Look who's talking: A comparison of automated and human-generated speaker tags in naturalistic daylong recordings

Federica Bulgarelli[1] & Elika Bergelson[1]

[1] Duke University

Author Note

Correspondence: Federica Bulgarelli, 417 Chapel Drive, Box 90086, Duke University, Durham, NC 27708-0086 USA Email: fedebul@gmail.com Phone: 919-684-9429 This work was supported by NIH grant to EB (DP5 OD019812-01). We wish to thank all of the research assistants at the University of Rochester and Duke University who coded the talker producing each utterance, allowing for the analyses in the current manuscript.

Correspondence concerning this article should be addressed to Federica Bulgarelli, 417 Chapel Drive. E-mail: fedebul@gmail.com

Abstract

The LENA system has revolutionized research on language acquisition, providing both a wearable device to collect daylong recordings of children's environments, and a set of automated outputs that process, identify, and classify speech using proprietary algorithms. This output includes information about input sources (e.g. adult male, electronics). While this system has been tested across a variety of settings, here we delve deeper into validating the accuracy and reliability of LENA's automated diarization, i.e. tags of who is talking. Specifically, we compare LENA's output with a gold standard set of manually-generated talker tags from a dataset of 88 daylong recordings, taken from 44 infants at 6 and 7 months, which includes 57,983 utterances. We compare accuracy across a range of classifications from the original Lena Technical Report, alongside a set of analyses examining classification accuracy by utterance type (e.g. declarative, singing). Consistent with previous validations, we find overall high agreement between the human and LENA-generated speaker tags for adult speech in particular, with poorer performance identifying child, overlap, noise, and electronic speech (accuracy range across all measures: 0-92%). We discuss several clear benefits of using this automated system alongside potential caveats based on the error patterns we observe, concluding with implications for research using LENA-generated speaker tags.

Keywords: LENA System, talker variability, LENA system reliability

Look who's talking: A comparison of automated and human-generated speaker tags in naturalistic daylong recordings

Introduction

Understanding the properties of children's linguistic input and how it shapes knowledge acquisition has been of interest to researchers for many decades (Hart and Risley, 1995; Williams, 1937; Taine, 1876). While lab-based experiments provide valuable information about what children know using tightly controlled experimental manipulations, information about naturalistic input is also critically important for understanding how children learn from their daily environment. The majority of observational research on language development has been conducted by collecting video and audio samples of child-caregiver interactions, alongside painstaking and labor-intensive manual transcription by trained researchers (Macwhinney, 2019; Nelson, 1973). The widely used Language ENvironment Analysis system (LENA, LENA Foundation, Boulder, CO, Greenwood et al., 2011) revolutionized this process, combining a lightweight wearable audio-recorder with a proprietary algorithm that processes the audio signal. The output of this algorithm then provides researchers and parents with estimates of a variety of information about the recorded linguistic input, including adult word counts, child vocalization counts, and conversational turns between the adult and the child wearing the recorder (i.e. the "target" child).

The LENA system was designed with research, intervention, and clinical settings in mind; its output can readily provide parents with feedback about the language their children hear. While a key focus of LENA users has been word counts and conversational turns (Gilkerson et al., 2017), the algorithm also exhaustively classifies the input into "utterances" across eight different talker categories: target child, other children, adult males, adult females, overlapping sounds, noise, electronic sounds, and silence. The source, quantity, and quality of input play an important role in language development, and indeed LENA output has been used to identify the relative proportion of speech to infants coming from speakers of different genders and ages, as well as from electronics (Christakis et al., 2009; Sosa, 2016; Richards et al., 2017).

One reason characterizing talkers in the input is important concerns early speech-sound learning. Indeed, an early challenge for young learners is identifying their language's speech sounds, which requires deducing the right consonant and vowel categories based on input that varies across and within talkers, and by phonetic context. Adding to this challenge, the same speech sound varies acoustically as a function of distinct vocal characteristics, alongside variables such as gender, age, topic or dialect (Liberman et al., 1967). Detecting the "invariant" (i.e. relatively stable and consistent) aspects of the input is an important part of learning language (Gogate and Hollich, 2010), one that is inevitably dependent on the amount and type of variability infants experience. As talker variability has been posited to be both beneficial (e.g. Rost and Mcmurray, 2009), and to pose a challenge (e.g. Mullennix et al., 1989; Jusczyk et al., 1992) for language learning, the speaker tags LENA provides are an important information source for moving theory forward.

However, before confidently using the LENA system's automated output to study talkers in children's input, it is necessary to establish its talker classification accuracy. That is, while the opportunity to crunch 1000s of hours of data in just dozens of hours with little human labor required is enticing, it is critical to understand the limitations of any automated approach, both for interpretive validity, and to help guide speech technology improvement. While many labs continue to use some method of manual annotation to look at variables of interest (e.g. Weisleder and Fernald, 2013; Soderstrom and Wittebolle, 2013; Bergelson and Aslin, 2017; Bergelson et al., 2018a), others use the output from the LENA software as ground truth (Johnson et al., 2014b). Especially since the LENA system has great potential for facilitating diagnosis and intervention for children at risk for language delays and deficits, it is imperative to understand the system's accuracy and error patterns in order to properly interpret research using LENA output.

Around LENA's initial release, Xu et al. (2009) published a LENA Technical Report (LTR-05-2) testing the software's accuracy on a test set consisting of one hour long segments from each of 70 test subjects ranging from 2-36 months from the LENA Natural Language Study (building on results in Xu et al., 2008). The hour long segments were made up of six 10-minute segments identified by an algorithm to include high levels of speech activity between the target child and an adult. The test set was analyzed by the LENA proprietary software, and by trained human transcribers. Xu et al. (2009) compared speaker tags generated by the LENA software to those generated by the trained human transcribers across four categories: adult speech, child speech, television, and other; the system attained 82%, 76%, 71%, and 76% accuracy, respectively. Overall, Xu et al. (2009) thus report high levels of agreement between the LENA proprietary software and trained human transcribers, noting false negatives for overlapping speech as the algorithm's greatest source of error.

In a similar endeavor, but using a different tack, Vandam and Silbert (2016) compared LENA's talker-tags with those generated by 23 trained judges. They obtained day-long LENA recordings from 26 families with 2.5 year old children, and extracted 30 "segments" (LENA's proxy for utterances) from three LENA categories of interest: adult male, adult female, and target child. These segments were systematically extracted over the course of the day, to avoid potential skew from oversampled contexts, environments, or times. All judges tagged each segment (in random order, i.e. without context) as child, male, female, or other. In this 4-way categorization of LENAs three categories, there was high agreement between the trained judges and the LENA software (weighted Fleiss $\kappa = .68$). Additionally, the authors were able to identify two key error-patterns in the LENA-generated tags. First, when a segment was tagged as "child" by judges but not by the LENA system, the LENA system generally tagged the segment as "female" rather than "male". Second, for segments tagged "female" by judges but not by the LENA system, the LENA system generally tagged the segments as male rather than child.

Another study (Lehet et al., 2018) investigated the LENA system's accuracy in classifying speech as speech, with particular interest in classifying adult speech at a fine granularity. They sampled 15 day-long audio recordings from children aged 7-33 months, analyzing approximately 30 minutes of audio sampled throughout the day from each recording. Each LENA segment was also coded by trained annotators as male, female, or child speech. These manual speaker tags were then compared to LENA-generated speaker tags every 50ms, revealing 70% agreement. Follow-up comparisons revealed that the LENA system was most accurate at classifying human speech (adult or child) from nonspeech (noise, 76-78% accuracy), but less accurate at differentiating between adult speech and speech from children or electronic devices (68% accuracy).

Taken together, these three studies (along with others, e.g. Mccauley et al. (2011); Soderstrom and Franz (2016)) provide consistent evidence that LENA's proprietary software is fairly accurate at classifying speech relative to trained human coders, while highlighting a variety of systematic mistakes. However, the literature to date leaves three clear gaps that the current work fills.

First, across these previous studies, the annotaters heard decontextualized clips and/or had little familiarity with the families. This critically differs from the infants' own experiences of their day, where activities and interactions have a coherent context and order, and are set against a firm basis of experience with particular key caretakers. To better approximate infants' experiences, we use manual annotations created by listening to the entire day in order (except nap-times), by researchers who know individual families well. This provides a contextual coherence to the tags, and protects against biases that emerge when listening to unknown talkers. For instance, someone familiar with a family may know that there is a toddler in addition to the target child, and that the grandmother, who is the primary caretaker, has a relatively deep voice. A naive annotator or algorithm couldn't know this information, and thus will likely make errors in attributing child vocalizations to the key vs. other child, or will use the (generally reliable) proxy that deeper voices belong to men rather than women.

Second, all of these previous studies used recordings from large age ranges (from 2-36 months), and either collapsed across all child categories (target vs. other child, Xu et al., 2009) or only investigated accuracy on one of the categories (target child, Vandam and Silbert, 2016). Across this developmental period children go from not producing any speech, to being active participants in conversation. Understanding the LENA system's accuracy in determining the source of child speech is critical, given that a primary goal of the LENA system is to collect information about child vocalizations and turn-taking to assess and promote language development. For example, if the LENA system has difficulty distinguishing between the target child and other children in the environment, these types of data can give a misleading assessment of the target child's vocal maturity in settings with more than one child present. This may be particularly problematic in low-SES settings, where family size tends to be larger, and caretaking more often involves multiple children (United States Statistics Division, 2015).

Lastly, while many previous evaluations of the LENA software focused on portions of recordings with a high density of speech, the type or content of this speech is not considered in identifying the speaker. A recent investigation of the availability of both child directed and adult directed speech in infants' input over the first two years of life (Bergelson et al., 2018b) found differences in accuracy in identifying speaker gender depending on child or adult directed speech. For child directed speech, the LENA algorithm misclassified a male speaker as female 10% of the time, but only misclassified a female speaker as male 4% of the time. Whereas for adult directed speech, a female speaker was mislabeled as male 34% of the time, while a male speaker was only mislabeled as female 22% of the time. These errors likely stem from child directed speech being characterized by overall higher pitch, making it more difficult for algorithms to differentiate child directed male speech from adult directed female speech.

Other aspects of the speech content itself can also potentially impact the algorithm's accuracy. For example, declarative statements and questions are marked by different intonational contours, which primarily include changes in fundamental frequency (Lieberman, 1967). As with child directed speech, it is therefore reasonable to expect that utterance-type may also impact talker classification by the LENA software. Indeed, different utterance-types have been proposed to serve different roles for language acquisition. For instance, words in single-word utterances or at the beginnings and ends of sentences (edges) have been proposed to scaffold segmentation (Brent and Siskind, 2001; Johnson et al., 2014a), while prosodic patterns of longer utterances can highlight syntactic boundaries (Nelson et al., 1989; see Soderstrom, 2007). Questions, in turn, have particular prosody, with yes/no questions in particular suggested to support auxiliary development (Gleitman et al., 1984). Situational contexts like reading and singing also have particular prosody and content. For instance, singing, a common caretaking activity with parallels to infant directed speech (Trehub et al., 1993), may pose a challenge for automated systems given its wider contour range. Finally, reading has been a particular focus in early language development, and features a distinctively wider range of words and grammatical constructions, and prosody (Montag et al., 2015; Debaryshe, 1993). Taken together, understanding how context, and inevitable variability in utterance-type, impact talker classification is relevant for language development more generally.

Filling these gaps, the current study uses a recently collected longitudinal corpus, the SEEDLingS corpus (Bergelson, 2017) to investigate LENA software-generated talker tags taken from a set of day-long longitudinal audio recordings of 44 typically-developing infants in a North American city. We restrict the current analyses to segments where trained researchers identifed that a noun was spoken to the target child, by a person, toy, or electronic device, and the type of utterance the noun was spoken in. We focus on instances of concrete nouns, given their high prevalence in early vocabulary (Braginsky et al., 2017)[1]. Furthermore, unlike previous investigations, we restrict our analysis to day-long recordings from 6 and 7 months of age, which allows us to investigate the child tags when the target child is not yet producing words, making it easier to identify patterns of mistakes in labeling target or other child utterances. Taken together, this paper goes beyond previous work in by comparing LENA algorithm speaker tags to those produced by trained researchers highly familiar with the context and individuals in the recordings, in a relatively large sample of pre-verbal infants.

Methods

Participants

Participants were 44 infants recruited for a large scale, year-long study of word learning. All infants were born full term (40±3 weeks), had no known vision or hearing problems, and heard English $\geq$ 75% of the time. 75% of the infants' mothers had a B.A. or higher, and 95% of the infants were Caucasian. Over the course of the year-long study starting when infants were 6 months of age, families were recorded using LENA once a month for an entire day, and video recorded once a month for an hour. For the purpose of the current study, only the audio recordings from 6 and 7 months were used, as these were the only portions of the data where the entire day was manually annotated. See Bergelson et al. (2018a) for a fuller description of the data and Bergelson (2017) to access the recordings directly.

[1] Further details about the generalizability of our noun-centric analysis is taken up in the Discussion

Procedure

Home recordings and Initial Data Processing. Researchers obtained monthly audio recordings capturing up to 16 hours of infants' language input each month. Parents were given small LENA audio recorders (LENA Foundation), and infant-sized vests with built-in pockets to house the LENA recorder. Parents were asked to have their child wear the vest and the LENA recorder from the time they woke up until they went to sleep for the night, except for naps and bath times. Parents were permitted to pause the recorder, but were asked to minimize these pauses.

Audio recordings were processed by LENA proprietary software, which segments each file and diarizes it (i.e. demarcates the onset and offset of every "utterance" and assigns it one of the eight talker-tags in its inventory, (Xu et al., 2008)).[2] The output from the LENA proprietary software was converted to CLAN format (MacWhinney and Wagner, 2010). In-house scripts were used to mark long periods of silence (such as naps) in the raw audio files, without information from the LENA software. Research assistants subsequently verified the edges of these long periods of silence using visual inspection of the waveform.[3] Subsequently these files were used for manual language annotation. Original audio recordings were modally 16 hrs (LENA's maximum capacity). After removing long silences, the recordings were ~10 hrs (Mode = 654 min, Mean = 603 min, SD = 106.8, Range = 385.2-951 min, see Bergelson et al. (2018a))

Manual Annotation. Trained researchers listened to the full daylong recording, and within each utterance delimited by the LENA software, annotated each concrete noun

[2] N.B. While the LENA technical report (Xu et al., 2009) states accuracy for the talker tags, as described in text, it does not report accuracy on the segment identification process, i.e. whether a human would agree with the utterance boundaries identified by LENA, regardless of talker.

[3] Process detailed here: https://bergelsonlab.gitbook.io/project/seedlings-annotations/audio-processing

said directly to or near the target child. Specifically, concrete noun tags were placed within timestamps delimited by the LENA software as utterances. However, multiple concrete nouns could occur within a single utterance delimited by LENA, or across utterance boundaries (in which case they were included in the timestamp where the majority of the word occurred). Based on the goals of the broader project, which examines noun acquisition (Bergelson and Aslin, 2017), trained researchers tagged easily imageable concrete nouns that could be visually represented, and included objects such as body parts (i.e. arm, leg) and foods (i.e. milk, cracker), but did not include occupations (e.g. teacher), or proper nouns.[4] Concrete nouns produced in the distance (such as faint background television) were not included. Each concrete noun instance was labeled alongside its utterance-type, a tag for whether the referent of the noun was present, and individual talker labels (see Bergelson et al., 2018a). The current analysis focuses primarily on the talker label, which tagged concrete nouns from any talker (live interlocutors and electronics), and on the utterance-type, which labeled the utterance as one of the following: declarative, imperative, reading, singing, short phrase (i.e. less than three words with no verb, see Bergelson et al. (2018a)).

Each talker was labeled with a unique identifier describing that specific talker. For example, mom was always MOT and maternal grandmother was always GRM, while other speakers' 3-letter codes indicated whether they were an adult or child, and male or female. The same label was used throughout the recordings for recurrent talkers (e.g. Aunt Sarah might be AFS for a given infant.) Unique 3-letter codes were also used when a word was spoken by multiple simultaneous talkers (e.g. mom and dad said "ball" at the same time). Each talker tag was created and checked by two different RAs initially. It then underwent a final check by a trained researcher highly familiar with each family (i.e. who could identify each individual talker present in the recordings and know, e.g. that a given family had 2 older brothers); this researcher confirmed the set of talker-tags for each child was accurate

[4] Further details here: https://bergelsonlab.gitbook.io/project/seedlings-annotations/annotation-notes-1

and consistent across recordings each month. The current dataset thus includes an average of 1,317.80 tags per child (SD = 620.05, Mode = 1,123.28, Range = 292-2726) for which we have both a LENA-generated and manual speaker tag.

Converting talker annotations to LENA-generated speaker tags. In order to compare the talker tags produced by trained research assistants with those produced by the proprietary LENA software, we reclassified our unique talker-tags to match those produced by the LENA software: female or male adult, target or other child, electronic, and overlap. Utterances labeled as electronic were produced exclusively by toys or television. The overlap category consisted of utterances produced by any two sources (e.g. two adults, a child and singing toy, etc.). Across the main set of analyses, we do not consider utterances labeled as noise or silence by the LENA algorithm, as our codes did not reflect this category. In the penultimate section of the results, we return to these to identify the types of utterances labeled as noise or silence by the LENA algorithm.[5] Finally, in order to assess inter-rater reliability for our human annotations, researchers blind to the existing tags coded 3150 concrete-noun instances (5% of the entire corpus) using speaker tags equivalent to those used by LENA: male adult, female adult, child, electronic or overlap. Reliability was high: accuracy = 96.56, kappa = 0.93.

[5] N.B. The LENA algorithm provides ’far’ and ’near’ versions of all tags except silence, LENAs own reported classification accuracy uses only near-field utterances, and we follow suit (Xu et al., 2009).

Data analysis

We used R and RStudio (Version 3.4.3; R Core Team, 2017)[6], to generate this manuscript, along with all figures and analyses. All code and data are already available (https://github.com/fedebul/BulgarelliBergelson_BehavioralResearchMethods2019).

In order to compare our results to those published in the original LENA technical report (LTR), we analyze the results of a series of confusion matrices. First, we analyze the four higher-level categories (adult, child, electronic, overlap), as in previous validations (Xu et al., 2009). Next we compare the LENA algorithm's performance on specific subsets of the data. Specifically, we look for cases where human coders and the LENA speaker tags agree that the speech segments are one of two categories: adult vs. child tags, male vs. female adult tags, target child vs. other child tags, and electronic vs. overlap tags. This allows us to investigate specific error patterns. For example, for the adult vs. child comparison we can ask: given agreement that the speaker is human, how accurate was the LENA algorithm at correctly identifying whether the speaker was an adult or child? For the electronic vs. overlap comparison we can ask: having established that the signal is not clear human speech, how accurate is the LENA algorithm at identifying its source? Further, as we only included segments that were identified by annotators as being spoken by a human, toy or electronic, we investigate LENA system's use of the noise and silence tags. Lastly, for each of these comparisons we investigate whether the LENA algorithm's accuracy is dependent on the type of utterance for each segment, based on the manual utterance-type tags (for which

[6] We used *bindrcpp* (Version 0.2.2; Müller, 2018), *broom* (Version 0.5.0; Robinson and Hayes, 2018), *caret* (Version 6.0.80; from Jed Wing et al., 2018), *childesr* (Version 0.1.0; Braginsky et al., 2018), *dplyr* (Version 0.8.0.1; Wickham et al., 2018), *ggplot2* (Version 3.1.0; Wickham, 2016), *ggpubr* (Version 0.2; Kassambara, 2018), *irr* (Version 0.84.1; Gamer et al., 2019), *janitor* (Version 1.1.1; Firke, 2018), *kableExtra* (Version 1.0.1; Zhu, 2019), *knitr* (Version 1.21; Xie, 2015), *magrittr* (Version 1.5; Bache and Wickham, 2014),*papaja* (Version 0.1.0.9842; Aust and Barth, 2018), *purrr* (Version 0.3.2; Henry and Wickham, 2018), and *tidyverse* (Version 1.2.1; Wickham, 2017).

there is no LENA system equivalent).

In all cases, manual tags are used as the gold standard against which the LENA-generated tags are assessed. We report accuracy (% agreement and Cohen's κ), alongside recall, precision and F1. Percent agreement reflects overall accuracy (# of *correct* tags/# of all tags), while Cohen's κ takes into account chance agreement due to randomly guessing, or always choosing a single response. Recall is operationalized as the rate of correct predictions divided by the total number of actual instances. Precision is our measure of correct identification. For example, for checking accuracy in classifying adult vs. child speech, recall would be: (# of *correct* LENA adult tags)/(# of Manual adult tags), while precision would be: (# of *correct* LENA adult tags)/(total # of LENA adult tags). Lastly, F1 is a weighted average of the recall and precision, with 1 reflecting perfect accuracy.

Results

Table 1 shows the number of utterances in each talker category as tagged manually and by the LENA software. Overall, LENA-generated talker tags and the manual talker tags were moderately correlated, (n = 57983, Kendall's τ=0.35, $p < .001$).

Table 1

Nouns in each category, by tag source

Speaker type	Human codes	LENA codes
Adult	51,097	39,532
Child	3,022	6,331
Electronic	3,165	2,702
Overlap	699	9,418
Total	57,983	57,983

Classifying LENA-generated vs. Human-generated adult, child, electronic, and overlap tags. We first analyzed accuracy for all of the speaker tags that were classified as adult, child, electronic or overlap. Across the four categories, the LENA system's overall accuracy was 0.72, Cohen's $\kappa = 0.28$. The confusion matrix results for these categories can be found in Figure 1 and Table 2. The LENA technical report (Xu et al., 2009) reports sensitivityin classifying each category, which here can be compared directly to recall from the confusion matrix. In all cases, our results show lower agreement percentages (by 1-38%) than the LENA technical report.

Table 2

Recall, precision and F1 for all four categories and comparison to Lena Technical Report sensitivity estimates

Type	Recall	Precision	F1	LTR Report
Adult	0.75	0.97	0.85	0.82
Child	0.57	0.27	0.37	0.76
Electronic	0.46	0.54	0.49	0.71
Overlap	0.38	0.03	0.05	0.76

Descriptively, when the LENA algorithm misclassified adult speech, it was most likely to classify it as overlap (15%). Similarly, when it misclassified child speech, it was most likely to classify it as overlap (23%) or adult speech (19%). Electronic speech was most likely to be misclassified as overlap (28%), and overlap speech was most likely to be misclassified as adult (29%). From these results, and consistent with the technical report, we can draw the preliminary conclusion that the LENA algorithm is overly sensitive to overlapping sounds, relative to human annotaters.

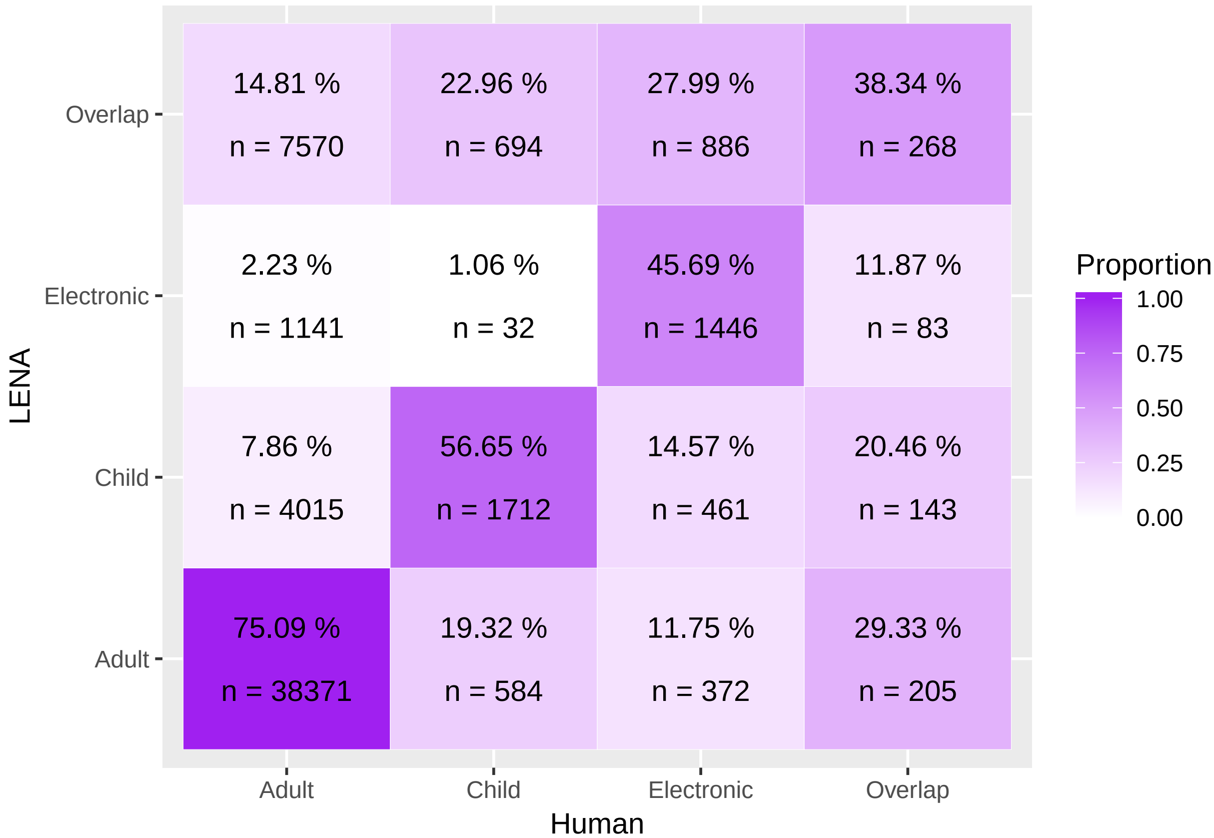

Figure 1. Confusion matrix displaying recall for LENA-generated labels compared to Human-generated labels. Each column constitutes all of the instances labeled by human coders as belonging to that category. Each cell displays how LENA software tags were labeled for each human category, as well the total number of segments in each cell. Darker colors represent a higher proportion of LENA software tags.

Despite lower agreement in the current dataset than in the LTR, we do find a significant (non-parametric) correlation across the proportion of the LENA system tags for each human tag category between these data and the percentages reported in the LTR for the equivalent confusion matrix (i.e. Figure 1), n= 16, Kendall's τ=0.74, $p < .001$.

Finally, we assessed whether accuracy varied as a function of utterance-type. Accuracy was operationalized as correct (scored "1") if the LENA-generated tag matched the human tag and incorrect (scored "0") if it did not. We then conducted a logistic regression with accuracy as the dependent variable and utterance-type (Declarative, Imperative, Short Phrase, Question, Reading or Singing) as a predictor. Utterance-type was significant $\chi^2 = (5, N = 57982)$ $p < .001$. As can be seen in Table 3 and Figure 2 the LENA software is incorrect nearly half of the time for Singing utterances, and most accurate on Reading utterances. We return to these descriptive differences in the discussion.

Table 3

Number of correctly and incorrectly classified segments by utterance-type and proportion correct across the four main categories: adult, child, electronic or overlap.

UtteranceType	Incorr	Corr	%Corr
Declarative	7,009	21,221	0.75
Imperative	1,106	2,405	0.68
Short Phrase	1,745	3,049	0.64
Question	2,953	8,252	0.74
Reading	995	3,661	0.79
Singing	2,377	3,209	0.57

Classifying adult vs. child tags. The next confusion matrix compared adult and child tags (excluding other LENA-generated or manual tags). Thus, this analysis investigates accuracy when both human coders and the LENA algorithm agree that the speaker is human, and omit overlap and electronic tags from consideration. The LENA system achieved 0.90 accuracy, Cohen's $\kappa = 0.38$. Recall for this classification is 0.90, while precision is 0.98. The F1 weighted score is 0.94. The error patterns reveal that the LENA system is more likely to misclassify child speech as adult than adult speech as child, see Figure 3. While the accuracy

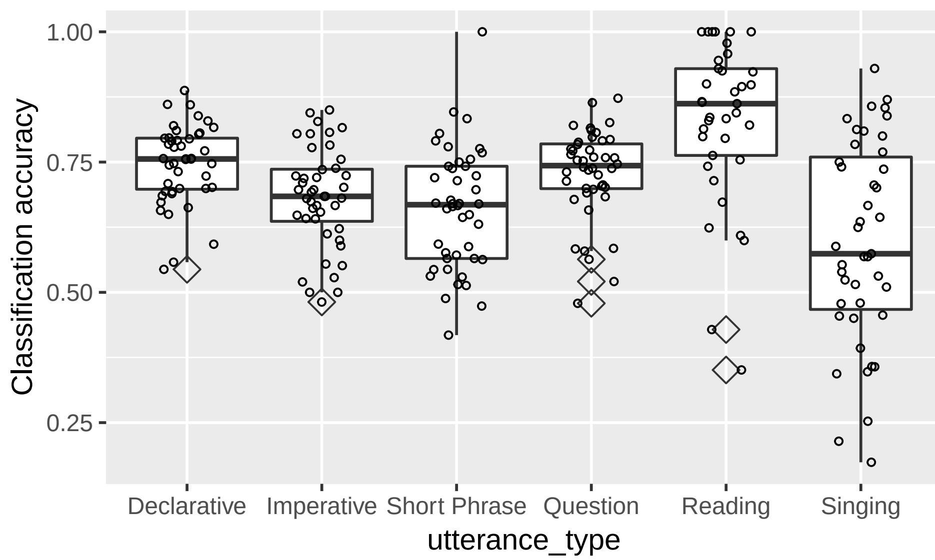

Figure 2. Classification accuracy distribution by utterance-type across the four main categories: adult, child, electronic or overlap. The box plot reflects the median of the means for each infant for each utterance-type. Each point (jittered horizontally) represents one child; diamonds (unjittered) indicate outliers.

for this classification is quite high, it's worth noting the large discrepency between accuracy and κ, which takes into account the chance of correctly guessing.

Here too, a logistic regression showed that utterance-type accounted for significant variance in classifying adult and child speech $\chi^2 = (5, N = 45014)$, $p < .001$. As can be seen in Table 4 and Figure 4 the LENA software is least correct at distinguishing between adult and child speech for singing utterances, and most correct for declaratives, though overall accuracy was quite high (80%– 92%).

Classifying male vs. female adult speech. We next investigated accuracy in labeling talker gender. This analysis only included tags labeled as male or female adults by both the LENA algorithm and human coders, and excluded children, electronics and overlap. The LENA system classified male and female speech with 0.90 accuracy, Cohen's $\kappa = 0.70$. Recall for this classification was 0.93, while precision was 0.94. The F1 weighted score was

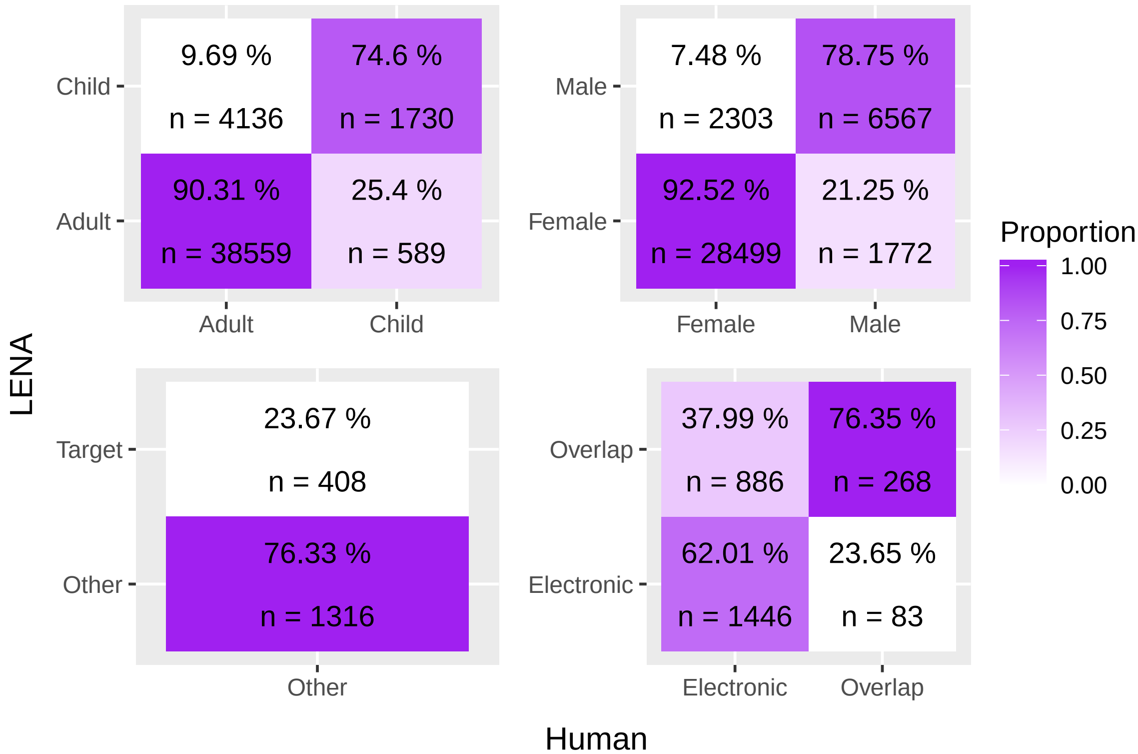

Figure 3. Confusion matrix displaying proportion correct (i.e. recall) for LENA-generated labels compared to Human-generated labels. Each column constitutes all of the instances labeled by human coders. Each cell displays how the LENA system tags were labeled for each human category, as well the total number of segments in each cell. Darker colors represent a higher proportion of LENA system tags.

0.93. The error patterns reveal that the LENA system is more likely to misclassify male speech as female speech than female speech as male speech. Indeed, female speech constitutes 79% of adult speech in the current data set (see Figure 3), a point we return to in the discussion.

Again, a logistic regression found that utterance-type accounted for significant variability in classification accuracy, here for male vs. female speech $\chi^2 = (5, N = 39141)$, $p <$ .001. While the effect of utterance-type was significant, as can be seen in Table 4 and Figure 4, the LENA software is quite accurate at distinguishing male and female speech. Given accuracy differences only ranging from 87% – 92%, utterance-type differences here should

Table 4

Number of correctly and incorrectly classified segments by utterance-type and proportion correct for all 2-way comparisons

	Adult vs. Child			Male vs. Female			Target vs. Other child			Electronic vs. Overlap		
UtteranceType	Incorr	Corr	%Corr	Incorr	Corr	%Corr	Incorr	Corr	%Corr	Incorr	Corr	%Corr
Declarative	1933	21136	0.92	2288	18350	0.89	175	654	0.79	112	88	0.44
Imperative	399	2384	0.86	228	2120	0.90	19	51	0.73	36	22	0.38
Short Phrase	592	3024	0.84	346	2338	0.87	139	306	0.69	96	26	0.21
Question	959	8232	0.90	615	7463	0.92	44	180	0.80	27	22	0.45
Reading	368	3656	0.91	391	3246	0.89	5	18	0.78	2	5	0.71
Singing	474	1857	0.80	207	1549	0.88	26	107	0.80	696	1551	0.69

probably be interpretted gingerly.

Classifying child speech. Our next analysis examined the LENA algorithm's target versus other child tags. Specifically, this analysis investigated tags labeled as children by both the LENA software and manual annotation. Notably, as the current data set only included target children at 6 and 7 months of age (well before word production has begun in even the most precocious talkers) there are no instances of concrete nouns tagged as the target child by human annotators. As a result, this analysis differs from the other confusion matrices, as it can only evaluate LENA agreement for tags labeled as other children by humans. As such, 0/410 of the LENA-generated target child tags were correct, since there were no nouns produced by the target child in the dataset. The LENA system classifies speech from the target child relative to other children with 0.76 accuracy. Recall for this classification is 0.76, while precision is 1, because all LENA-generated "other child" tags were correct. The F1 weighted score is 0.87. See Figure 3.

A logistic regression investigating whether utterance-type accounts for significant variance again found that it did so, here for classifying target vs. other child speech $\chi^2 = (5, N = 1724)$, $p = .001$. As can be seen in in Table 4 and Figure 4, the LENA

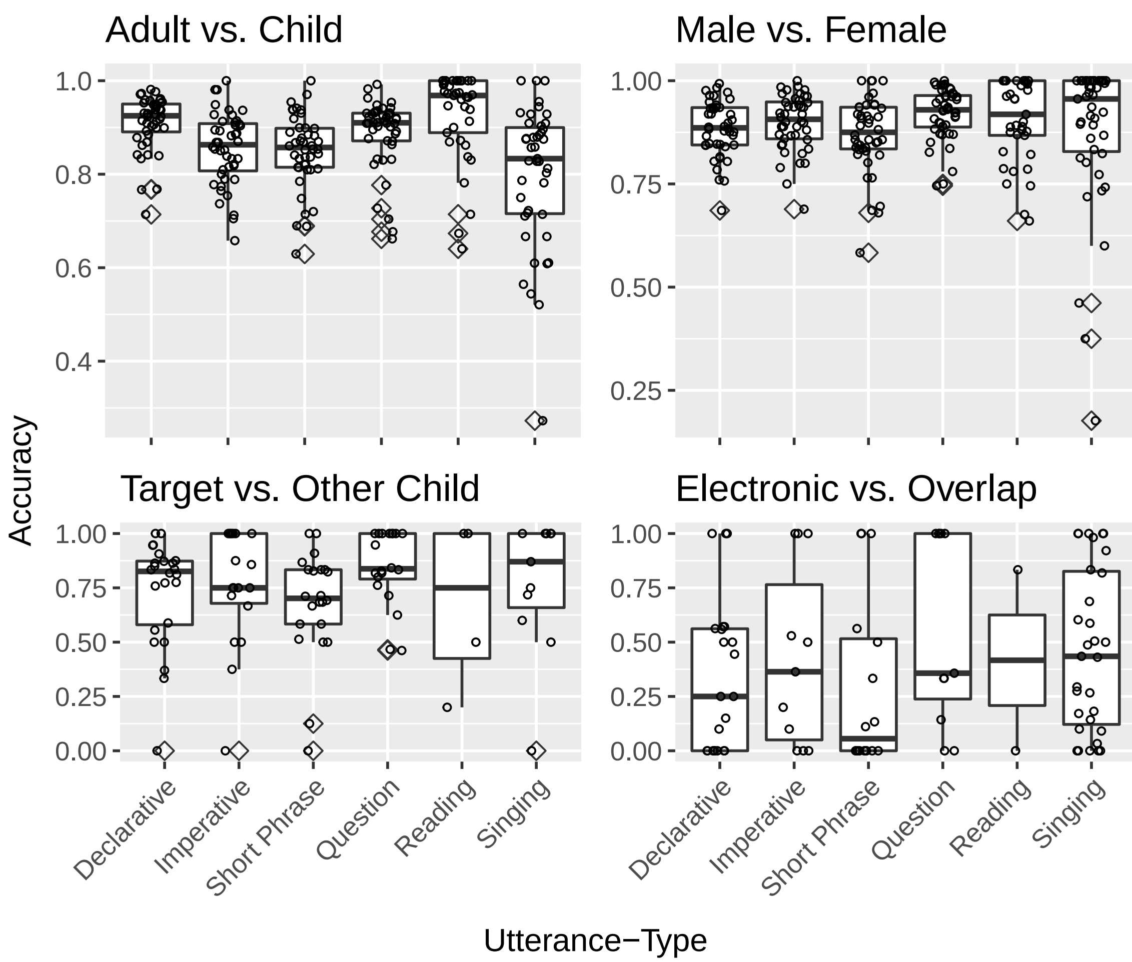

Figure 4. Classification accuracy distribution by utterance-type. Each point (jittered horizontally) represents one child; diamonds (unjittered) indicate outliers. N.B. not all participants contributed data to each utterance type for each comparison.

algorithm was least correct at distinguishing between target and other child speech for words in short phrases (69%), and most correct for questions (80%) and singing (80%).

Classifying electronic and overlap categories. Our next analysis investigated classification accuracy for instances labeled as electronic or overlap by both the LENA system and human coders. Thus, this analysis addresses the LENA system's accuracy at classifying speech coming from a source other than a single live talker. The LENA algorithm classifies speech from the electronic category relative to the overlap category with 0.64 accuracy, Cohen's $\kappa = 0.19$. Recall for this classification is 0.62, while precision is 0.95. The

F1 weighted score is 0.75, see Figure 3.

Electronic vs. overlap speech classification accuracy too was significantly predicted by utterance-type in a logistic regression $\chi^2 = (5, N = 2683)$, $p < .001$. As can be seen in Table 4 and Figure 4, LENA-generated tag accuracy was lowest when distinguishing between electronic and overlap speech for short phrases, and highest for reading (21% and 71%, respectively).

Table 5

Human-generated speaker tag for LENA-generated noise and silence categories

Type	Noise	Silence
Adult	97	322
Child	1	8
Electronic	24	67
Overlap	2	1
Total	124	398

Classifying LENA-generated noise and silence tags. All of the analyses thus far have excluded instances that were classified as noise or silence by the LENA software, which total 522 instances, i.e. 0.90% of the total data. As the trained human coders did not use these categories, we now investigate who was talking when these tags were used by the LENA algorithm. As can be seen in Table 5, the majority of the time the LENA algorithm labeled an utterance as noise or silence it was labeled as an adult utterance by trained researchers. In their technical report, Xu et al. (2009) acknowledge that human coders are likely to be better at identifying human speech in noise, and therefore excluded any tags labeled as noise from their analyses. Our results offer convergent support for this hypothesis, though we note that as only a small portion of data falls in this category (<1%), this seems largely unproblematic for the LENA system's speaker-tag validity.

Establishing Viability of Concrete Nouns as a Proxy for All Input. Given that the current dataset includes only instances of concrete nouns, it is worth assessing whether concrete nouns are a reasonable proxy for language input in the context of talker classification. We first examined noun prevalence within the Brent corpus. We find that 5.60% of utterances in Brent contain a noun (concrete or otherwise) (Sanchez et al., 2019; Brent and Siskind, 2001), and that nouns represent 13.40% of word tokens. Convergently, using LENA's automated Adult Word Count (AWC) estimates as a proxy for word tokens in the current dataset, we find that the concrete nouns we include are ~3.49% of the total word tokens. In order to establish whether concrete nouns are representative of the day-long recordings despite being a small proportion of the input relative to, e.g. function words, we conducted a further series of comparisons. First, the number of concrete nouns we tagged in each recording as produced by adults was strongly correlated with the AWC estimates reported by LENA, Pearson's $r(42) = 0.73$, $p < .001$. Second, the proportion of concrete nouns produced by female (0.79) vs. male speakers (0.21) were highly correlated with the overall proportion of words produced by female (0.71) vs. male (0.29) speakers identified by LENA, Pearson's $r(44) = 0.73$, $p < .001$. Third, we find that the distribution of utterance-types are convergent with those reported by Soderstrom et al. (2008), who used similar utterance-type categories to analyze speech between mothers and preverbal infants. Finally, the talker-tags we use from LENA were for full utterances that included the concrete nouns that the human tags were based on. Together, this raises our confidence that this subset of the data is representative of the sample as a whole.

Discussion

In the current work, we investigated the LENA algorithm speaker tag accuracy in a sample of 44 North-American infants at 6 and 7 months of age. LENA-generated speaker tags for all instances of concrete nouns spoken to the infants were compared to manual speaker tags generated by trained human annotators well familiar with each family, who listened to the recordings in chronological order as the day unfolded. Consistent with previous validations of the LENA software, we found moderate overall agreement between the human-generated codes and the LENA-generated codes, even when limiting our analyses to utterances containing a specific early-produced part of speech: concrete nouns. To summarize, accuracy on the four way comparison (adult, child, electronic, overlap) was reasonably strong (0.72), while accuracy was quite good for the adult and child comparison (0.90) and the male and female comparison (0.90). While overall performance was reasonably strong for the target vs. other child comparison (0.76), its worth reiterating that one category (target child) was only used in error by the software. Finally, accuracy was relatively less strong for the comparison between electronics and overlap (0.64), a notably difficult distinction. It's also noteworthy that despite moderate accuracy overall, there was a very large range of accuracies across the different categories we examined. This merits further investigation in future validation efforts, and ideally, in further iterations of language environment analysis algorithms, which may fruitfully take into account a broader range or larger contiguous stretches of time within the training data.

Across all four main categories (adult, child, electronic or overlap), the LENA software was most accurate at classifying adult speech as adult speech, and was overly reliant on the overlap category. Indeed, our human ability to ignore noise is remarkable, and unsurprisingly difficult for automated analyses: this was clearly acknowledged by Xu et al. (2009) in the original LENA Technical Report. Overreliance on the overlap category was also particularly notable in the electronic vs. overlap comparisons, where speech coded as electronic by

Human coders was labeled as overlap by the LENA software 40% of the time. As also noted by Xu et al. (2008), differentiating electronic speech from human speech can be quite challenging, particularly with improving digital media in recent years. Speculatively, since the LENA system's central goal is to capture human speech, it's possible the system is less well-tuned or trained to electronic sound detection, which may also be sparser or less consistent across instances and recordings. This may in turn lead to overuse of the "overlap" category, especially since if the child wearing the recorder is interacting with electronic sounds, they are likely also generating noise themselves (either vocally, or in playing with e.g. an iPad). Future research is needed to understand what factors might impact electronic vs. overlap errors, e.g. loudness, especially given increasing research centered on understanding children's media use (Christakis et al., 2009).

Throughout the results above, Cohen's κ values were often lower than accuracy. This is almost certainly due to the predominance of certain categories across our comparisons. For example, as base rates for different speakers and speaker categories vary, tagging every single utterance as "adult" would result in >50% accuracy. In contrast, κ values account for this sort of bias in the underlying data distribution when assessing performance.

We found lower overall agreement relative to previous validations of LENA's proprietary software. One possible explanation for this is that we used a larger amount of data than previous validation efforts, and that LENA software's accuracy falls off over longer samples, perhaps due to the wider variability in acoustic environments and situations such lengthy samples engender. For example, the original Lena Technical Report (Xu et al., 2008) analyzed one hour of data from 70 participants, while we analyzed an average of 10 hours of data, from two separate days, for 44 infants, resulting in a difference of 70 hours vs. ~880 hours. Relatedly, in the current corpus, the number of speakers ranges from 4-22 across participants, which may reduce accuracy by introducing larger ranges of non-systematic acoustic variability. While we do not know the number of speakers present in previous corpora used for LENA system validations, given the shorter samples used, it was likely fewer than considered here. The demographic characteristics of our participant sample also differed from those reported in the Lena Technical Report (Xu et al., 2009), specifically with respect to mother's education, which was more variable in the original technical report. While we find it unlikely that this would have a large impact on our results, wider validation efforts with more representative populations would be an important and welcome addition to this literature.

Our further classification comparisons revealed more details about the error patterns made by LENA's proprietary software. The algorithm was found to be highly accurate for classifying adult and child speech, and male and female speech, though when it did make mistakes it was more likely to misclassify a child as an adult and female speech as male speech than the opposite. As it has recently been demonstrated that the LENA system was more likely to classify male speakers as female when they were using child directed speech (Bergelson et al., 2018b), it is possible that these errors patterns reflect register differences used by the speaker. This may also extend to classifying child speech as adult speech; as children have been shown to adapt their speech based on their interlocutors (Syrett and Kawahara, 2014; Tomasello et al., 1984), children speaking to adults may sound more adult-like. While the LENA system does not currently tag child directed speech vs. adult directed speech, this would be a fruitful future direction for algorithmic approaches (cf. Schuster et al., 2014).

In contrast, classification of child speech (target child vs. other children) was relatively inaccurate, particularly given the age of the target child (which is information the LENA system gathers before data processing). Specifically, the algorithm misclassified 410 tokens of speech produced by other children as being produced by the target child. By limiting our sample to just infants at 6 and 7 months of age, we could be sure that the target children in our sample were not producing words, much less concrete nouns which were the focus of the current dataset. Nonetheless, as the misclassified tokens make up 24% of tokens classified as children by either human coders or the LENA system in the current sample, it is important for future research to be aware of these types of mistakes, particularly when the age range of participants varies widely and it is likely that some portion of participants are not yet producing words and contributing to the conversation. To be fair, the LENA algorithm seeks to tag all child vocalizations, not just words. By focusing only on utterances containing words (and not e.g. babble), we limit our assessment of LENA's target vs. other-child tag accuracy to a lexical context, rather than examining all child vocalizations. Given a large focus on early vocabulary differences across populations, we felt this was a worthwhile analysis to include, but acknowledge that for other research questions, accuracy when considering the full range of early vocalizations remains important to establish.

One avenue of improvement in automated analyses would be a way to take the target child's vocal maturity into account more explicitly, or, complementarily, adding an explicit parameter that incorporates family-provided information about how many children are in the recordings. This may be particularly relevant for gathering accurate information about language input from families with more children, or in which caretaking responsibilities include other children (as is particularly the case for low-SES homes, United States Statistics Division, 2015).

Across all comparisons, we also found that utterance-type significantly predicted accuracy, though ranges in accuracy were too tight in some cases to merit interpretation. For the four-way comparison (across adult, child, electronic, and overlap tags), reading and declarative utterances resulted in the highest classification accuracy, whereas singing and short phrases resulted in the lowest classification accuracy. This pattern was consistent for a subset of other comparisons, likely because reading and declaratives capture a similar set of intonational contours across age and gender. In contrast, singing is intrinsically particularly dynamic in pitch and contour. Thus, while we did not find wholly consistent results across utterance-types across comparisons, these results do highlight an explanatory role for utterance-type in classification accuracy. This is important for researchers to keep in mind, as a benefit of the LENA software is that it allows for day-long audio recordings, which are inevitably going to contain variability in utterance-types.

Returning again to our focus on concrete nouns, it remains in principle possible that this would systematically reduce accuracy in talker tags. However, the analyses above suggest that concrete nouns are representative of utterance-type and adult word count distributions more broadly. Furthermore, given the virtual unavoidability of nouns in conversational speech, and the prevalence of concrete nouns in input to infants (Bergelson et al., 2018a; Roy et al., 2015), we believe that a high proportion of speech segments used in previous validations is also likely to contain concrete nouns. Thus, one contribution of the present work is that we provide results at the daylong scale, across a large range of talkers, in a specific lexical class.

Practical implications

To conclude, we want to first reiterate the difficulty faced by speech processing softwares, and the ways the LENA software has revolutionized the field of language acquisition. Without LENA, collecting and processing naturalistic recordings of children's daily environments would be impossible for many researchers. Despite the immense benefits, we have identified some limitations of the LENA talker tags, which researchers may want to consider when deciding whether human annotations are necessary to accurately address their research questions.

For researchers interested in the *relative* proportions of speech produced from males or females, or even from adults and children, the output created by the LENA software is likely sufficiently accurate without a need to manually annotate the input.

In contrast, for researchers interested in child vocalization counts or conversational turns between caregivers and the target child, manually checking target child vocalizations may be necessary to draw valid conclusions. While the restricted age range in the current data set does not allow us to explore whether utterances produced by the target child are mislabeled as produced by other children, it is reasonable to believe that this classification error is bidirectional, particularly as target children get older. Future research is needed to continue to understand this error pattern, and whether it is more likely to occur in specific contexts (child directed vs. adult directed speech, reading vs. singing, louder vs. quieter environments, etc.).

The overreliance on the overlap category may be particularly problematic for researchers interested in the presence of electronics in the input. Considering a large proportion of electronic speech in the current dataset was mistaken for overlap, the proportion of electronic input may be largely underestimated.

One other limitation of the LENA software generally is that it does not identify individual speakers, and effectively collapses across all adult speakers of the same perceived gender, and all non-target children. As such, researchers interested in the number of talkers present in the input, the amount of speech produced by different talkers, or comparing talker variability between- and within- families will need to manually code the input to obtain this type of information.

Lastly, we want to draw attention to how these results might impact other automatic measurements produced by LENA, such as adult word counts and child vocalization counts. As we found that the LENA software was quite accurate at identifying adult relative to child speech, overall adult word counts estimates reported by LENA are likely to be largely unaffected by mistakes in classification accuracy. As noted above, we did not include any child vocalizations which were not concrete nouns, and thus we cannot speak to the accuracy of the LENA system identifying child vocalizations broadly construed. However, the errors found here for identifying target child speech suggest that child vocalization counts may be inflated, particularly for younger children.

Taken together, the analyses presented in the current manuscript reiterate the moderate reliability of the LENA software, while also highlighting patterns of mistakes that researchers should keep in mind as they use the LENA system to collect naturalistic day-long recordings. Knowing about the types of systematic errors the software is likely to make allows researchers to focus their efforts on manually annotating variables of interest, while trusting the software to automate the rest of the process. Despite these error patterns, we maintain that the LENA system has more advantages than drawbacks, and remains a revolutionary data collection tool.

References

Aust, F. and Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. R package version 0.1.0.9842.

Bache, S. M. and Wickham, H. (2014). *magrittr: A Forward-Pipe Operator for R*. R package version 1.5.

Bergelson, E. (2017). Bergelson Seedlings HomeBank Corpus.

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., and Tor, S. (2018a). Day by day , hour by hour : Naturalistic language input to infants. *Developmental Science*, (June):1–10.

Bergelson, E. and Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, page 201712966.

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., and Amatuni, A. (2018b). What Do North American Babies Hear? A large-scale cross-corpus analysis. *Developmental Science*, (June 2018):1–12.

Braginsky, M., Sanchez, A., and Yurovsky, D. (2018). *childesr: Accessing the 'CHILDES' Database*. R package version 0.1.0.

Braginsky, M., Yurovsky, D., Marchman, V., and Frank, M. C. (2017). *Consistency and variability in word learning across languages.*

Brent, M. R. and Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2):33–44.

Christakis, D. A., Gilkerson, J., Richards, J. A., Zimmerman, F. J., Garrison, M. M., Xu, D., Gray, S., and Yapanel, U. (2009). Audible Television and Decreased Adult Words, Infant Vocalizations, and Conversational Turns. *Archives of Pediatrics & Adolescent Medicine*, 163(6):554.

Debaryshe, B. D. (1993). Joint picture-book reading correlates of early oral language skill. *Journal of Child Language*, 20(02):455–461.

Firke, S. (2018). *janitor: Simple Tools for Examining and Cleaning Dirty Data.* R package version 1.1.1.

from Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., and Hunt., T. (2018). *caret: Classification and Regression Training.* R package version 6.0-80.

Gamer, M., Lemon, J., and <puspendra.pusp22@gmail.com>, I. F. P. S. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement.* R package version 0.84.1.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., and Paul, T. D. (2017). Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis. *American Journal of Speech-Language Pathology*, 26(2):248.

Gleitman, L. R., Newport, E. L., and Gleitman, H. (1984). The current status of the motherese hypothesis. *Journal of Child Language*, 11(01).

Gogate, L. J. and Hollich, G. (2010). Invariance detection within an interactive system: a perceptual gateway to language development. *Psychological review*, 117(2):496–516.

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., and Gilkerson, J. (2011). Assessing Children's Home Language Environments Using Automatic Speech Recognition Technology. *Communication Disorders Quarterly*, 32(2):83–92.

Hart, B. and Risley, T. R. (1995). Meaningful differences in the everyday life of America's children. *Baltimore, MD: Paul Brookes.*

Henry, L. and Wickham, H. (2018). *purrr: Functional Programming Tools*. R package version 0.2.5.

Johnson, E. K., Seidl, A., and Tyler, M. D. (2014a). The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PloS one*, 9(1):e83546.

Johnson, K., Caskey, M., Rand, K., Tucker, R., and Vohr, B. (2014b). Gender differences in adult-infant communication in the first months of life. *Pediatrics*, 134(6):e1603–10.

Jusczyk, P. W., Pisoni, D. B., and Mullennix, J. W. (1992). Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition*, 43:253–291.

Kassambara, A. (2018). *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.2.

Lehet, M., Arjmandi, M. K., Dilley, L. C., Roy, S., and Houston, D. (2018). Fidelity of automatic speech processing for adult speech classifications using the Language ENvironment Analysis (LENA) system. *Proceedings of Interspeech*, pages 3–7.

Liberman, A. M., Coopers, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6).

Lieberman, P. (1967). Intonation, perception, and language. *M.I.T. Research Monograph*, xiii(210).

Macwhinney, B. (2019). Tools for Analyzing Talk Part 2: The CLAN Program.

MacWhinney, B. and Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprachsforschung : Online-Zeitschrift zur verbalen Interaktion*, 11:154–173.

Mccauley, A., Esposito, M., and Cook, M. (2011). Language Environment Analysis of Preschoolers with Autism: Validity and Application. *Poster*.

Montag, J. L., Jones, M. N., and Smith, L. B. (2015). The Words Children Hear. *Psychological Science*, 26(9):1489–1496.

Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, 85(1):365–378.

Müller, K. (2018). *bindrcpp: An 'Rcpp' Interface to Active Bindings*. R package version 0.2.2.

Nelson, D. G. K., Hirsh-Pasek, K., Jusczyk, P. W., and Cassidy, K. W. (1989). How the prosodic cues in motherese might assist language learning. *Journal of Child Language*, 16(01):55.

Nelson, K. (1973). Structure and Strategy in Learning to Talk. *Monographs of the Society for Research in Child Development*, 38(1/2):1.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richards, J. A., Xu, D., Gilkerson, J., Yapanel, U., Gray, S., and Paul, T. (2017). Automated Assessment of Child Vocalization Development Using LENA. *Journal of Speech Language and Hearing Research*, 60(7):2047.

Robinson, D. and Hayes, A. (2018). *broom: Convert Statistical Analysis Objects into Tidy Tibbles*. R package version 0.5.0.

Rost, G. C. and Mcmurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, 12(2):339–349.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., and Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences of the United States of America*, 112(41):12663–8.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., and Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, pages 1–14.

Schuster, S., Pancoast, S., Ganjoo, M., Frank, M. C., and Jurafsky, D. (2014). Speaker-independent detection of child-directed speech. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 366–371. IEEE.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27:501–532.

Soderstrom, M., Blossom, M., Foyg El, R., and Morgan, J. L. (2008). Acoustical cues and grammatical units in speech to two preverbal infants*. *Journal of Child Langauge*, 35:869–902.

Soderstrom, M. and Franz, W. (2016). Comparing Human- and Machine- Annotated Language Input Across Childcare Settings. *Talk presented at ICIS 2016*.

Soderstrom, M. and 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.

Sosa, A. V. (2016). Association of the Type of Toy Used During Play With the Quantity and Quality of Parent-Infant Communication. *JAMA Pediatrics*, 170(2):132.

Syrett, K. and Kawahara, S. (2014). Production and perception of listener-oriented clear speech in child language. *Journal of Child Language*, 41(06):1373–1389.

Taine, H. (1876). Note sur l'acquisition du language chez les enfants et dan l'espece humaine. *Revue Philosophique de la France et de l'Ètrangers*, pages 5–23.

Tomasello, M., Farrar, M. J., and Dines, J. (1984). Children's Speech Revisions for a

Familiar and an Unfamiliar Adult. *Journal of Speech, Language, and Hearing Research*, 27(3):359–363.

Trehub, S. E., Unyk, A. M., and Trainor, L. J. (1993). Maternal singing in cross-cultural perspective. *Infant Behavior and Development*, 16(3):285–295.

United States Statistics Division (2015). Birth rate in the United States in 2015, by household income.

Vandam, M. and Silbert, N. H. (2016). Fidelity of Automatic Speech Processing for Adult and Child Talker Classifications. *PLoS ONE*, 11(8):1–13.

Weisleder, A. and Fernald, A. (2013). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science*, 24(11):2143–2152.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.

Wickham, H., François, R., Henry, L., and Müller, K. (2018). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.6.

Williams, H. M. (1937). An analytical study of language achievement in preschooler children. *University of Iowa Studies in Child Welfare*, 13:9–18.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xu, D., Yapanel, U., and Gray, S. (2009). Reliability of the LENA Language Environment Analysis System in Young Children ' s Natural Home Environment. *LENA Foundation Technical Report*, (February):1–16.

Xu, D., Yapanel, U., Gray, S., and Gilkerson, J. (2008). Signal processing for young child speech language development. *Wocci: First Workshop on Child, Computer and Interaction.*

Zhu, H. (2019). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.0.1.