finding suggests substantial variability in Adult Word Count accuracy may occur in the LENA system, though this remains largely unexplored.

A handful of studies have evaluated LENA's accuracy at classifying audio, as opposed to Adult Word Count accuracy. Perhaps the most widely cited example is the non-peer-reviewed study provided by LENA discussed above by Xu et al. (2009; Xu, Yapanel, Gray, Gilkerson, et al., 2008b), which has been cited many times in support of the claim of reliability of LENA classification (Ambrose, VanDam, & Moeller, 2014; Caskey & Vohr, 2013; Dykstra et al., 2013; Gilkerson, Richards, & Topping, 2017a; Gilkerson, Richards, Warren, et al., 2017b; Greenwood et al., 2017; Greenwood et al., 2011; Johnson et al., 2014; Marchman, Martínez, Hurtado, Grüter, & Fernald, 2017; Ota & Austin, 2013; Ramírez-Esparza, García-Sierra, & Kuhl, 2017; Richards, Gilkerson, Xu, & Topping, 2017a; Richards, Xu, et al., 2017b; Sangwan, Hansen, Irvin, Crutchfield, & Greenwood, 2015; Thiemann-Bourque et al., 2014; VanDam, Ambrose, & Moeller, 2012; Warlaumont et al., 2010; Warlaumont et al., 2014; Xu, Gilkerson, Richards, Yapanel, & Gray, 2009a; Xu, Richards, et al., 2009b; Zhang et al., 2015). The classification accuracy data reported by Xu et al., which was re-reported in Christakis et al., (2009), Zimmerman et al. (2009), and Warren et al. (2010), was based on human coding generated for another unpublished study (Gilkerson et al., 2008). Xu et al. reported classification accuracy for LENA of 82%, 76%, and 76% for adult, child, and other segments, respectively, based on a frame size of 10 ms, as reported in Warren et al. (2010) as opposed to the original study by Xu et al. (2009). This data set has also been analyzed in great detail for the accuracy of child vocalization classification (Oller et al., 2010). However, Xu et al. (2009; p. 5) state that their algorithm for sampling the audio for use in the analysis "was designed to automatically detect high levels of speech activity between the key child and an adult", leaving unclear whether their sampling procedure might have introduced bias into estimates of accuracy in cases when sampling did not rely solely on portions automatically determined by LENA to involve high levels of speech activity.

Groups outside of the LENA organization have also investigated classification by LENA. Ko, Seidl, Cristia, Reimchen, and Soderstrom (2016) randomly selected LENA-defined segments (50 FAN and 50 CHN) from 14 recordings (1400 total segments). Humans then manually coded these segments. LENA's mean accuracy was 84%; however, accuracy ranged between 51% and 93% across recordings, indicating a great deal of variability in accuracy. A similar recent analysis of classification accuracy (Seidl et al., 2018) had human listeners code 1384 LENA-defined FAN and CHN segments. They found overall accuracy of 72% with confusion between FAN and CHN segments occurring 15% of the time.

VanDam and Silbert (2013, 2016) elaborated upon other classification results by determining factors in the audio that predict accuracy in LENA. They selected 30 segments each from 26 recordings that LENA had classified as FAN, MAN, or CHN. Human listeners classified these LENA-defined segments as mother, father, child, or other. Human listeners classified segments LENA identified as FAN or MAN as adult speech 80% of the time. They further found evidence that LENA's classification relied on fundamental frequency (F_0) and duration as major criteria for deciding among adult male, adult female, or child talkers. Missing from studies of LENA's audio classification reliability, among other things, are robust assessments of LENA's false negative rate (since many studies have focused only on stretches of audio that LENA had identified as a talker), a thorough characterization of variability in accuracy across multiple families, and identifying how classification error carries over to LENA's Adult Word Count.

Further, none of the studies mentioned above assessed whether there are systematic biases in accuracy of LENA's classification of audio or Adult Word Count estimates across adult talkers or situations. Given VanDam and Silbert's (2016) finding that LENA appears to rely heavily on F_0 and duration to classify a talker as a man, woman, or child, it is notable that F_0 varies considerably as a function of many factors, including talker gender, speaker size, emotional state, and/or communicative intent (Bachorowski, 1999; Benders, 2013; Fernald, 1989; Pisanski et al., 2014; Pisanski & Rendall, 2011; Podesva, 2007; Porritt, Zinser, Bachorowski, & Kaplan, 2014). Situation-specific speech register could potentially affect accuracy in LENA, something especially important for clinical and research issues in child language. Adults often adopt an infant-directed (ID) speech register when speaking with young children, typically characterized by higher and more variable F_0 (i.e., dynamic pitch) and slower rate (i.e., longer durations) relative to an adult-directed (AD) register, along with shifts in other kinds of acoustic cues (e.g., distributions of vowel formants; Cristia

& Seidl, 2013; Kondaurova, Bergeson, & Dilley, 2012; Wieland, Burnham, Kondaurova, Bergeson, & Dilley, 2015). Therefore, the intended *addressee* –child or adult – can have implications for distributions of acoustic cues – especially F_0 and duration – in ID vs. AD speech, potentially systematically affecting LENA performance. The gender of a talker and the addressee of a segment of speech – whether addressing a child or an adult – could in theory systematically affect accuracy of LENA's measures. Ensuring the consistency and comparability of metrics in this widely-used device is important for ensuring the soundness of theoretical claims or clinical guidance made on LENA's output.

The present study was designed to provide important new data regarding variability and consistency in LENA's accuracy for quantifying children's language environments across families in English. We focused on LENA's accuracy in a corpus of recordings comprised of many families with children at risk for language disorder/delay due to hearing loss. There is a high need for accurate and valid quantification of language input to at-risk children, both for research purposes and since the data might be used to guide clinical decision-making and assess effectiveness of caregiver-centered interventions. Prior studies have shown that caregivers use acoustically similar speech behaviors when speaking to children with hearing impairment as to typical-hearing children (Wieland et al., 2015; Bergeson, Miller, & McCune, 2006), or else that caregivers may pronounce speech in a manner similar to how they speak to adults (Kuhl et al., 1997; Lam & Kitamura, 2010). Caregivers may also make minor acoustic modifications when speaking to children who have hearing impairment, many of which may make speech acoustically clearer and words more detectable (Lam & Kitamura, 2012). For example, they may use a slightly higher second formant (F2) for vowels (Wieland et al., 2015), or they may raise their pitch slightly more than they would to a typical-hearing child of the same age (Bergeson, Miller, & McCune 2006). If LENA shows high error in a sample where children are at risk for delay and/or disorder under conditions where adults are expected to be speaking at least as – if not more – clearly than they would speak to typically developing children or adults, then this would be important information to bring to light for interventionists and behavioral researchers to consider in determining whether and when to deploy LENA technology. To the extent that LENA is being deployed as a technology used to detect speech vocalizations by adults (and children), it is important to show that it is capable of robust, high, accurate performance in quantifying caregiver utterances produced in cases of at-risk child populations.

We asked whether differences in classification accuracy for human vocalizations could explain differences in Adult Word Count accuracy. In our study, we selected audio from cases where LENA had and had not identified speech in order to evaluate LENA's accuracy more thoroughly than prior studies. Finally, an important goal was to quantify how accuracy in LENA's classifications and Adult Word Count might differ based on the gender of the talker and the addressee in ID vs. AD speech. Previewing our results, we found that LENA's classification and Adult Word Count accuracy depended on both the gender (male vs. female) of the talker and the addressee (ID vs. AD).

Methods

The present study was conducted as part of initial phases of a larger NIH-funded project at Ohio State University and Michigan State University focused on investigating how the amount and quality of language input in a child's environment predicts language development. This study was an initial validation test and assessment of whether LENA's Adult Word Count was suitable as a primary dependent measure for our broader project. Specifically, we asked (1) whether error in LENA's Adult Word Count was small and consistent across families; (2) whether this error was unbiased across and robust to conditions of interest, i.e., ID vs. AD speech; and (3) whether the amount of error was affected by extraneous factors, such as whether talkers were male vs. female. Satisfying (1), (2), and (3) were necessary preconditions for using LENA's Adult Word Count as a primary metric for planned individual differences research. The study was also designed to permit identifying systematic sources of inaccuracy or bias in LENA classification steps that might help explain downstream inaccuracies in calculation of the LENA Adult Word Count.

Participants LENA recordings used in the present study were collected in pilot and initial stages of the larger NIH-funded project described above. Families gave permission to participate and to have their child wear a LENA system for at least one day. The research was approved by the institutional review boards at The Ohio State University and Michigan State University. The present study was based on a single day-long recording from each of a total of 23 enrolled families who had completed at least one day-long LENA recording at the time of initiation of the present study. If an enrolled family had completed more than just one LENA recording, as called for under the broader grant protocol, then the first LENA recording made was included in the present study. Within each family, one child under 3 years of age was given a LENA vest to wear for the recording (M = 1 year 8 months, SD = 8.8 months; range: 4 months to 2 years 8 months at the time of recording). (See Appendix Table 11 for details.) For four families, the target children (i.e., those wearing the LENA device) had normal hearing, for eight families, the target child had hearing aids, for two families, the target child had a cochlear implant in one ear and a hearing aid in the other, and for nine families, the target child had bilateral cochlear implants.

Children with cochlear implants had 3–22 months (M = 10 months, SD = 7.54 months) of post-implantation hearing experience. The first available recording from each family where the child had hearing experience was used.

General research design and selection of audio Our approach involved the following steps: (1) running LENA's software on the entire day-long audio recording of family language environments; (2) sampling audio from the day-long recording; (3) enlisting human coders to (a) identify times during sampled audio when they heard speech vocalizations, and, for adults' speech, determine whether it was child- or adult-directed, and (b) count the number of words in adult speech utterances; (4) parceling sampled audio into 100-ms frames, then for each frame, compare the code from humans with that from LENA; and (5) compare human word counts and LENA's Adult Word Count estimates.

Prior published studies of LENA's classification accuracy have not estimated the proportion of intelligible speech which LENA inaccurately classifies as non-speech (i.e., the false negative rate). Our study thus sought to estimate a false negative rate in part by sampling pause units, i.e., portions of audio which LENA had classified as *not* containing near-field speech, as well as from conversational blocks, i.e., portions of audio which LENA had classified as containing near-field speech (although see Schwarz et al., 2017; Soderstrom & Wittebolle, 2013 for analysis of Adult Word Count accuracy that included audio from LENAdefined pauses). Thus, unlike prior classification studies (e.g., VanDam & Silbert, 2016), our design permitted estimation of LENA's classification rates of true positives, true negatives, false positives, and false negatives for categories like speech vs. non-speech.

LENA's algorithm was first deployed on the entire day-long audio recording using its offline Advanced Data Extractor (ADEX) classification software (v. 1.1.3-5r10725). From each family's recording, we then excluded audio for which the child was asleep based on context in the audio which evidenced prolonged heavy breathing, the parents saying goodnight, and/or other contextually based cues to naps, since there was no communicative relevance for the child of any adult speech during those times. Next, we selected the first and last 30 "adult-speech" conversational blocks, i.e., those that had been classified by LENA's ADEX classification software as involving at least one adult talker – female (FAN) or male (MAN) – as a primary participant[1]. The selection of conversational blocks containing adult speech was motivated by the desire to use LENA's Adult Word Count metric, which is only calculated for segments of adult speech. In total, samples of approximately 30 minutes of audio ("sampled audio") were drawn from the beginnings and endings of each recording. These times were selected because family members were likely to be at home and engaged in routine, child-centered activities, e.g., waking up, eating morning or evening meals, and getting ready for bed. As such, this audio was deemed likely to be a fairer test of LENA's capabilities as it was deemed likely to directly assess the home environment without variability introduced by families engaging in a wide-ranging set of daily activities. Additionally, given our priority of maximizing reliable determination of when ID vs. AD speech was happening from context, sampling audio from the beginning and end of the day had the benefit of enhancing continuity of understanding situational contexts of communicative interactions, which other sampling methods might not have afforded. If the total duration of either the first 30 or last 30 adult speech conversational blocks was less than 10 minutes of audio length, then for whichever portion(s) – first or last – that fell below 10 minutes of audio, we included the next (or preceding, respectively) consecutive adult speech conversational block until the 10-minute audio length minimum was reached. This yielded a minimum of 20 minutes of sampled audio from adult speech conversational blocks per recording (M = 22.98 min, SD = 5.36 min, *range*: 20.02–44.33 min). There was considerable variability in individual conversational block durations across recordings that collectively constituted sampled audio (M = 10.65 sec, SD = 21.07 sec, *median* = 4.17 sec, *range*: 0.6–529.97 sec).

The sampled audio also included approximately 9 minutes of short chunks of audio from pause units (i.e., audio that LENA had identified as not containing near-field speech), which were interleaved between audio portions of adult speech conversational blocks from the beginning and end of the day that had been selected as described above. The mean portion of sampled audio from pause units was M = 9.31 minutes (SD = 0.43 min, *range*: 8.75–9.96 min). Sampled audio from pause units was selected by first dividing pause units that fell between selected adult conversational blocks into 5-second chunks; chunks were then randomly selected for study inclusion until 5 minutes total duration from pause units was selected at the beginning and at the end of the file. Any portions of sampled audio that incidentally overlapped with a conversational block consisting of primarily child talkers were excluded. After this exclusion, if the total duration of sampled audio from pause units was less than 4 minutes, then additional 5-second chunks of pause units were randomly included in the sample until a minimum of 4 minutes from pause units was achieved. Durations of pause intervals between the sampled conversational blocks with adult speech varied considerably (M = 31.9 sec, SD = 231.0 sec, *median* = 10.9 se., *range*: 2.3–12062.9 sec). Across all

[1] Each conversational block was classified via LENA's black-box methods as consisting of up to three "primary" participants named by the title corresponding to the block type code. For instance, one conversational block type we selected for sampling was Adult Female with Key Child (AICF), but such blocks may have had a few frames classified as a male adult talker (MAN). The full list of conversational block type included in this study and their named correspondences is listed in Appendix Table 12.

selected recordings, sampled audio for analysis (from conversational blocks and pause units) consisted of a mean of 32.29 minutes of audio per participating family (*SD* = 5.42; *range*: 29.07–54.15 min). The total size of the sample was 735 minutes of total coded audio, which compares favorably with the amount of audio examined in the other studies listed in Table 1. Independent variables of primary interest for statistical hypothesis testing were (1) the gender of adult speakers identified by human listeners (male vs. female) and (2) addressee (ID vs. AD). Due to the spontaneous nature of speech, not all conditions were represented for all families.

Coding of human communicative vocalizations by human analysts In this study, ten trained human analysts identified entire stretches of human communicative vocalizations (i.e., speech or speech-like vocalizations by adult male, adult female, or child talkers) within sampled audio and marked the temporal starts and ends of these entire intervals on the relevant textgrid tier (see Fig. 1) in Praat (Boersma & Weenink, 2017). For stretches judged to have been adult speech, analysts indicated whether the speech was directed to a child (ID speech), to an adult (AD speech), or neither (e.g., self-directed speech, pet-directed speech) based on context. Laughs, burps, sighs, and other non-speech noises made in the throat (e.g., in surprise) were not treated as speech nor as speech-like vocalizations. Stretches of speech that were unintelligible, due to being, e.g., very faint or distant, were likewise not identified nor labeled, consistent with LENA's goal of excluding "far-field" speech unlikely to contribute to child language development. Analysts counted all words within a contiguous stretch of speech attributed to a single adult talker and typed a number into the relevant Praat textgrid interval.

Additional details of the coding procedure ensured that LENA was given "the benefit of the doubt", e.g., concerning handling of acoustic overlap of talkers, and that minor temporal discrepancies did not count against agreement. First, recall that LENA assigns a single ADEX segment code to each successive chunk of audio. This is potentially problematic for LENA in the case of overlapping sound sources such as overlapped human vocalizations. In these cases, LENA is forced to "choose" between a single talker code (MAN, FAN, CHN, or CXN) or else a multi-talker code, (OLN), which stands for "overlapped speech or noise" during which no adult words are estimated. During OLN intervals, LENA does not increment the Adult Word Count. Recognizing LENA's classification algorithms might handle such cases unreliably, coders were instructed that, in the special case that they detected overlapped speech in a portion of signal, they should consider LENA's labels and favor a coding consistent with LENA's interpretation. In particular, if the coder heard the portion as containing two or more overlapping talkers but LENA assigned a single talker code for that portion of signal, then in the case that the coder perceived the portion to have contained speech from a talker consistent with the single talker code which LENA had indicated, coders were instructed to attribute that portion to the single talker which LENA has identified by marking the interval on a tier for that talker type. If, on the other hand, LENA had assigned the multi-talker OLN code to the overlapped speech, then coders were instructed to mark the portion of overlapped speech on the "*Excluded Vocalizations*" tier (tier 9 in Fig. 1). Conversely, if the coder heard the portion as containing two or more overlapping talkers but LENA assigned a single talker code that did not correspond to either of the talkers the coder heard, then the coder transcribed that portion on the "*Excluded Vocalizations*" tier. Given that OLN codes and "*Excluded Vocalizations*" entailed no increment to LENA's Adult Word Count, this handling had the effect of ensuring that speech frames identified as consisting of overlapped human vocalizations which were coded as OLN were essentially not treated in our analysis as speech (since they were not attributed to a single adult male, adult female, or child talker,

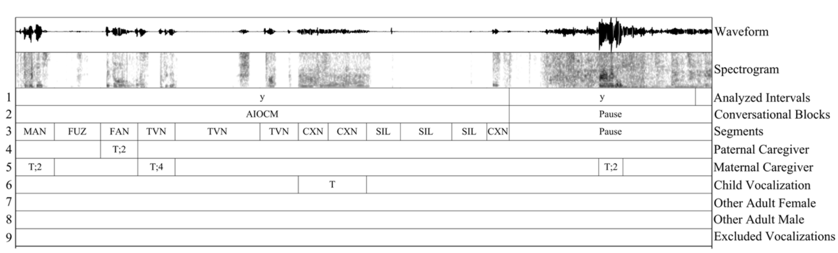

Fig. 1 Annotation scheme used by human analysts to identify human communicative vocalizations. The top of the display shows the waveform and spectrogram. Textgrid tiers provided for the following information (top to bottom): (1) The Analyzed Intervals tier indicated sampled audio portions (given with a "y"). (2) The Conversational Blocks tier depicted starts and ends of conversational blocks (e.g., AIOCM for adult male with target child) and pause units. (3) The Segments tier depicted LENA's segment level ADEX code from among its sound categories. Analysts also indicated starting and ending points of human communicative vocalizations as shown above for: (4) the paternal caregiver, (5) the maternal caregiver, (6) a child, (7) another adult female, or (8) another adult male. For tiers corresponding to adult speech, i.e., (4), (5), (7), and (8), analysts indicated the addressee (e.g., "T" for child or "A" for adult), and they also typed a number corresponding to the judged number of intelligible words in each adult speech interval. Finally, (9) the Excluded Vocalizations tier was used to mark speech that significantly overlapped with other noises, other speech, or speech-like vocalizations and were marked as overlap noise by LENA. Analysts determined the tier that audio should be assigned to based exclusively on their judgment of the audio recording

consistent with LENA's handling). Second, to further advantage LENA and speed coding, analysts copied the temporal boundaries of LENA's ADEX codes by default to mark the starts and ends of speech events, only changing those times relative to LENA if LENA was incorrect by more than 100 ms (a value in line with prior estimates of LENA's temporal accuracy: Ko et al., 2016). This meant that minor (< 100 ms) discrepancies that LENA may have had with the actual start or end of vocalization did not count against LENA in our agreement quantification algorithms.

Analyses of agreement between human and LENA classification Following humans' labeling continuous human speech vocalizations as above, our general approach to determining when LENA and human coders agreed was to: (i) divide sampled audio into short 100-ms frames; (ii) determine the human-derived category characterizing each frame (see below); (iii) determine the LENA classification code characterizing each frame; and then to (iv) determine, for each frame, whether the category implied a match between the LENA code and the human-derived category. Accuracy (and error) was then calculated as a percentage of frames showing consistency (or inconsistency) between the LENA code and the human-derived category.

(i) *Divide sampled audio into short frames*. Each textgrid annotating the sampled audio was first divided into a sequence of frames using MATLAB R2017b (The MathWorks website [http://www.mathworks.com]) and the *mPraat toolbox* (Bořil & Skarnitzl, 2016), following prior work (Atal & Rabiner, 1976; Deller, Hansen, & Proakis, 2000; Dubey, Sangwan, & Hansen, 2018a, 2018b; Ephraim & Malah, 1984; Proakis, Deller, & Hansen, 1993; Rabiner & Juang, 1993). A *100-ms* frame length was chosen first based on the granularity of LENA segmentation accuracy in prior literature (for instance, Ko et al., 2016); secondly, based on the instructions to human coders regarding the granularity of their decisions when LENA segment boundaries deviated from perceived audio; and finally, based on the observation that 100 ms is one sixth the size of the smallest LENA segment (600 ms) and one twelfth the size of the average segment (M = 1260 ms, SD = 760 ms), providing a meaningful resolution for sampling LENA's classification of audio. Frames containing audio outside of sampled audio were discarded.

(ii) *Determine the human-derived category characterizing each frame*. Next, for each frame, we determined the human label that best characterized that frame (*adult male*, *adult female*, *child,* or *other*). This step was necessary, given the possibility that the start or end of the continuous stretch of human vocalization that a coder had identified might have started or ended part-way through the 100-ms frame; as a result, one part of the frame might have been identified by a coder as speech from one category, while the other part of the frame was identified by a coder as from the other category. Therefore, the label that was assigned to the frame was the one which corresponded to the label that took up the greatest temporal extent (i.e., 50 ms or more) of the frame. For instance, if 90% of a frame's temporal extent was identified as an adult male talker and 10% as an adult female talker, the frame was classified as an *adult male speech* frame. Regions coded by humans as either "paternal caregiver" or "other adult male" were treated as *adult male speech*, and regions coded by humans as either "maternal caregiver" or "other adult female" were treated as *adult female speech* (see Table 2). Frames of *adult speech* (either by a male or female) were characterized as an "AD" or "ID" frame if 50% or greater of the frame's temporal extent had been annotated as adult-directed or infant-directed, respectively (or neither in the case of pet-directed or self-directed speech).

(iii) *Determine the LENA classification code characterizing each frame*. Next, a single label derived from LENA segment codes was assigned to each frame, corresponding to the one taking up the greatest temporal extent of the frame (i.e., 50 ms or more).

(iv) *Determine, for each frame, whether the LENA classification code implied a match with the human-derived category*. We computed several different analyses of agreement based on comparisons between human-derived categories implied by the human labels and LENA classification codes for frames; see Table 2. The first analysis addressed agreement about when speech vocalizations were happening and who was talking; it was based on a four-way category distinction: *male adult speech*, *female adult speech*, *child vocalization*, or *other*. The second analysis addressed agreement about whether a frame constituted some kind of speech vocalization or not; it was based on a two-way category distinction: *speech* vs. *non-speech*. Finally, the third analysis addressed agreement about when adult speech was happening or not; it was based on a two-way category distinction: *adult speech* vs. *everything else*. The third analysis was expected to be most pertinent to accuracy of LENA's Adult Word Count, since this measure is based on the frames classified by LENA as adult speech (i.e., as MAN or FAN). Agreement (or error) was quantified as the percentage of frames classified correctly (or incorrectly), given the category implied by human annotation.

Table 2 Types of analyses assessing agreement between classifications by humans and by LENA for a given 100-ms audio frame

Analysis #1: Female Adult vs. Male Adult vs. Child vs. Other				
Label source	Category 1: "Female Adult"	Category 2: "Male Adult"	Category 3: "Child"	Category 4: "Other"
Human	*female adult speech*	*male adult speech*	*child vocalization*	(no label)
LENA	FAN	MAN	CHN, CXN	NON, OLN, TVN, FUZ, and SIL/"Faint"
Analysis #2: Speech vs. Non-speech				
Label source	Category 1: "Speech"		Category 2: "Non-Speech"	
Human	*female adult speech, male adult speech, child vocalization*		(no label)	
LENA	MAN, FAN, CHN, or CXN		NON, OLN, TVN, FUZ, and SIL/"Faint"	
Analysis #3: adult speech vs. everything else				
Label source	Category 1: "Adult Speech"		Category 2: "Everything Else"	
Human	*female adult speech, male adult speech*		*child vocalization* (no label)	
LENA	MAN, FAN		CHN, CXN, NON, OLN, TVN, FUZ, and SIL/"Faint"	

Note. Female adult speech refers to a frame which was marked as speech on the Maternal Caregiver or Other Adult Female tier. *Male adult speech* refers to a frame which was marked as speech on the Paternal Caregiver or Other Adult Male tier

Analyses of adult word count accuracy Two approaches were taken to calculating error in LENA's Adult Word Count. First, we calculated the ratio of total Adult Word Count for sampled audio from each family's file (determined by summing Adult Word Counts in sampled audio from the ADEX file) to the total adult word count identified by humans within sampled audio. This ratio was a measure of the degree of over- or underestimation by the LENA metric for each family, where correlations between these quantities would not have revealed patterns of error as fully. Second, we assigned a fractional signed error in Adult Word Count to each frame. To calculate the fractional signed error, a fractional LENA Adult Word Count was first assigned to each 100-ms frame by identifying the Adult Word Count of the LENA segment(s) that the frame overlapped with, then multiplying by the proportion of the corresponding LENA segment duration that temporally overlapped with the frame. Next, a fractional human word count was analogously determined for each frame; this was calculated by multiplying the human adult word count of the adult speech portion that the frame overlapped with by the proportion of the duration of the speech portion that temporally overlapped with the frame. The fractional signed error for the frame was then calculated by subtracting the fractional human adult word count from the fractional LENA Adult Word Count. This fractional word count error was a dependent variable in statistical analyses testing whether categorical predictor variables (ID vs. AD, female vs. male speech, and correct vs. incorrect classification) associated with frames significantly influenced fractional signed error in Adult Word Count.

Human inter-rater reliability Inter-rater reliability was assessed through re-coding a total of about 3.6 minutes of audio randomly selected from each of ten recordings, which were also randomly selected, including 2.4 min of audio from adult conversational blocks and 1.2 minutes from pause units drawn equally from the beginning and end of the recording. For each 100-ms frame within audio selected for the inter-rater reliability analysis (approximately 21,600 frames in total), the frame's classification by each analyst was determined by assigning each frame to a category (*male adult*, *female adult*, *child*, or *other*) for each coder following the rule described above using the largest portion of the frame's temporal extent. Cohen's kappa (Carletta, 1996; Kuhl et al., 1997) was then used to determine agreement between pairs of codes. Further, a value of kappa was calculated to assess the agreement in labeling ID and AD speech within the subset of frames for which the frame had been classified as adult speech in both the original and reliability coding.

Results

Human inter-rater reliability Our first step was to establish inter-rater reliability for coding by human analysts. Results showed high inter-rater reliability among humans for distinctions of interest. The average κ values indicate very good to outstanding agreement (Breen, Dilley, Kraemer, & Gibson, 2012; Krippendorff, 1980; Landis & Koch, 1977; Rietveld & van Hout, 1993; Syrdal & McGory, 2000). For the four-way classification of frames as adult male, adult female, child vocalization, and anything else (i.e., other), human analysts agreed with mean $\kappa = 0.77$ ($SD = 0.08$). For the speech vs. non-speech distinction, human analysts agreed with mean $\kappa = 0.67$ ($SD = 0.12$). For the adult speech vs. everything else distinction, human analysts agreed with mean $\kappa = 0.81$ ($SD = 0.08$). For adult speech frames, human analysts agreed on whether speech was AD or ID with mean $\kappa = 0.90$ ($SD = 0.18$). Further, accuracy of human word counts showed a strong correlation between the two sets of coded files, $r(8) = .96$, $p < .001$. This consistent across-the-board agreement suggests the robustness of human judgments about when speech was happening/not happening, who was talking, whether the adults were talking to a child or to an adult, and how many words the adult spoke. The remaining analyses used these human judgments as the basis of determinations of LENA's accuracy.

Classification accuracy achieved by LENA for identifying speech vocalizations and attributing these to the correct talkers Throughout the following, italic font is used to indicate a frame's classification as assigned by humans. Table 3 shows counts of frames classified by humans as *female adult speech*, *male adult speech*, *child vocalizations*, or *other* in rows; LENA's classifications of frames are shown across the columns.[2] A chi-square test of independence showed that there was a significant relation between human classification and how LENA classified across the eight categories [$\chi^2(21, N = 440{,}802)$, $p < .001$], suggesting that, as expected, LENA's rates of classification for categories like FAN, MAN, etc. differed as a function of information captured by human classifications (e.g., acoustic properties). Importantly, while there are many on-diagonal entries (i.e., correct classifications), there are many off-diagonal entries (i.e., incorrect classifications). For example, 59% of all *female adult speech* frames were correctly classified by LENA as "FAN", such that, by extension, 41% of *female adult speech* frames were misclassified; it is noteworthy that 12% of these misclassifications were misattributions of a *female adult speech* frame to a child talker (CHN or CXN). By contrast, 57% of all *male adult speech* frames were correctly classified as "MAN", such that, by extension, 43% of *male adult speech* frames were misclassified; however, just 4% of the misclassified *male adult speech* frames were misattributions by LENA to a child talker (CHN or CXN). These observations preview our finding of an interaction between talker gender (male vs. female) and speech style (infant-directed vs. adult-directed), something discussed below.

Table 4 shows LENA's classification accuracy as an overall percentage of frames correctly classified by LENA within each family's recording, averaged across families. Human-identified *female adult speech*, *male adult speech*, *child vocalization*, and *other* frames were classified correctly by LENA at average rates of 59%, 60%, 63%, and 82%, respectively (which corresponded in turn to error rates of 41%, 40%, 37% and 18%, respectively).[3] We also conducted a simple one-sample t test to determine whether, across recordings, classification accuracy was reliably higher than the theoretically defined value of chance, assuming four analysis categories and an unbiased classification method (i.e., ¼ or 25%). LENA's classification was statistically above this value (i.e., 25%) for frames of each of the four classification categories for all 23 families [*adult female*: $t(22) = 15.79$, $p < .001$, *adult male*[4]: $t(21) = 9.25$, $p < .001$; *child vocalizations*: $t(22) = 18.19$, $p < .001$; *other*: $t(22) = 36.24$, $p < .001$].[5] We also used tests of proportions for each family individually to investigate whether LENA's classification accuracy for the four classification categories was significantly above chance (25%) for that family. For one family (*Family 5*), classification accuracy for *male adult speech* was statistically at chance levels ($z = -.64$, $p = .26$); based on this sample, male adult speech for this family was more likely to have been misclassified as female adult speech (27/52 frames) or as child speech

[2] Although the present paper focused on evaluating LENA's accuracy for adult speech measures, our method gave data on LENA's classification accuracy for frames classified by humans as *child speech vocalization* ($N = 66{,}158$). Of frames classified by humans as from the target child ($N = 51{,}334$), LENA classifications were as follows: 3315 FAN (6%), 322 MAN (1%), 15,179 CXN (30%), 18,660 CHN (36%), 280 NON (1%), 4830 OLN (9%), 426 TVN (1%), 3039 FUZ (6%), and 5283 SIL or "faint" (10%). Of frames classified by humans as from another child ($N = 14{,}824$), LENA classifications were as follows: 1246 FAN (8%), 77 MAN (1%), 7020 CXN (47%), 1049 CHN (7%), 75 NON (1%), 2651 OLN (18%), 139 TVN (1%), 839 FUZ (6%), and 1728 SIL or "faint" (12%).

[3] Humans classified frames of target child speech vocalization in all recordings and frames of other child speech vocalization for all but four recordings (for Families 3, 5, 9, and 14). The mean overall percentage of frames correctly classified by LENA within each family's recording, averaged across families, for target child speech vocalization was 39% ($SD = 17\%$) and for other child speech vocalization was 37% ($SD = 27\%$).

[4] One family did not have adult male speech in the selected audio.

[5] We also conducted one-sample t tests using measures of chance based on prevalence of frames from the four categories (Table 3) as classified (i) by humans (*adult female*: 18%, *adult male*: 9%, *child*: 15%, *other*: 59%) and (ii) by LENA (FAN: 16%, MAN: 8%, CHN/CXN: 18%, all other codes: 58%). Results remained reliably higher than chance across families for method (i) [*adult female*: $t(22) = 19.06$, $p < .001$; *adult male*: $t(21) = 13.25$, $p < .001$; *child*: $t(22) = 23.03$, $p < .001$; *other*: $t(22) = 14.67$, $p < .001$], and for method (ii) [FAN: $t(22) = 19.99$, $p < .001$; MAN: $t(21) = 13.71$, $p < .001$; CHN/CXN: $t(22) = 21.58$, $p < .001$; all other codes: $t(22) = 13.31$, $p < .001$].

Table 3 Counts of frames given human analysts' classifications (rows) and LENA classifications (columns). Percentages are given reflecting the number of frames assigned one or more specific labels by LENA, divided by the total number of frames within that human classification category. Percentage values across rows thus sum to 100% (with any deviations due to rounding)

		LENA classifications								
		FAN	MAN	CHN *or* CXN	NON	OLN	TVN	FUZ	SIL or "faint"	Totals
Human classifications	*female adult speech*	**46,011 (59%)**	3951 (5%)	9068 (12%)	249 (0%)	5954 (8%)	1264 (2%)	5747 (7%)	5126 (7%)	77,370
	male adult speech	6770 (18%)	**21,790 (57%)**	1482 (4%)	46 (0%)	2151 (6%)	895 (2%)	2768 (7%)	2182 (6%)	38,084
	child vocalization	4561(7%)	399 (1%)	**41,908 (63%)**	355 (1%)	7481 (11%)	565 (1%)	3878 (6%)	7011 (11%)	66,158
	other	11400 (4%)	8855 (3%)	27715 (11%)	**2603 (1%)**	**18781 (7%)**	**6777 (3%)**	**40,284 (16%)**	**14,2775 (55%)**	25,9190
	Totals	68,742	34,995	80,173	3253	34,367	9501	52,677	15,7094	44,0802

Note. Counts in boldface font were considered correct classifications

(11/52 frames) than to have been correctly classified as male adult speech (just 11/52 frames).[6]

Classifications of "speech": LENA false negative and false positive rates Next, we assessed LENA's accuracy at classifying *speech* and *non-speech* frames as "speech" vs. "non-speech" (cf. Table 2), which gives an index of LENA's ability to detect human vocal activity and is an important step in all automatic speech classification systems (Kaushik, Sangwan, & Hansen, 2018). Figure 2a shows a boxplot for LENA accuracy in classifying *speech* frames across families. Mean accuracy for classifying frames as "speech" was 74%; this corresponded to a false negative rate (i.e., LENA misclassifying *speech* as "non-speech") of 26% (SD = 7%). Classification accuracy for *speech* frames varied widely, from 53% to 86% across families (corresponding to 14% to 47% false negative rates). Table 5 presents error rates across families for classification analyses, and shows that all families had over 10% error rate for false negatives.

Figure 2b shows a boxplot for LENA accuracy in classifying *non-speech* frames across families. Mean accuracy for "non-speech" classifications was 82%; this corresponded to a false positive rate (i.e., LENA misclassifying *non-speech* frames as "speech") of 18% (SD = 8%). Classification accuracy for *non-speech* frames ranged from 64% to 91% (corresponding to a range between 9% to 36% false positives); Table 5 shows a substantial majority (91%) of families had over 10% error rate for false positives.

The lowest bar for evaluating LENA's classification relates to whether it performed better than chance. Classification for both *speech* and *non-speech* frames was better than chance (50%) by a significant statistical margin across families [*speech*: $t(22)$ = 17.531, p <.001; *non-speech*: $t(22)$ = 20.382, p < .001].[7] Tests of proportions were also calculated for each family individually to investigate whether LENA's classification accuracy for *speech* vs. *non-speech* was above chance for that family. Classification rates for *speech* vs. *non-speech* exceeded chance levels (50%) for all families' recordings (α = .05).

Classifications of "adult speech": LENA false negative and false positive rates We next assessed LENA's accuracy at classifying *adult speech* frames (i.e., frames humans identified as an adult female or adult male talker) as "adult speech" (i.e., FAN or MAN); see Table 2. Figure 2c shows a boxplot for LENA accuracy in classifying adult speech frames across

[6] With the lower values of chance from Table 3, 11/52 was significantly higher than 9% for human-identified male frames, z = −3.21, p < .01, and 8% for LENA-identified male (MAN) frames, z = −3.52, p < .001.

[7] We also conducted one-sample t tests based on prevalence of assigning frames into categories (Table 3) as classified (i) by humans (*speech*: 41%, *non-speech*: 59%) and (ii) by LENA (speech: 42%, non-speech: 58%) as measures of evaluating chance. Results remained significant across families for method (i) [*speech*: $t(22)$ = 23.98, p < .001; *non-speech*: $t(22)$ = 14.67, p < .001], and for method (ii) [speech: $t(22)$ = 23.26, p < .001; non-speech: $t(22)$ = 15.31, p < .001] for these values of chance also.

Table 4 Mean classification rates for LENA across families, relative to four-way classification by human analysts

		LENA classifications			
		FAN	MAN	CHN/CXN	Other
Human classifications	Female adult	**59 (10)**	5 (7)	11 (8)	25 (9)
	Male adult	14 (14)	**60 (18)**	4 (5)	22 (11)
	Child	7 (5)	0 (1)	**63 (10)**	30 (11)
	Other	4 (2)	3 (6)	10 (7)	**82 (8)**

Note. The standard deviations across families are given in parentheses. Values in boldface font reflect correct classifications

families. Mean accuracy for *adult speech* frames was 67%, corresponding to a mean false negative rate of 33% (SD = 9%) (i.e., LENA misclassifying *adult speech* frames as "everything else"). Classification accuracy for *adult speech* varied widely across families, ranging from 45% to 82% (corresponding to a range between 18% to 55% false negative rates). Table 5 shows that all families had over 10% error rate for false negatives.

Figure 2d shows a boxplot for LENA accuracy in classifying frames of *everything else* across families. Mean accuracy for classifying *everything else* (i.e., the complement of *adult speech*) was 92%, corresponding to a mean false positive rate of 8% (SD = 6%) (i.e., LENA misclassifying *everything else* frames as "adult speech"). Classification accuracy for *everything else* frames varied from 70% to 97% (corresponding to 3% to 30% false positive rates). Table 5 shows that a substantive minority (17%) of families had over 10% error rates for false positives.

LENA's classification accuracy was significantly better than chance (i.e., 50% for two categories) at classifying both *adult speech* frames [$t(22) = 8.865$, $p < .001$] and *everything else* [$t(22) = 35.808$, $p < .001$].[8] Tests of proportions were calculated for each family individually to investigate whether LENA's classification accuracy for *adult speech* vs. *everything else* was above chance for that family. This analysis revealed that for two families, LENA's machine classifications were significantly *below* chance levels of accuracy for "adult speech" classification with $\alpha = .05$ (*Family 10*: $z = -5.09$; *Family 18*: $z = -5.19$). In both of these cases, intelligible frames of live adult speech were frequently miscoded by LENA as noise or recorded content (including OLN, TVN, and FUZ).

This variability across families in *adult speech* false positive and false negative rates might be less worrisome if there was consistency in LENA's accuracy within a family's recording from one time point to the next. We therefore conducted a statistical test of the null hypothesis that there was consistency in LENA's accuracy levels across our two sampling time points, i.e., no difference in LENA classification accuracy for *adult speech* between samples drawn from the beginning vs. the end of the day. A mixed effects model with a logit linking function was created to predict accuracy across frames (incorrect frames coded as 0, correct coded as 1; incorrect set as baseline) based on the fixed factor of Time with two levels (beginning vs. end, with beginning set as the baseline) and a random intercept for each family. This statistical test showed that the null hypothesis was not supported. Instead, LENA showed systematically *lower* accuracy for frames drawn from the end of the day than frames at the beginning of the day [beginning (baseline): $\beta = 1.90$, $z = 29.02$, $p < .001$, odds ≅ 6.7:1; end vs. beginning, $\beta = -.11$, $z = -12.65$, $p < .001$, odds ≅ 0.9:1]. Thus, not only is there a lack of consistency in classification accuracy/error rates for LENA *across* families' recordings, but there is a lack of consistency in classification accuracy/error rates *within* families' recordings as well. We return to this point in the Discussion.

Effects of talker gender (male, female) and addressee (ID vs. AD) on "adult speech" classification accuracy The remaining analyses focused on *adult speech* frames only. Table 6 shows how the gender of the talker (*male* vs. *female*) as well as the addressee (*ID* vs. *AD* speech) affected patterns of LENA classification for *adult speech* frames[9]. Both "FAN" and "MAN" classifications result in increments to LENA's Adult Word Count estimates, while frames classified in any other way do not. Values in the third data column, which collapses instances which LENA-classified *adult speech* frames as either "FAN" or "MAN'", therefore reflect correct classifications as "adult speech" of some type (even if the talker's gender was misclassified).

The patterns in Table 6 suggest that classification accuracy of *adult speech* may indeed depend on both talker gender and the addressee (ID vs. AD). For instance, for female adult talkers, a higher percentage of frames was accurately classified *in AD* (72%) than *in ID* (64%), with the latter condition involving a lot of misclassifications as a child (14%). Figure 3

[8] We also conducted one-sample t tests drawing on prevalence of frames in each category (Table 3) classified (i) by humans (*adult speech*: 26%, *everything else*: 74%) and (ii) by LENA (adult speech: 24%, everything else: 76%) as measures of evaluating chance. Results remained significant across families for method (i) [*adult speech*: $t(22) = 21.61$, $p < .001$; *everything else*: $t(22) = 15.47$, $p < .001$], and for method (ii) [adult speech: $t(22) = 22.67$, $p < .001$; everything else: $t(22) = 13.77$, $p < .001$] for these values of chance also.

[9] Frames identified as *adult speech* but as directed to individuals other than an adult or child, such as pets or oneself, were excluded (6276 frames, or approximately 5% of *adult speech* frames).

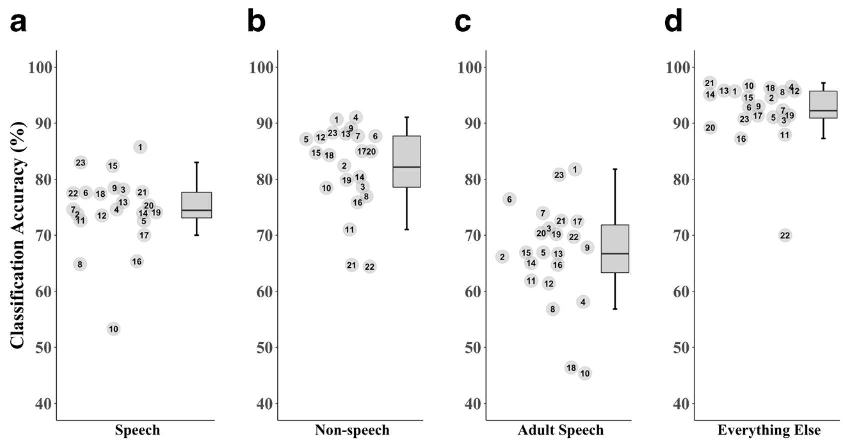

Fig. 2 Box plots showing variability in classification accuracy as a percentage of frames across each family's recording for LENA classifying human-labeled (**a**) speech and (**b**) non-speech or as (**c**) adult speech, and (**d**) everything else (see Table 2 for how these categories are defined). Data from individual families are shown in scatterplots for each classification. Overlaid numbers identify families' recordings across analyses, further illustrating variability

shows rates of correct classification of *adult speech* frames as "adult speech" for each family broken out as a function of Talker Gender (female, male) and Addressee (AD, ID). There is tremendous variability in how accurately *adult speech* was detected across families' recordings, and this accuracy varies as a function of the gender and addressee.

To construct a statistical test of whether there were systematic effects of Talker Gender or Addressee (ID vs. AD) on accuracy of classification of *adult speech* frames, we constructed a mixed effects logistic regression model with a binomially distributed dependent variable of accuracy of classification as "adult speech" (Agresti, 2002; Barr, Levy, Scheepers, & Tily, 2013; Jaeger, 2008; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017; Quené & Van den Bergh, 2008). This statistical approach shows robustness to imbalanced numbers of data points across grouping factors, as well as to missing observations (Th. Gries, 2015). The dependent variable value for each frame was set to 0 if that frame's LENA classification for "adult speech" was incorrect (i.e., if LENA classified the *adult speech* frame as anything other than FAN or MAN), and as 1 if its classification for "adult speech" was correct (i.e., if LENA classified the frame as either FAN or MAN, even if it got the gender wrong). The model (implemented in R; Bates, Mächler, Bolker, & Walker, 2015; R Development Core Team, 2015) included categorical predictor variables of Gender (female vs. male, with female set as the baseline) and Addressee (AD vs. ID, with AD set as the baseline), as well as their interaction, plus a random intercept term for the effect of each family.[10]

As shown in Table 7, statistical modeling revealed statistically significant effects of both Gender and Addressee (and a significant interaction between these) on LENA's ability to classify *adult speech* frames accurately as "adult speech." The significant effect of Addressee ($p < .001$) indicates better classification as "adult speech" for *adult female AD* speech (i.e., the baseline, $M = 68\%$, $SD = 15\%$; odds of correct classification ~2.4:1 [=exp(0.855)]) than for *adult female ID* speech ($M = 63\%$, $SD = 12\%$; odds of correct classification ~1.8:1 [=exp(0.855)*exp(-0.261)]). The significant effect of Gender ($p < .001$) indicates there was better classification for *adult male AD* speech ($M = 76\%$, $SD = 19\%$; odds of correct classification ~2.7:1 [=exp(0.855)*exp(0.129)]) than *adult female AD* speech (odds of 2.4:1 [=exp(0.855)]). Finally, the significant interaction between Gender and Addressee ($p < .001$) indicates that Addressee did not affect equally *adult male* and *adult female* speech. Rather, *adult male* ID speech had an odds of correct classification of about 3.3:1 [=exp(0.855)*exp(0.129)*exp(-0.261)*exp(0.477)], which was significantly more accurate classification than would be expected based on the independent effects of ID addressee on female speech and the effect of being male rather than female. This meant that *adult male* ID speech was classified correctly as "adult speech" *almost twice* as often as expected based on independent effects of being ID and male (3.3/1.8 ≅ 1.83). In summary, accuracy of classifying adult speech frames as "adult speech" was significantly affected by both the gender of the talker, and whether the speech was AD or ID.

Gender classification accuracy: Effects of addressee (ID vs. AD)

To recap, statistical tests revealed significant differences in LENA's classification accuracy for "adult speech" as a function of talker gender and addressee. *Female ID* speech

[10] Following current best practices in statistical modeling, we did not include random slopes in the model, due to the fact that these were not warranted under the naturalistic research design (Barr et al., 2013; Matuschek et al., 2017). This is because not all families had observations for both levels of the two factors, and some families had highly imbalanced data across levels of the factors. Thus, including extra complexity in random factors modeling would have led to less reliable estimation of the main factors of interest.

Table 5 Error rate frequencies across analyses

Error rate	*Speech vs. Non-speech*		*Adult speech vs. Everything Else*		*Adult Word Count*
	False negatives: Frequency (%)	False positives: Frequency (%)	False negatives: Frequency (%)	False positives: Frequency (%)	|LENA – Human| Frequency (%)
>5%	23/23 (100%)	23/23 (100%)	23/23 (100%)	14/23 (61%)	22/23 (96%)
>10%	23/23 (100%)	21/23 (91%)	23/23 (100%)	4/23 (17%)	22/23 (96%)
>20%	20/23 (87%)	8/23 (35%)	21/23 (91%)	1/23 (4%)	16/23 (70%)
>30%	3/23 (13%)	2/23 (9%)	14/23 (61%)	1/23 (4%)	13/23 (57%)
>40%	1/23 (4%)	0/23 (0%)	4/23 (17%)	0/23 (0%)	9/23 (30%)

Note. Frequencies and percentages reflect the number of families (out of 23) that had classification error rates greater than the error rate on each row. Adult word counts reflect absolute percent overestimation or underestimation by LENA

produced the worst classification performance, while *male ID* speech produced the best classification performance. Yet, these conditions were associated with notable error patterns (cf. Table 6); for instance, frames in the *female ID* condition were disproportionately misclassified as a child. Misclassifications as a child were far less common in the other three conditions. The *male ID* condition further showed an apparently disproportionate misclassification of the *gender* of the talker, and the *female AD* speech condition was also associated with a large number of gender misclassifications. Given these error patterns, we further investigated LENA's accuracy in classifying talker gender. Figure 4 depicts LENA's accuracy, for frames of *adult speech*, at correctly classifying the *gender* of an adult talker, broken out by the talker's human-identified gender (male vs. female) and the addressee condition (ID vs. AD).

Rigorous statistical testing bears out what is apparent in the figure, i.e., differential error in LENA's classification of the gender of an adult talker and addressee condition. The statistical analysis was done on the subset of *adult speech* frames which were correctly classified by LENA as "adult speech" (i.e., FAN or MAN). We constructed a mixed effects logistic regression model with a categorical, binomially distributed dependent variable in which, for each human-identified *adult speech* frame which LENA had classified as adult speech (FAN or MAN), the dependent variable value was coded as 1 if LENA correctly classified the gender as the same that humans had identified, and as 0 otherwise. Our model also included categorical predictor variables of (human-identified) talker gender (with female set as the baseline) and addressee (ID vs. AD; with AD set as the baseline). A random intercept term for the effect of each family was also included.

As shown in Table 8, statistical modeling revealed that gender classification for *adult speech* frames was significantly affected both by Gender and Addressee, and by a significant interaction between these. The significant effect of Addressee ($p < .001$) suggested that classifying gender for *adult female ID* speech was *eight times* better (with odds of correct classification of ~53:1 [=exp(1.851)*exp(2.126)]) than for *adult female AD* speech (with odds of ~6:1 [=exp(1.851)]). Further, the significant effect of Gender ($p < .001$) suggested that classification of gender for *adult male AD* speech was *four times* better (odds of ~27:1 [=exp(1.851)*exp(1.438)]) than for *adult female AD* speech (odds of ~6:1 [=exp(0.855)]). Finally, the significant interaction between Gender and Addressee (p < .001) meant that Addressee did not affect the relative accuracy of gender classification equally for *adult male* and *adult female* speech. Rather, *adult male* ID speech had an odds of correct gender classification of about 3.6:1 [=exp(1.851)*exp(1.438)*exp(2.126)*exp(-4.124)]. As such, LENA classified gender for adult male ID speech more poorly than any other condition; the odds of correct gender classification for *adult female* AD speech being two times higher; for *adult male* AD speech, seven times higher; and for *adult female* ID speech, *14 times* higher, than accuracy of gender classification in the adult male ID speech condition.

Accuracy in LENA's Adult Word Count measure Figure 5 presents the percent over- or underestimation of LENA's Adult Word Count compared to the word count from human listeners for sampled audio. A value of 100% means LENA's Adult Word Count perfectly agrees with human word count. The mean percent of overestimation for LENA Adult Word Count was $M = 147\%$, indicating an average 47% overestimation in word count by LENA relative to human word counts. The median overestimation was 31% (the difference between the mean and the median is largely driven by three families – excluding these families generated a mean = 29% overestimation, a value in line with the median). LENA word counts ranged from 83% to 310% of human word counts (SD = 56%). Table 5 shows that 22/23 families had greater than 10% difference between LENA's Adult Word Count and human word counts (either over- or underestimation). Surprisingly, for 7/23 (30%) of the families, the overestimation was greater than 50%. Nevertheless, in keeping with prior findings, LENA Adult Word Count and human adult word

Table 6 LENA classifications of human-identified adult speech frames, shown separately as a function of talker gender (female or male) and type of addressee (ID vs. AD). Top number represents frame counts. Shown in parentheses is the percentage of frames for that row, out of the row total. The third column of LENA classifications (shown in boldface font) represents the sum of the first two columns and gives the count and percentage of frames classified as adult speech, collapsing across gender. See text

Human classifications		LENA classifications: FAN	MAN	FAN *or* MAN	CHN *or* CXN	NON	OLN	TVN	FUZ	SIL *or* faint	Totals
AD	Female adult speech	7280 (54%)	2391 (18%)	**9671 (72%)**	448 (3%)	43 (<1%)	1592 (12%)	266 (2%)	906 (7%)	589 (4%)	13,515
	Male adult speech	259 (4%)	4138 (70%)	**4397 (73%)**	59 (1%)	0 (0%)	337 (6%)	193 (3%)	482 (8%)	529 (9%)	5997
	Totals	7539	6529	14,068	507	43	1929	459	1388	1118	19,512
ID	Female adult speech	36,463 (62%)	1405 (2%)	**37,868 (64%)**	8162 (14%)	175 (<1%)	3978 (7%)	846 (1%)	4216 (7%)	3942 (7%)	59,187
	Male adult speech	6241 (20%)	16,892 (56%)	**23,133 (76%)**	1264 (4%)	32 (<1%)	1642 (5%)	557 (2%)	1960 (6%)	1868 (6%)	30,456
	Totals	42,704	18,297	61,001	9426	207	5620	1403	6176	5810	89,643

Note. Percentages of frames are rounded to the nearest percent. Values in boldface reflect LENA's correctly classifying a human-identified adult speech frame as adult speech, regardless of how it classified

counts were correlated with one another [$r(21) = .86$, $p < .001$][11], meaning that both the human count and LENA count tended to rise together, in spite of the overestimation by LENA. Similarly, the ranking of participants based on LENA's Adult Word Count and human word counts showed a significant correlation ($rs(21) = .46$, $p = .029$), suggesting that the ranked order of word counts from humans and LENA were somewhat consistent despite the large and variable errors we observed.

Relationship between classification accuracy and Adult Word Count Given that the Adult Word Count is preceded by, and depends on, the classification step, we expected that accuracy for classifying frames as "adult speech" would significantly influence the accuracy of LENA's Adult Word Count. However, no prior published study has tested or shown such a dependency. To test this, we constructed a generalized linear model in R (using *glm*) to test the extent to which, across families, the percentage of correct classifications of *adult speech* and *everything else* frames (or their interaction) predicted the percentage of over- or underestimation for the LENA Adult Word Count (see above). All variables were scaled and centered. Table 9 shows the results of this statistical modeling. Accuracy of classification of *everything else* frames significantly predicted Adult Word Count accuracy, with a large effect size ($r = -0.77$).[12] There were no other significant effects; we return to this point in the Discussion.[13]

Given this finding relating overall *everything else* classification accuracy to overall Adult Word Count error, we sought to identify how classification accuracy interacted with the additional factors of gender and addressee on a frame by frame

[11] As pointed out in the Introduction, correlations are not optimal tools for comparing methods. However, the correlation is provided for comparison with values from prior LENA reliability studies (see Table 1).

[12] A plot of *everything else* classification accuracy against Adult Word Count classification accuracy suggested that *Family 22* was something of an outlier. To test whether Family 22 was driving significance for the generalized linear model reported in Table 10, we re-ran the model but removing *Family 22*. The results were similar. The statistically significant effect of *everything else* classification accuracy on LENA Adult Word Count accuracy persisted (β estimate = −0.523, st. error = 0.20, $t = -2.61$, $p = .018$), with no other significant effect or interaction, as before. Further, the effect size for the relationship between *everything else* classification accuracy to Adult Word Count accuracy remained strong ($r = 0.58$). These results support the robustness of the statistical relationship between *everything else* classification accuracy and Adult Word Count accuracy and suggest the results are not due to an outlier.

[13] The architecture of LENA's algorithms for Adult Word Count calculations entail that Adult Word Count is only incremented when stretches of audio are classified as "adult speech", as opposed to any kind of "speech" in general. Consistent with this, a generalized linear model was constructed for LENA Adult Word Count accuracy with predictor variables of accuracy of speech and non-speech classification (and their interaction); neither variable, nor the interaction, showed a significant effect (all p's > 0.58). This additional modeling underscores the dependency of LENA's Adult Word Count classification accuracy on "adult speech" classification decisions per se, rather than *all* speech (or speech-like) vocalization decisions.

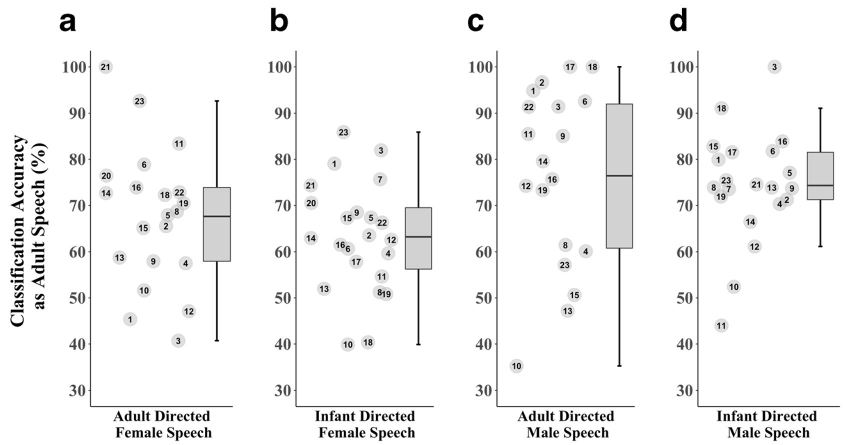

Fig. 3 Effects of Talker Gender (female, male) and Addressee (AD, ID) on accuracy in classification of adult speech as adult speech. Boxplots and associated scatter plots highlight mean accuracy and variability across families (indicated by numbers in the scatterplot)

basis. Therefore, a generalized linear mixed effect regression model was constructed to predict the continuous dependent variable of signed per-frame Adult Word Count error. The model included categorical predictor variables for each *adult speech* frame consisting of Talker Gender (with female as the baseline), Addressee (ID vs. AD, with AD as baseline), Classification Accuracy (incorrect vs. correct, with incorrect as baseline), and all possible interactions (see Method). The model included a random intercept-only effect term to account for clustering by family. This model was reduced through iterative elimination of nonsignificant interaction terms starting with the three-way interaction until a likelihood ratio test revealed that the next simpler model was a significantly worse fit, following current best practices for model fitting (Gries, 2016), including an assumption of convergence of the *t* and *z* distributions.

The final model (Table 10) showed that Adult Word Count accuracy was significantly affected by Classification Accuracy, which had a large effect on the amount of per-frame signed error; there were also smaller, but still significant, effects of Talker Gender and Addressee, and significant interactions between Talker Gender and Addressee and between Classification Accuracy and Gender. A value of "0" for per-frame signed error would indicate perfect agreement in proportional word counts by humans and LENA. First, incorrectly classified AD frames engendered more negative signed error ($b_{\text{female}} = -0.3$, $b_{\text{male}} = -0.28$), that is, a greater deviation in the direction of under-counting, than incorrectly classified ID frames ($b_{\text{female}} = -0.23$, $b_{\text{male}} = -0.24$). Moreover, correctly classified frames engendered positive signed error, i.e., over-counting, of a magnitude that depended on the Talker Gender and Addressee. Correctly classified AD frames engendered positive signed error ($b_{\text{female}} = +0.06$, $b_{\text{male}} = +0.05$) which was nevertheless smaller in magnitude than the error of ID frames ($b_{\text{female}} = +0.13$, $b_{\text{male}} = +0.09$).

Taken together, these results reveal that LENA showed systematically more error in detecting and correctly classifying speech of adult females than speech of adult males. Even under conditions when LENA had accurately classified frames of adult talkers as "adult speech", LENA was less accurate in registering and counting words of adult females than in counting words of adult males, showing systematically greater undercounting of words of adult females than words of adult males. Finally, there were significantly higher error rates for the LENA Adult Word Count when adult females were directing their utterances to children (i.e., ID condition), compared with any other condition.

Table 7 Statistical model of effects of *Addressee* and speaker *Gender* on accuracy of classification of adult speech frames as "adult speech" (i.e., FAN or MAN)

	β Estimate	St. Error	z	Pr(>|z|)
(intercept)	0.855	0.084	10.202	< .001**
Addressee	−0.261	0.024	−10.828	< .001**
Gender	0.129	0.035	3.662	< .001**
Addressee:Gender	0.477	0.041	11.695	< .001**

Note. ** indicates statistical significance at $\alpha = .001$

Discussion

This study presented an independent assessment of reliability in classification and Adult Word Count from LENA at-home recordings. Independent assessment (i.e., analyses not funded by the LENA Foundation) is a requisite for clinicians and researchers to use this tool with confidence. The current analysis focused on accuracy of audio classification by LENA, accuracy of LENA's Adult Word Count metric, and the implications of classification errors on Adult Word Count estimates. Our focus on these metrics was due to the developmental importance of quantity and quality of environmental speech and the importance of child-directed speech for language development (e.g., Hoff & Naigles, 2002;

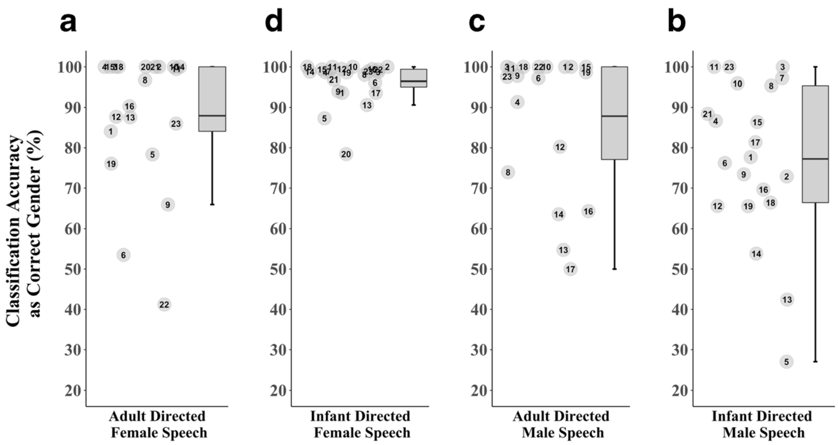

Fig. 4 Effects of Talker Gender (Female, Male) and Addressee (AD, ID) on accuracy of gender classification within the subset of human-identified adult speech frames correctly classified as adult speech by LENA. Boxplots and associated scatter plots highlight mean accuracy and variability across families (indicated by numbers in the scatterplot)

Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991; Shneidman et al., 2013). LENA's automatic analysis of Adult Word Count has become widely used to assess the quantity and quality (assessed through addressee) of speech in children's environments (Romeo et al., 2018; Weisleder & Fernald, 2013). Given the shift towards its use, we sought to provide an independent evaluation of LENA by having human analysts (i) identify when a man, woman, or child produced a speech vocalization; (ii) indicate, for adult talkers, whether the utterance was child-directed or adult-directed; and (iii) count the number of intelligible adult words. We therefore posed the question: What amount of error is "acceptable," for both research and clinical purposes, for ensuring standards of validity and reliability in order to justify reliance on automatic, machine-based decisions about the amount of language input in a child's environment?

LENA showed variable – and in some cases quite large – errors in classifying audio as the correct talker (man, woman, or child). The average false negative rate for *adult speech* frames was 33% (range: 18–55% missed frames). For all 23 families in our sample, LENA was in error on more than 10% of intelligible *adult speech* frames. Classification was highest (92% accuracy) for audio that did not contain adult speech (*everything else*). In contrast, human-identified adult female speech was correctly identified by LENA as female speech only 59% of the time.

Table 8 Statistical model of effects of *Addressee* and speaker *Gender* on accuracy of classification of adult speech frames according to the correct gender (i.e., adult female as "FAN" and adult male as "MAN")

	Estimate	St. Error	*z*	Pr(>|z|)
(intercept)	1.851	0.234	7.911	< .001 **
Addressee	2.126	0.070	20.626	< .001 **
Gender	1.438	0.046	46.053	< .001 **
Addressee:Gender	−4.124	0.083	−49.701	< .001 **

Note. ** indicates statistical significance at α = .001.

Further, both human-identified gender (male vs. female) and addressee (ID vs. AD) significantly affected the accuracy of LENA's audio classification. LENA was overall statistically better at classifying frames of adult speech as "adult speech" for male voices compared with female voices. LENA showed especially high error at classifying adult female speech in ID condition, in which LENA disproportionately classified frames as a child talker (and thus not as adult speech). Even when LENA accurately identifies audio as adult speech, gender and addressee still affect accurate classification of talker gender. Within correctly classified *adult speech* frames we found that accuracy for ID speech was high for women but low for men, whereas AD speech accuracy was consistent across genders. Thus, even in cases when LENA accurately identifies the *amount* of adult speech, variability due to addressee and gender may lead to attribution of adult word count to the incorrect gender. Given typical goals of many research studies and clinical situations to attribute input to the correct talker or talker group, ascribing that input to the wrong talker(s) could lead to incorrect conclusions by researchers or clinicians about who was providing input to the child, and how much input they provided. As one example, if words produced by a male caregiver (say, a single father) were misattributed to a female caregiver, this could lead a clinician to doubt the father's reports about how much input he was providing to a child, where this could potentially further lead to needless, ineffectual interventions aiming to increase the amount of input the father provided to the child.

We also showed that these systematic classification errors significantly impacted the accuracy of LENA's Adult Word Count. On average, LENA overestimated Adult Word Counts by 47% (ranging from undercounting words by 17% to overcounting words by 208%). The correlation observed between human word counts and LENA's Adult Word Count (r = .86)

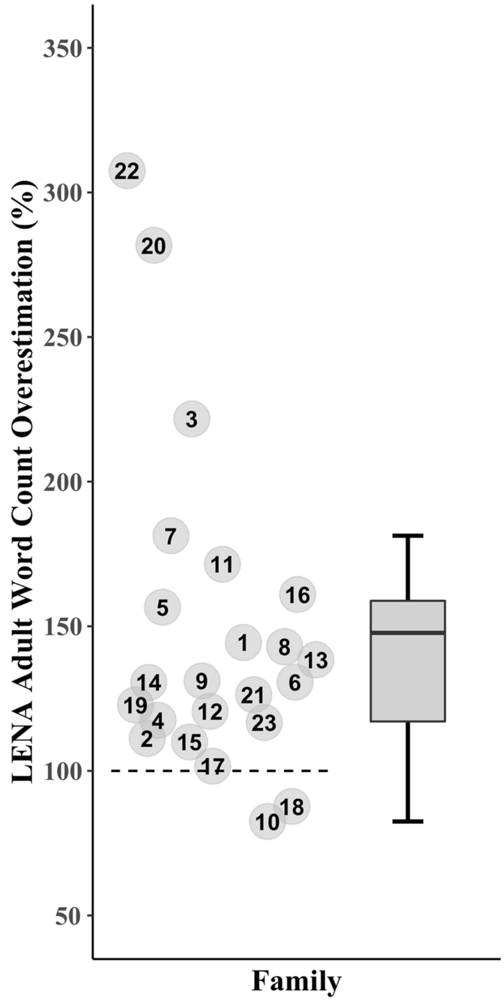

Fig. 5 Box plots showing variability in error between LENA Adult Word Count and human adult word count. Values represent the percent of over or underestimation by LENA (LENA/Human) such that the dashed line at 100% represented perfect agreement between LENA and human word counts. Values below this line represent underestimation and values above represent overestimation

was well within the range of values reported for prior studies (see Table 1), and the Spearman correlation ($rs = .46$) showed similar rankings between LENA and human coders. However, the variability in error for Adult Word Count estimates we identify are concerning, and these significant correlations obscure the problematic overestimation by LENA we observed, highlighting the inadequacy of correlations for assessing reliability.

This is the first paper to have identified the speaker gender and intended addressee as variables that directly affect accuracy of segment classification and Adult Word Count. The fact that our study included both of these dependent measures, while also measuring the false negative ("miss") rate, allowed us to evaluate how the these variables were related to one another, thus going substantially beyond prior work which has examined only subsets of these variables (e.g., Bergelson, Casillas, Soderstrom, Seidl, Warlaumont, Amatuni, 2019; VanDam & Silbert, 2016). Gender and addressee both interacted with classification accuracy to predict word count error. Interestingly, the relative amount of error across ID vs. AD conditions depended on whether frames had been correctly classified as adult speech. In particular, when human identified adult speech frames were missed by LENA, the Adult Word Count showed greater error (i.e., more undercounting) when frames were AD compared with when they were ID. However, when adult speech frames were correctly classified, the Adult Word Count showed greater error (i.e., more overcounting) when frames were ID compared with when they were AD. Further, frames of male adult speech generated significantly less error in Adult Word Count than frames of female adult speech for three out of four conditions; only inaccurately classified ID frames showed less error for female than male speech. The patterns we identified suggest that LENA misattributes or misses adult words as a function of the talker's gender and speech style in part due to systematic errors in classification, and this is especially problematic for ID speech from adult female speakers.

Therefore, a main finding was that adult females talking in ID register were particularly likely to have their speech "missed" (i.e., LENA failed to detect it) for purposes of Adult Word Count; such speech was disproportionately attributed to children. LENA very rarely misattributed the gender of female adult talkers who were addressing children (ID speech). In other words, when female ID speech was accurately identified to be from an adult (as opposed to mistakenly

Table 9 Statistical model of effects of adult speech and everything else classification accuracy on LENA Adult Word Count accuracy

	β Estimate	St. Error	t	Pr(>|t|)
(intercept)	0.043	0.169	0.268	0.79174
adult speech	0.029	0.234	0.125	0.90181
everything else	−0.781	0.168	−4.662	< .001**
adult speech: everything else	0.188	0.350	0.539	0.59618

Note. ** indicates statistical significance at $\alpha = .001$

Table 10 Statistical model of effects of Addressee, Gender and Classification Accuracy on fractional Adult Word Count Error per frame

	β Estimate	St. Error	t	Pr(>|t|)
(intercept)	−0.302	0.009	−32.07	< .001**
Addressee	0.070	0.002	34.12	< .001**
Gender	0.022	0.003	6.13	< .001**
Accuracy	0.363	0.001	245.98	< .001**
Addressee:Gender	−0.028	0.003	−8.03	< .001**
Gender:Accuracy	−0.028	0.003	−10.42	< .001**

Note. ** indicates statistical significance at $\alpha = .001$

attributed to a child), this adult speech was assigned to the correct gender ("female") with high accuracy. Adult male speech showed a generally opposite pattern – better detection accuracy but worse gender classification. That is, adult male speech was much more readily detected as "adult speech" (and tended to be more faithfully reflected in Adult Word Counts), but gender classification was quite poor, with male ID speech was misattributed to females 14 times more often than the reverse (female adult ID speech being attributed to a male adult).

Across all results of classification and Adult Word Count accuracy, we see striking variability between families. Some of the variability across families in the accuracy with which adults' speech was classified as "adult speech" depended upon the gender and addressee of the speaker. Speakers of a given gender differ in their typical fundamental frequency ranges; for instance, the distribution of mean F_0 values for adult female speakers – even in AD register – ranges from statistically quite low and overlapping with higher-pitched males, to statistically quite high and overlapping with the typical F_0 values of children (Hanson, 1997; Hanson & Chuang, 1999; Iseli, Shue, & Alwan, 2006). Classification of speech given this variability is further complicated by variable usage of ID and/or AD registers between speakers. Given prior research suggesting a dependency of LENA's classification accuracy on F_0 (VanDam & Silbert, 2016), we speculate that female talkers who naturally have lower F_0 may have produced speech which was better detected than female talkers with higher F_0.

Varying degrees of competing environmental noise sources presumably also account for some of the variability in classification error, although this is speculative. Classification errors where TV or young siblings are misclassified as adult speech could significantly alter Adult Word Count accuracy, a concern that may underlie our finding that the rate of correct classification of *everything else* frames in the "adult speech" analysis significantly predicted Adult Word Count error. In keeping with this idea, we observed that for two of the families with more than 100% overestimation in LENA's Adult Word Count – relative to the human word count – the error seemed to have been driven by misclassification of TV, while in the third case, it appeared to be due to misclassification of sibling speech as adult speech. Overestimation by LENA has been observed in prior studies due to TV (Xu, Yapanel, et al., 2009c), and during activities in the home (Burgess et al., 2013; or in Table 2 of Soderstrom & Wittebolle, 2013). Additional work will be needed to determine if LENA generates accurate language estimates in particular auditory environments (e.g., quiet settings with little reverberation).

It may be instructive to compare the present LENA results to current industry standards for speech and/or voice activity detection. Speech/voice activity detection (S/VAD) algorithms classify speech segments, discriminating them from non-speech ones. S/VAD is considered an essential preprocessing step for automatic speech recognition, but achieving robust accuracy is challenging in naturalistic settings, due to the variety of undesirable variables such as additive noise and channel distortions (Hansen & Hasan, 2015). Recently, benchmarking studies have quantified performance of algorithms for speech/voice activity detection using a variety of corpora of noisy, naturalistic recordings under comparable conditions. In order for such studies to facilitate comparisons across algorithms, performance has sometimes been evaluated with reference to both the probability of false alarms, P_{fa}, as well as the probability of misses, P_{miss}, combining both of these into a single weighted *detection cost function* (DCF) equal to $a{*}P_{fa} + b{*}P_{miss}$, using weighting factors a and b, such that $a + b = 1$. Clearly, lower DCF indicates better classification performance. Two recent studies found DCF values of 7.4% (Sharma, Das, & Li, 2019) and 11.7% (Hansen, Joglekar, Shekhar, Kothapally, Yu, Kaushik, & Sangwan, 2019) (using $a = 0.25$ and $b = 0.75$) in S/VAD for recordings from the Apollo 11 space travel mission (Kaushik et al., 2018). Another study (Dubey, Sangwan, & Hansen, 2018) evaluated algorithms using a corpus of noisy recordings from degraded military communication channels and reported DCF values (also with $a = 0.25$, $b = 0.75$) ranging from 4.3% to 8.9% (mean = 6.1%) across five novel algorithms (averaging across degraded channel conditions), where this constituted comparable performance in relation to baseline algorithms (e.g., Sholokhov, Sahidullah, & Kinnunen, 2018). This study further reported DCF values (here with $a = 0.5$, $b = 0.5$) of 2.0% to 7.2% (mean = 3.5%) for a corpus of naturalistic recordings of students interacting in peer-led team learning, substantially outperforming the baseline algorithms, each of which had DCF of 28% or higher.

We compared the results of such benchmarking studies to the present study of LENA S/VAD performance, recognizing caveats associated with the distinct corpora on which evaluations were carried out. To achieve a rough comparison, we used rates of P_{fa} and P_{miss} for each recording for speech vs. non-speech classification to calculate a mean DCF in VAD across all 23 recordings from the present study. (Note that we could not find any studies which evaluated algorithms in a manner comparable to our adult speech vs. everything else binary classification.) We calculated DCF values for LENA performance which ranged from 22% to 24% (for $a = 0.5$ vs. $a = 0.25$, respectively). Although any comparison must be predicated on the fact that different corpora were used in prior studies in comparison to the current one, to the extent that such a comparison may be valid, the present results suggest that LENA's performance on S/VAD is well below current industry benchmark standards.

There were several limitations of our study. First, sampled audio came specifically from the beginning and end of the day-long recordings; this could be considered a strength and/or a weakness. These times were chosen to provide a fairer test of LENA's measurement of the home environment because family members were likely to be at home engaging with the child in routine activities. In addition to providing a sample that we thought would provide consistency across families, it allowed us to compare accuracy across multiple times of the day. It also allowed coders to have context necessary for identifying addressee. Our samples also included audio judged by LENA to have adult speech, plus random samples of portions judged by LENA to have no near-field adult or child vocalizations, allowing estimation of false negative rates. Given the sampling approach, results from our sampling method are representative of LENA's performance early and late in children's days. The activities and genders of the speakers in these samples may not be representative of the entire day – for example, there may be more male speech in the selected samples. We did not randomly sample from the entire recording – e.g., times when the children might have been in noisy daycare environments, or in cars on the freeway. It is unclear whether such sampling would yield worse or better accuracy estimates. It should be noted that other studies – including the well-cited study by Xu et al. 2009 – used non-random sampling methods.

Another limitation is that our coding system was designed to identify only adult speech and child speech (or speech-like vocalizations), rather than any other kinds of audio sources. While the coding system permitted us to efficiently assess specifically what we cared about – LENA's accuracy at identifying speech vocalizations and Adult Word Count – it nevertheless left us unable to assess other reasons why LENA may have missed speech vocalizations, or incorrectly classified audio as speech vocalizations when it was not. The extent to which some classifications decisions show ceiling effects while others show extensive variability across families demonstrates the strength of our coding system; however, we cannot determine whether TV might have been a frequent source of error for LENA.

Moreover, the fact that we included children with a variety of hearing statuses is both a strength and a limitation. Assessing available families' recordings regardless of hearing status was undertaken as a specific targeted goal of our study, due to our need to be able to generalize LENA's accuracy across our heterogeneous population with a variety of hearing statuses. We therefore viewed this as a strength, because the results were not dependent on any particular hearing status; however, our study was not designed to assess the effects of hearing status, which would have involved an entirely different design (e.g., matching groups on potentially extraneous variables, and larger samples for each hearing status).

Finally, due to our overarching research interest in variability in language environment provided by adults, our study was designed to analyze adult speech classification and Adult Word Count accuracy. It was not designed to analyze other LENA metrics such as conversational turns or child vocalization counts. Nevertheless, our findings that females talking in ID speech register were often misclassified as children, where this misclassification happened significantly more often than when females were talking in an AD speech register should give users of child vocalization LENA metrics pause. This is because LENA is often used as a device for estimating the amount of language input to children, which is often produced by female adult talkers who are talking to these children in ID register; hence, our results suggest that LENA will disproportionately and significantly undercount speech of a type – ID speech by adult females – for which it is frequently deployed to measure. Given the wide variability in performance across families in our sample, applying a uniform correction factor to all families to offset undercounting would not be a feasible way of achieving enhanced accuracy at the level of individual families' recordings. The results from this study also provide reason for concern about LENA's conversational turn counts, given that such counts rely on prior steps involving classification of audio into vocalization events by the mother and child. Given that mothers can often be expected to use an ID register, our LENA results imply systematic error in detecting mothers' speech turns, suggesting consequently that LENA may significantly misrepresent counts of turns, a topic we are investigating in ongoing studies.

Despite these limitations, the findings reported here raise concerns for research which may seek to make theoretical claims based on observations of individual differences from LENA estimates of word counts across recordings in a corpus. First, our findings of systematic error in audio classification for adult female speech in ID register raises concerns for research relying on LENA-defined segmentation to select audio for analysis (e.g., Ko et al., 2016; Seidl et al., 2018), especially if infant-directed female speech is of interest. Furthermore, our findings that Adult Word Count is of variable accuracy across families should give pause to a class of studies that rely on raw reports of Adult Word Count to make conclusions (e.g., Irvin, Hume, Boyd, McBee, & Odom, 2013; Marchman et al., 2017; Sacks et al., 2014; VanDam et al., 2012). Our finding that this metric was of variable accuracy across families suggests that individual differences in the Adult Word Count metric in such studies may reflect the actual speech environment or may represent measurement error between and among families. Our study further raises validity concerns about various approaches used in prior studies to identify samples for research comparison where sample identification is predicated on LENA automated outputs. For instance, approaches involving finding portions of audio across recordings that are as similar as possible on the basis of LENA outputs – such as identifying the hour with the most recorded vocal interaction as in Romeo et al. (2018) – are a step in the right direction but entail an assumption of equivalent accuracy across recordings in identifying speech vocalization events which the present study did not support. Similarly, using LENA to identify samples likely to contain speech and then transcribing those samples (Garcia-Sierra, Ramírez-Esparza, & Kuhl, 2016; Oller et al., 2010; Ramírez-Esparza et al.,

2014) may be a well-supported approach if the research goal does not entail comparison of quantities of speech from LENA outputs, but instead merely on identification of who spoke and what was spoken through use of human transcription, which is considered the “gold standard” for identifying talkers and words. Our finding that classification and Adult Word Count accuracy in each frame are significantly affected by gender and addressee (with an ID or AD register) further presents theoretical concerns for research making claims specifically about infant-directed speech or about the roles of male or female caregivers based solely on LENA derived metrics given that we found different degrees of measurement error for LENA across men vs. women using ID vs. AD. For example, a methodological approach in a recent study (Weisleder & Fernald, 2013) involved classifying 5-minute portions of LENA recordings as infant-directed or adult-directed. The LENA-generated Adult Word Count within each portion was then binned as either ID or AD. Our results suggest that such binning is problematic given systematic differences in Adult Word Count error rates for ID versus AD speech. Such differences could not only create spurious relationships, but also obscure real relationships. The issues highlighted here are theoretically problematic for researchers making claims about individual differences between children in the LENA Adult Word Count metric, especially when the size of those differences is within the range of LENA’s measurement error.

Ultimately, we recommend that any study considering reliance on LENA automated outputs for either selection of samples and/or quantification of speech and/or words should employ a rigorous evaluation step aimed at quantifying LENA’s accuracy within their particular corpora. Such an evaluation step should necessarily assess any LENA outputs which might affect sample selection and/or speech quantification and will likely include assessments of classification accuracy, including both false positive and false negative rates (preferably via random sampling from recordings to include portions of both conversational block and “pause” units), not to mention accuracy of counts of intelligible words. Prior to the use of LENA automated outputs, it will be essential for researchers to identify levels of acceptable error (e.g., false positives or negatives) in their dependent measures in order to minimize validity threats to study conclusions. In the future, such evaluations may be assisted through use of the methods for human coding developed in the current study.[14]

Further, consideration of study sample sizes emerges as a central concern for planning how future studies might use LENA. Our results suggest that, *on average*, LENA’s automated output measures may often be statistically be correlated with actual values. However, when LENA measures are to be used as proxies for other metrics (e.g., estimates of numbers of words perceived by humans), then uncovering statistical relationships among variables of interest in the face of systematic and random errors in LENA outputs, such as those uncovered here, will depend on study sample sizes in relation to individuals or groups under investigation. Given that, in any case, LENA has value as a wearable child-oriented digital recording tool capable of capturing a full day of audio, researchers are empowered to make decisions tailored to their individual study designs about which software and/or algorithms are best applied to any recordings that are acquired. It is hoped that future studies may identify recording settings for deploying LENA and/or best practices for its use that will assure high accuracy for its automated outputs.

Clinicians should also be aware of the implications of our results. If the quantification of a speech environment LENA reports is inaccurate, then clinical guidance will correspondingly be misguided. This is especially concerning given the widespread use of LENA as a clinical assessment tool, such as in the Providence Talks city-wide language exposure intervention for at-risk children (Talbot, 2015; Wong et al., 2018). Clinical intervention requires working with individual families to determine what speech the child is hearing (i.e., Pae et al., 2016; Suskind, Graf, et al., 2016a; Suskind, Leffel, et al., 2016b; Zhang et al., 2015). Given the variability across families observed in the present study, the guidance clinicians offer may be distorted based on factors such as the register that mothers use. That the inaccuracies are unpredictable a priori across families compounds the problematic clinical implications of our present findings. A worst-case scenario suggested from our results is that undercounting in ID female speech could lead female caretakers to appear to clinicians to provide less speech than they actually do provide. However, shifting to less pronounced (more AD-like) speech would lead to female speakers getting more credit from clinicians despite the speech being less helpful for children learning language. These concerns – and identifying solutions to address them – should be a priority for anyone focused on clinical interventions.

Overall, these findings suggest that relying solely on LENA’s Adult Word Count to infer who is talking, and how much they are talking, is not a best practice for either clinical use or research. These findings cast doubt on the value of LENA-generated metrics as a basis of clinical recommendations for individual families or for use in individual-differences research – where these data on LENA’s unreliability have prompted our research team’s return to hand-coding of child language environments. LENA’s accuracy varies greatly from family to family, or from one time to another. Much adult speech that is intelligible to humans is missed by LENA, especially female infant-directed speech. If the goal is to use a very large set of recordings to identify general trends from correlations, then LENA may be a reasonable tool for this purpose. However, these data provide evidence relying

[14] The coding manual and raw data files for the current project are available at https://osf.io/2dz4y/.

solely on LENA's Adult Word Count to infer the amount of language spoken by caregivers in children's home environments is not a best practice, since doing so may lead to invalid clinical judgments and/or research conclusions.

Acknowledgements We gratefully acknowledge the support of NIH grant R01 DC008581 to D. Houston and L. Dilley. The researchers would like to acknowledge help of Jessica Reed and Yuanyuan Wang for their help in data collection and Somnath Roy for assistance with analyses. We would also like to thank James Chen, Elizabeth Remy, Josh Zhao, Chitra Lakshumanan, Courtney Cameron, Sophia Stevens, Nikaela Losievski, Riley Reed, Kayli Silverstein, and Kelsey Dods for their diligent work coding audio. Thanks to Melanie Soderstrom for sharing previous analysis of LENA reliability with us and for many useful discussions.

Open practices statement This study was not formally preregistered. The data files and coding manual have been made available on a permanent third-party archive at https://osf.io/2dz4y/.

Appendix

Table 11 Demographic characteristics of children in participating families; CI = cochlear implant, HA = hearing aid, and NH = normal hearing. For Family 10, the recording was made prior to the child's device fitting

Family	Chronological age at recording (years;months)	Gender	Hearing status	Time with hearing device (years;months)
1	2;8	F	CI	1;11
2	2;7	M	CI	0;10
3	2;8	M	CI	1;4
4	1;8	M	CI	1;5
5	1;11	F	CI	1;8
6	2;9	M	CI	1;0
7	2;2	M	HA	0;5
8	1;8	M	HA	1;0
9	1;6	F	CI	1;3
10	0;9	M	CI	0;0
11	1;3	F	HA	0;4
12	1;10	M	CI	1;7
13	1;10	F	CI	1;6
14	0;11	F	HA	0;11
15	0;8	F	NH	N/A
16	0;9	F	HA	0;2
17	2;8	F	NH	N/A
18	1;3	F	HA	0;4
19	0;4	F	NH	N/A
20	1;8	M	CI	1;2
21	1;3	F	NH	N/A
22	0;6	F	HA	0;3
23	1;3	M	HA	0;2

Table 12 Conversation block code types and designations selected for the present study

Conversational Block Code	Designation
CIC	Key Child with Adult Female/Male
CIOCAX	Key Child with Other Child and Adult Female/Male
AMF	Adult Female Only
AICF	Adult Female with Key Child
AIOCF	Adult Female with Other Child
AIOCCXF	Adult Female with Key Child and Other Child
AMM	Adult Male Only
AICM	Adult Male with Key Child
AIOCM	Adult Male with Other Child
AIOCCXM	Adult Male with Key Child and Other Child
XIOCA	Other Child with Adult Female/Male
XIC	Key Child and Adult Female/Male with Turns
XIOCAC	Other Child with Key Child and Adult Female/Male with No Turns

References

Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Ambrose, S., VanDam, M., & Moeller, M. P. (2014). Linguistic input, electronic media, and communication outcomes of toddlers with hearing loss. *Ear and hearing, 35*(2), 139.

Ambrose, S., Walker, E., Unflat-Berry, L., Oleson, J., & Moeller, M. P. (2015). Quantity and quality of caregivers' linguistic input to 18-month and 3-year-old children who are hard of hearing. *Ear and Hearing, 36*(1), 48S-59S. doi:https://doi.org/10.1097/AUD.0000000000000209

Atal, B., & Rabiner, L. (1976). A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 24*(3), 201–212.

Bachorowski, J. A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychologycal Science, 8*, 53–57.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278. doi:https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Benders, T. (2013). Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent. *Infant Behavior and Development, 36*(4), 847–862.

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do north American babies hear? A large-scale cross-corpus analysis. *Developmental Science*, 22(1), e12724.

Bergeson, T. R., Miller, R. J., & McCune, K. (2006). Mothers' speech to hearing-impaired infants and children with cochlear implants. *Infancy, 10*(3), 221–240.

Bland, J. M., & Altman, D. G. J. L (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. *1*(8476), 307–310.

Boersma, D. C., & Weenink, D. (2017). Praat: Doing phonetics by computer (Version 6.0.29) (Version 6.0.29). Retrieved from http://www.praat.org/

Bořil, T., & Skarnitzl, R. (2016). *Tools rPraat and mPraat.* Paper presented at the International Conference on Text, Speech, and Dialogue.

Breen, M., Dilley, L. C., Kraemer, J., & Gibson, E. (2012). Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory, 8*(2), 277–312. doi:https://doi.org/10.1515/cllt-2012-0011

Burgess, S., Audet, L., & Harjusola-Webb, S. (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. doi:https://doi.org/10.1016/j.jcomdis.2013.09.003

Busch, T., Sangen, A., Vanpoucke, F., & van Wieringen, A. (2017). Correlation and agreement between Language ENvironment Analysis (LENA™) and manual transcription for Dutch natural language recordings. *Behavior Research Methods*. doi:https://doi.org/10.3758/s13428-017-0960-0

Canault, M., Le Normand, M. T., Foudil, S., Loundon, N., & Thai-Van, H. (2016). Reliability of the Language ENvironment Analysis system (LENA™) in European French. *Behavior Research Methods, 48*(3), 1109–1124.

Carletta, J. (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics, 22*(2), 249–254.

Caskey, M., Stephens, B., Tucker, R., & Vohr, B. (2011). Importance of parent talk on the development of preterm infant vocalizations. *Pediatrics, 128*(5), 910–916. doi:https://doi.org/10.1542/peds.2011-0609

Caskey, M., Stephens, B., Tucker, R., & Vohr, B. (2014). Adult talk in the NICU with preterm infants and developmental outcomes. *Pediatrics, 133*(3), e578-584. doi:https://doi.org/10.1542/peds.2013-0104

Caskey, M., & Vohr, B. (2013). Assessing language and language environment of high-risk infants and children: A new approach. *Acta Paediatrica, 102*(5), 451–461. doi:https://doi.org/10.1111/apa.12195

Christakis, D. A., Gilkerson, J., Richards, J. A., Zimmerman, F. J., Garrison, M. M., Xu, D., … Yapanel, U. (2009). Audible television and decreased adult words, infant vocalizations, and conversational turns: a population-based study. *Arch Pediatr Adolesc Med, 163*(6), 554–558. doi:https://doi.org/10.1001/archpediatrics.2009.61

Cristia, A., & Seidl, A. (2013). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language, 41*(4). doi:https://doi.org/10.1017/S0305000912000669

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J. & Bergelson, E. A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods,* in press.

Deller, J. R., Hansen, J. H. L., & Proakis, J. G. (2000). Discrete-time processing of speech signals.

Dubey, H., Sangwan, A., & Hansen, J. H. (2018a). Leveraging Frequency-Dependent Kernel and DIP-Based Clustering for Robust Speech Activity Detection in Naturalistic Audio Streams. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26*(11), 2056–2071.

Dubey, H., Sangwan, A., & Hansen, J. H. (2018b). Robust Speaker Clustering using Mixtures of von Mises-Fisher Distributions for Naturalistic Audio Streams. *arXiv preprint arXiv:1808.06045*.

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.

Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 32*(6), 1109–1121.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development, 60*(6), 1497–1510.

Ford, M., Baer, C. T., Xu, D., Yapanel, U., & Gray, S. (2008). The LENATM Language environment analysis system: Audio specifications of the DLP-0121. *LENA Foundation*.

Garcia-Sierra, A., Ramírez-Esparza, N., & Kuhl, P. K. (2016). Relationships between quantity of language input and brain responses in bilingual and monolingual infants. *International Journal of Psychophysiology, 110*, 1–17.

Gilkerson, J., Coulter, K., & Richards, J. A. (2008). Transcriptional analyses of the LENA natural language corpus. *LENA Foundation*.

Gilkerson, J., & Richards, J. A. (2008). The LENA natural language study. *LENA Foundation*.

Gilkerson, J., Richards, J. A., & Topping, K. J. (2017a). The impact of book reading in the early years on parent–child language interaction. *Journal of Early Childhood Literacy, 17*(1), 92–110. doi:https://doi.org/10.1177/1468798415608907

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Oller, D. K., … Paul, T. D. (2017b). Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis. *Am J Speech Lang Pathol, 26*(2), 248–265. doi:https://doi.org/10.1044/2016_AJSLP-15-0169

Gilkerson, J., Richards, J. A., Warren, S. F., Oller, D. K., Russo, R., & Vohr, B. (2018). Language experience in the second year of life and language outcomes in late childhood. *Pediatrics, 142*(4), e20174276. doi:https://doi.org/10.1542/peds.2017-4276

Gilkerson, J., Zhang, Y., Xu, D., Richards, J. A., Xu, X., Jiang, F., … Topping, K. J. (2015). Evaluating Language Environment Analysis system performance for Chinese: A pilot study in Shanghai. *Journal of Speech, Language, and Hearing Research, 58*(2), 445–452. doi:https://doi.org/10.1044/2015_JSLHR-L-14-0014

Greenwood, C. R., Carta, J. J., Walker, D., Watson-Thompson, J., Gilkerson, J., Larson, A. L., & Schnitz, A. (2017). Conceptualizing a public health prevention intervention for bridging the 30 million word gap. *Clinical Child and Family Psychology Review, 20*(1), 3–24.

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & Gilkerson, J. (2011). Assessing children's home language environments using automatic speech recognition technology. *Communication Disorders Quarterly, 32*(2), 83–92. doi:https://doi.org/10.1177/1525740110367826

Gries, S. T. (2016). *Quantitative corpus linguistics with R: A practical introduction*. Taylor & Francis.

Hansen, J. H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine, 32*(6), 74–99.

Hansen, J. H., Joglekar, A., Shekhar, M. C., Kothapally, V., Yu, C., Kaushik, L., & Sangwan, A. (2019). The 2019 inaugural Fearless Steps Challenge: A giant leap for naturalistic audio. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association* (Interspeech 2019), 1851–1855.

Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America, 101*(1), 466–481.

Hanson, H. M., & Chuang, E. S. (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *Journal of the Acoustical Society of America, 106*(2), 1064–1077.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes.

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child development, 73*(2), 418–433. doi:https://doi.org/10.1111/1467-8624.00415

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27*(2), 236–248.

Irvin, D. W., Hume, K., Boyd, B. A., McBee, M. T., & Odom, S. L. (2013). Child and classroom characteristics associated with the adult language provided to preschoolers with autism spectrum disorder. *Research in Autism Spectrum Disorders, 7*(8), 947–955.

Iseli, M., Shue, Y.-L., & Alwan, A. (2006). *Age-and gender-dependent analysis of voice source characteristics.* Paper presented at the Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on.

Jaeger, F. T. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446. doi:https://doi.org/10.1016/j.jml.2007.11.007

Johnson, K., Caskey, M., Rand, K., Tucker, R., & Vohr, B. (2014). Gender differences in adult-infant communication in the first months of life. *Pediatrics, 134*(6), e1603–1610. doi:https://doi.org/10.1542/peds.2013-4289

Kaushik, L., Sangwan, A., & Hansen, J. H. (2018). Speech Activity Detection in Naturalistic Audio Environments: Fearless Steps Apollo Corpus. *IEEE Signal Processing Letters*, 25(9), 1290–1294.

Ko, E.-S., Seidl, A., Cristia, A., Reimchen, M., & Soderstrom, M. (2016). Entrainment of prosody in the interaction of mothers with their young children. *Journal of Child Language, 43*(2), 1–26. doi:https://doi.org/10.1017/S0305000915000203

Kondaurova, M. V., Bergeson, T. R., & Dilley, L. C. (2012). Effects of deafness on acoustic characteristics of American English tense/lax vowels in maternal speech to infants. *Journal of the Acoustical Society of America, 132*(2), 1039-1049. doi:https://doi.org/10.1121/1.4728169

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*: Sage Publications.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., … Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277*, 684–686. doi:https://doi.org/10.1126/science.277.5326.684

Lam, C., & Kitamura, C. (2010). Maternal interactions with a hearing and hearing-impaired twin: Similarities and differences in speech input, interaction quality, and word production. *Journal of Speech, Language, and Hearing Research, 53*, 543–555.

Lam, C., & Kitamura, C. (2012). Mommy, speak clearly: Induced hearing loss shapes vowel hyperarticulation. *Developmental Science, 15*(2), 212–221.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174. doi:https://doi.org/10.2307/2529310

Ludbrook, J. J. C. a. E. P. a. P. (1997). Special article comparing methods of measurement. *24*(2), 193–203.

Marchman, V. A., Martínez, L. Z., Hurtado, N., Grüter, T., & Fernald, A. (2017). Caregiver talk to young Spanish-English bilinguals: comparing direct observation and parent-report measures of dual-language exposure. *Developmental science, 20*(1).

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315.

McCauley, A., Esposito, M., & Cook, M. (2011). *Language environment of preschoolers with autism: Validity and applications.* Paper presented at the LENA Users Conference, Denver, CO.

Montag, J. L., Jones, M. N., & Smith, L. B. (2018). Quantity and diversity: Simulating early word learning environments. *Cognitive science, 42*, 375–412.

Oetting, J. B., Hartfield, L. R., & Pruitt, J. S. (2009). Exploring LENA as a tool for researchers and clinicians. *The ASHA Leader, 14*(6), 20–22. doi:https://doi.org/10.1044/leader.ftr3.14062009.20

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, 107*(30), 13354–13359. doi:https://doi.org/10.1073/pnas.1003882107

Ota, C. L., & Austin, A. M. B. (2013). Training and mentoring: Family child care providers' use of linguistic inputs in conversations with children. *Early Childhood Research Quarterly, 28*(4), 972–983.

Pae, S., Yoon, H., Seol, A., Gilkerson, J., Richards, J. A., Ma, L., & Topping, K. J. (2016). Effects of feedback on parent–child language with infants and toddlers in Korea. *First Language, 36*(6), 549–569. doi:https://doi.org/10.1177/0142723716649273

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J., Röder, S., Andrews, P. W., … Feinberg, D. R. (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour, 95*, 89–99.

Pisanski, K., & Rendall, D. (2011). The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. *Journal of Acoustical Society of America, 129*(4), 2201–2212. doi:https://doi.org/10.1121/1.3552866

Podesva, R. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics, 11*(4), 478–504. doi:https://doi.org/10.1111/j.1467-9841.2007.00334.x

Porritt, L., Zinser, M., Bachorowski, J.-A., & Kaplan, P. (2014). Depression Diagnoses and Fundamental Frequency-Based Acoustic Cues in Maternal Infant-Directed Speech. In (Vol. 2014, pp. 51–67).

Proakis, J., Deller, J., & Hansen, J. (1993). Discrete-time processing of speech signals. *New York, Macrnillan Pub. Co.*

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*(4), 413–425.

R Development Core Team. (2015). R: A language and environment for statistical computing.

Rabiner, L. R., & Juang, B.-H. (1993). *Fundamentals of speech recognition* (Vol. 14): PTR Prentice Hall Englewood Cliffs.

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2017). The Impact of Early Social Interactions on Later Language Development in Spanish-English Bilingual Infants. *Child Development, 88*(4), 1216–1234. doi:https://doi.org/10.1111/cdev.12648

Ramírez-Esparza, N., García-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.

Richards, J. A., Gilkerson, J., Xu, D., & Topping, K. (2017a). How much do parents think they talk to their child? *Journal of Early Intervention, 39*(3), 163–179.

Richards, J. A., Xu, D., Gilkerson, J., Yapanel, U., Gray, S., & Paul, T. (2017b). Automated assessment of child vocalization development using LENA. *Journal of Speech, Language, and Hearing Research, 60*(7), 2047–2063.

Rietveld, T., & van Hout, R. (1993). *Statistical techniques for the study of language and language behavior*: Mouton de Gruyter.

Roberts, M. Y., & Kaiser, A. P. (2011). The effectiveness of parent-implemented language interventions: A meta-analysis. *American Journal of Speech-Language Pathology*.

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-million-word gap: Children's conversational exposure is associated with language-related brain function. *Psychological Science, 29*(5), 700–710. doi:https://doi.org/10.1177/0956797617742725

Rowe, M. L. (2012). Recording, transcribing, and coding interaction. *Research methods in child language: A practical guide*, 191–207. doi:https://doi.org/10.1002/9781444344035.ch13

Sacks, C., Shay, S., Repplinger, L., Leffel, K. R., Sapolich, S. G., Suskind, E., … Suskind, D. L. (2014). Pilot testing of a parent-directed intervention (Project ASPIRE) for underserved children who are deaf or hard of hearing. *Child Language Teaching and Therapy, 30*(1), 91–102. doi:https://doi.org/10.1177/0265659013494873

Sangwan, A., Hansen, J. H. L., Irvin, D. W., Crutchfield, S., & Greenwood, C. R. (2015). *Studying the relationship between physical and language environments of children: Who's speaking to whom and where?* Paper presented at the Signal Processing and Signal Processing Education Workshop (SP/SPE), 2015 IEEE.

Schwarz, I.-C., Botros, N., Lord, A., Marcusson, A., Tidelius, H., & Marklund, E. (2017). *The LENATM system applied to Swedish: Reliability of the Adult Word Count estimate.* Paper presented at the Interspeech 2017.

Seidl, A., Cristia, A., Soderstrom, M., Ko, E.-S., Abel, E. A., Kellerman, A., & Schwichtenberg, A. (2018). Infant–mother acoustic–prosodic alignment and developmental risk. *Journal of Speech, Language, and Hearing Research, 61*(6), 1369–1380.

Sharma, B., Das, R. K., & Li, H. (2019). Multi-level adaptive speech activity detector for speech in naturalistic environments. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association* (Interspeech 2019), 2015–2019.

Shneidman, L. A., Arroyo, M. E., Levine, S. C., & Goldin-Meadow, S. (2013). What counts as effective input for word learning? *Journal of Child Language, 40*(3), 672–686.

Sholokhov, A., Sahidullah, M., & Kinnunen, T. (2018). Semi-supervised speech activity detection with an application to automatic speaker verification. *Computer Speech and Language*, 47, 132–156.

Soderstrom, M., & Wittebolle, K. (2013). When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. *PLoS One, 8*(11), e80646. doi:https://doi.org/10.1371/journal.pone.0080646

Suskind, D. L., Graf, E., Leffel, K. R., Hernandez, M. W., Suskind, E., Webber, R., … Nevins, M. E. (2016a). Project ASPIRE: Spoken language intervention curriculum for parents of low-socioeconomic status and their Deaf and Hard-of-Hearing Children. *Otology & Neurotology, 37*(2), e110–e117.

Suskind, D. L., Leffel, K. R., Graf, E., Hernandez, M. W., Gunderson, E. A., Sapolich, S. G., … Levine, S. C. (2016b). A parent-directed language intervention for children of low socioeconomic status: A randomized controlled pilot study. *Journal of child language, 43*(2), 366–406. doi:https://doi.org/10.1017/S0305000915000033

Syrdal, A. K., & McGory, J. (2000). *Inter-transcriber reliability of ToBI prosodic labeling.* Paper presented at the International Conference on Spoken Language Processing, Beijing, China.

Talbot, M. (2015). The talking cure. *The New Yorker, 90*, 43.

Th. Gries, S. (2015). The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora, 10*(1), 95–125.

Thiemann-Bourque, K., Warren, S. F., Brady, N., Gilkerson, J., & Richards, J. A. (2014). Vocal interaction between children with Down syndrome and their parents. *American Journal of Speech-Language Pathology, 23*(3), 474–485. doi:https://doi.org/10.1044/2014_AJSLP-12-0010

VanDam, M., Ambrose, S., & Moeller, M. P. (2012). Quantity of parental language in the home environments of hard-of-hearing 2-year-olds. *J Deaf Stud Deaf Educ, 17*(4), 402–420. doi:https://doi.org/10.1093/deafed/ens025

VanDam, M., & Silbert, N. H. (2013). *Precision and error of automatic speech recognition.* Paper presented at the Proceedings of Meetings on Acoustics ICA2013.

VanDam, M., & Silbert, N. H. (2016). Fidelity of automatic speech processing for adult and child talker classifications. *PLoS One, 11*(8), e0160588. doi:https://doi.org/10.1371/journal.pone.0160588

Vigil, D. C., Hodges, J., & Klee, T. (2005). Quantity and quality of parental language input to late-talking toddlers during play. *Child Language Teaching and Therapy, 21*(2), 107–122.

Wang, Y., Hartman, M., Aziz, N. A. A., Arora, S., Shi, L., & Tunison, E. (2017). A systematic review of the use of LENA technology. *American Annals of the Deaf, 162*(3), 295–311.

Warlaumont, A. S., Oller, D. K., Dale, R., Richards, J. A., Gilkerson, J., & Xu, D. (2010). *Vocal interaction dynamics of children with and without autism.* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. *Psychological Science, 25*(7), 1314–1324. doi:https://doi.org/10.1177/0956797614531023

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. doi:https://doi.org/10.1007/s10803-009-0902-5

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science, 24*(11), 2143-2152. doi:https://doi.org/10.1177/0956797613488145

Weizman, Z. O., & Snow, C. E. (2001). Lexical output as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology, 37*(2), 265–279. doi:https://doi.org/10.1037/0012-1649.37.2.265

Wieland, E., Burnham, E., Kondaurova, M. V., Bergeson, T. R., & Dilley, L. C. (2015). Vowel space characteristics of speech directed to children with and without hearing loss. *Journal of Speech, Language and Hearing Research, 58*(2), 254–267. doi:https://doi.org/10.1044/2015_JSLHR-S-13-0250

Wong, K., Boben, M., & Thomas, C. (2018). *Disrupting the early learning status quo: Providence Talks as an innovative policy in diverse urban communities.*

Xu, D., Gilkerson, J., Richards, J., Yapanel, U., & Gray, S. (2009a). *Child vocalization composition as discriminant information for automatic autism detection.* Paper presented at the Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE.

Xu, D., Richards, J. A., Gilkerson, J., Yapanel, U., Gray, S., & Hansen, J. (2009b). *Automatic childhood autism detection by vocalization decomposition with phone-like units.* Paper presented at the Proceedings of the 2nd Workshop on Child, Computer and Interaction.

Xu, D., Yapanel, U., & Gray, S. (2009c). *Reliability of the LENA™ Language Environment Analysis System in young children's natural home environment (LENA Technical Report LTR-05-2).* Retrieved from Boulder, CO: http://lena.org/wp-content/uploads/2016/07/LTR-05-2_Reliability.pdf

Xu, D., Yapanel, U., Gray, S., & Baer, C. T. (2008a). The LENA Language Environment Analysis System: The interpretive time segments (ITS) file. *LENA Research Foundation Technical Report LTR-04-2.*

Xu, D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J. A., & Hansen, J. H. L. (2008b). *Signal processing for young child speech language development.* Paper presented at the First Workshop on Child, Computer and Interaction.

Zhang, Y., Xu, X., Jiang, F., Gilkerson, J., Xu, D., Richards, J. A., … Topping, K. J. (2015). Effects of quantitative linguistic feedback to caregivers of young children: A pilot study in China. *Communication Disorders Quarterly, 37*(1), 16–24. doi:https://doi.org/10.1177/1525740115575771

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. doi:https://doi.org/10.1542/peds.2008-2267