Fidelity of automatic speech processing for adult speech classifications using the Language ENvironment Analysis (LENA) system

Matthew Lehet[1], *Meisam K. Arjmandi*[1], *Laura C. Dilley*[1], *Somnath Roy*[2], *Derek Houston*[3]

[1]Michigan State University, Communicative Sciences and Disorders Department, USA
[2]Centre for Linguistics, Jawaharlal Nehru University, India
[3]Ohio State University, Wexler Medical Center Department of Otolaryngology, USA

lehetmat@msu.edu, khalilar@msu.edu, ldilley@msu.edu, somnathroy86@gmail.com, derek.houston@osumc.edu

Abstract

The Language ENvironment Analysis (LENA) system is a wearable audio recorder that collects daylong recordings; it identifies and classifies speech, providing automated measures of adult word counts and other vocalization metrics. Clinicians and researchers adopted LENA for analysis of the at-home language and acoustic environments of children at risk for speech-language delay or disorder. A primary issue for researchers and clinicians is the reliability of LENA derived speech classification and adult word count (AWC). We tested classification and AWC reliability in LENA recordings from 15 families with young children who were typically developing, hearing impaired with a hearing aid, or were profoundly deaf with a cochlear implant. The analysis focused on samples of audio classified by LENA as containing speech and samples classified as non-speech within one recording from each family. Human listeners identified start and end points of speech by adult female or male talkers and child vocalizations, as well as the number of words produced by adult talkers during speech intervals. Our results suggest marginal reliability for LENA's classifications, with approximately a 2:1 ratio of adult speech found vs. missed. These results suggest that LENA's automatic measures of AWC and other vocalization metrics should be interpreted cautiously.

Index Terms: Automatic speech recognition (ASR), human-computer interaction, Language ENvironment Analysis (LENA) system, computational paralinguistics

1. Introduction

1.1. Variability in children's language and auditory environments, and effects on language development

Young children's speech and language development depends in part on frequent daily experience with speech-language input in their auditory environment [1-3]. Children can experience dramatically different numbers of words produced by adult caregivers, e.g., as a function of household socioeconomic status, where this impacts the rate of language development [4, 5]. In turn, the amount of speech heard by young children affects their rate of language growth.

Given findings of variability in the amount of speech experienced by children, there is considerable interest and value in identifying means of efficiently collecting accurate estimates of how much high-quality adult speech is experienced by young children on a daily basis. Knowing about a child's language environment could be particularly beneficial for children at-risk for speech-language disorder or delay, such as children with moderate to profound hearing loss [6, 7]. Of particular interest is the language development of children who receive cochlear implant surgical intervention for severe to profound hearing loss [8, 9]. Cochlear implants are auditory prostheses, which perform a frequency analysis on auditory input; they transmit frequency information by way of an electrode array, surgically implanted into the cochlea, that provides electrical stimulation directly to the auditory nerve. Understanding how much speech children experience– especially those with cochlear implants – can provide valuable information for clinicians and parents in planning to maximize their speech-language outcomes.

1.2. The LENA system: An automatic speech processing (ASP) device for understanding children's language environments

Automatic speech processing (ASP) technology has been applied in recent decades to understanding variability in speech and auditory input in children's environments [10]. A proprietary ASP device that has gained prominence among clinicians and researchers interested in children's language development is known as the Language ENvironment Analysis (LENA™; LENA Research Foundation, Boulder, CO) system [6, 11-17]. The LENA system consists of a body-worn audio recorder designed to be worn unobtrusively on the body of a child, together with ASP processing software. The system hardware has been designed to collect audio recordings of up to 16 hours.

Audio collected by LENA's hardware is uploaded to a computer and processed off-line in several steps. The software segments and assigns labels using standard methods like Gaussian mixture and hidden Markov models. In the first step, the audio is segmented into short audio portions (typically ~600-1000 ms in length) based on extraction of acoustic features used to partition the audio stream. In a subsequent step, these short audio portions are assigned a preliminary sound category classification consisting broadly of one of two types of categories: (i) vocal tract activity by humans (speech by adult female or male speech and child vocalizations; and (ii) other sound events, including environmental sounds (like noise or television) and silence. After that, the goodness-of-fit associated with the preliminary sound category classification is compared with the goodness-of-fit associated with a Silence model. If the latter fit is better than the former, then the preliminary sound category classification is replaced with a corresponding "Faint" or secondary category; this step is meant

to distinguish sound events which are in the near field from "faint" sounds in the far field.

In a final step, the software uses the classifications of the short audio portions to group together successive short audio portions into alternations of *conversational blocks* and *pause units*. A conversational block corresponds to a successive sequence of short audio portions identified as near-field human vocal tract activity separated by less than five seconds. In contrast, each pause unit is a stretch of audio bounded by a pair of conversational blocks separated by over five seconds.

Importantly, short audio portions within conversational blocks that are classified as near-field human vocal tract activity are used to derive a number of estimates of the child's language environment. In particular, short audio portions classified as near-field adult male or adult female speech are used to derive automated estimates adult word count. The extent to which LENA derives an accurate automatic estimate of the number of adult words in a child's environment therefore depends on the extent of its accuracy in classifying short audio portions as speech – specifically, as adult speech. The present paper assesses accuracy of LENA's classifications of short audio portions as speech (including adult vs. child speech), as opposed to not speech.

1.3. Prior studies of LENA reliability and accuracy

Several studies of reliability of LENA's automatic language estimates have been reported [18-21]. However, most of these studies did not appeared in peer-reviewed research journals and/or certain study details lead to questions about its method of assessing reliability/accuracy and/or generalizability of the results. For example, in a well-cited unpublished technical report from LENA Corporation, Xu et al. [19] stated "false negative classifications…were simply excluded from the final estimates" of accuracy (p. 4). Another study by Canault et al. [20] reported reliability based on calculations made from non-independent sample subsets and indicated considerable variability in accuracy across recordings.

Even fewer studies have focused on LENA's accuracy, as opposed to its reliability [21]. Existing studies lend little insight into how classification accuracy might vary across different kinds of distinctions which could impact LENA's automatic language estimates. (See also [22].) LENA's accuracy in classifying speech vs. non-speech has not been carefully evaluated. Most especially for LENA's automatic estimates of word count, LENA's accuracy in classifying adult speech has not been evaluated.

1.4. The current study: LENA classification accuracy for speech, specifically adult speech

LENA's ability to derive accurate automatic developmental language estimates depends on its ability to *accurately classify* short audio portions as adult speech and/or as communicative child vocalizations. Any failure of the LENA ASP system to correctly classify speech as speech – or adult speech as adult speech – will lead to error in its developmental language estimates.

For the present study, we focused on LENA's accuracy in classifying speech as speech, and its accuracy in classifying adult speech as adult speech. Our approach prioritized "broad" sampling from many different participants to obtain an understanding of variability in accuracy across recordings and home language environmental contexts. We employed a method whereby human listeners judged the start and end times of adult speech and of communicative child vocalizations, thereby providing ground-truth determinations of when meaningful human communications were occurring as distinct from all other kinds of auditory events and environments. We then evaluated accuracy of LENA's classifications of speech (vs. non-speech) and of adult speech (vs. all other sound events and environments).

2. Method

2.1. Participants

LENA samples for the present study came from a database of LENA recordings made as part of an ongoing project at the Ohio State Wexler Medical Center Department of Otolaryngology on the impacts of home environmental input on language development in children with hearing loss. Participating families in the broader project agreed to their child wearing a LENA system for one or more days as part of a study of their home language environment. The present sample came from a single day-long recording from each of 15 randomly-selected participating families which had a target child aged 7 – 33 months at the time of the recording. Target children in participating families had a range of hearing statuses: 8 had a cochlear implant, 5 had hearing aids, and 2 were normal hearing. Target children with cochlear implants had 3 – 21 months post-implantation hearing experience.

2.2. Audio selection

From within each recording, speech was selected from a range of points during the day for analysis, according to the following procedure. First, the first and last 30 conversational blocks of the recording that occurred while the child was awake were included. Next, if either the first or last 30 conversational blocks totaled less than 10 minutes of speech from the beginning or end of the file, additional conversational blocks were included from the beginning or end, respectively, until a minimum of 10 minutes of selected speech was reached. Additionally, short, randomly selected 5-second portions of audio file that LENA had coded as pause units totaling at least 2 minutes were selected for analysis for each selected recording. Across all selected recordings, we analyzed a mean of 30.2 minutes of audio per participant family (SD = 6.3 minutes; range = 23.2 – 51.8 minutes).

2.3. Automatic analysis of speech by LENA

LENA automated analysis was exported for all conversational blocks within each sample. Segments of speech were labeled by LENA as female adult near (FAN), male adult near (MAN), key child (CHN), other child (CXN), overlapping vocals (OLN), TV/electronic media (TVN), noise (NON), silence (SIL), and uncertain/fuzzy (FUZ). The automatic segmentation by LENA inserted into a Praat [23] TextGrid that human coders could reference as they analyzed the speech content in each sample.

2.4. Analysis of speech by human listeners

Human listeners used Praat to demarcate when adult male, adult female, and child vocalizations occurred within each sample. In cases where LENA coded speech as overlap (OLN), and there was overlapping noise to interfere with signal processing, the segment was marked for exclusion.

2.5. Comparison between LENA and human listeners

In order to characterize LENA's classification accuracy, the analyzed audio was split into 50 ms frames. For each frame, the segment code given by LENA was compared to the classification for that frame given by the human coder. In cases where a bin straddled two labels, the label associated with the larger temporal proportion of the bin was taken to characterize the entire frame. In our analysis, we focus on LENA's accuracy in classifying frames which humans coded as consisting of human communicative vocal activity by adults and/or children. Frames which humans classified as containing human communicative vocalizations should be coded by LENA as either FAN or MAN (for adult female or adult male speech, respectively) or as CHN or CXN (for communicative vocalizations by the target child or other child, respectively).

3. Results

3.1. LENA performance in the 4 by 8 classification

We first assessed LENA's accuracy in classification relative to human listeners. Table 1 presents the number of 50 ms frames classified by human listeners as adult female speech, adult male speech, child vocalizations, or other (on rows) and the corresponding LENA classification code for these frames (in columns). Overall, LENA accurately classified 70.5% of frames.

3.2. Four-way classification of speech

Next, LENA codes were collapsed to four categories, and the percentage of frames accurately classified for each family was calculated. Table 2 shows percentages for each category averaged across participants. The mean percentage of correct classifications for human communicative vocalizations ranged from 60% for female adult speech to 64% for male adult speech.

3.3. Speech vs. non-speech binary classification

Table 3 reports accuracy for two categories: speech (adult female speech, adult male speech, or child vocalizations) vs. nonspeech. Table 3 shows that human identified speech and non-speech frames were classified as speech or non-speech by LENA an average of 76% and 78% across participants, respectively. This binary classification scheme permitted expressing accuracy for speech vs. nonspeech classifications as an *odds ratio*, namely, the ratio of percentage of correct to incorrect classifications by LENA. For example, Table 3 shows that, LENA correctly classified 76% of frames classified by human coders, and incorrectly classified nonspeech" 24% of the time, for an odds ratio of 3.1:1. There was considerable variability in LENA's accuracy of the speech vs. nonspeech classification. For human identified speech, LENA's classification accuracy expressed as an odds ratio ranged from 1.8:1 to 6.5:1; conversely, for ground-truth nonspeech, LENA's classification accuracy expressed as an odds ratio ranged from 1.3:1 to 10.1:1.

3.4. Adult speech vs. not adult speech binary classification

Table 4 reports accuracy for two categories: adult speech (female or male) vs. not adult speech (i.e., all other kinds of human or environmental sound code). Table 4 shows that ground-truth adult speech, as identified by human coders, was correctly classified as adult speech by LENA an average of 68% of the time across participants (odds ratio 2.1:1), while frames identified as not consisting of adult speech were accurately classified by LENA an average of 89% of the time across participants (odds ratio 8.1:1). For ground-truth adult speech, LENA's classification accuracy expressed as an odds ratio ranged from 0.9:1 to 4.8:1; conversely, for cases identified ground-truth as not adult speech, LENA's classification accuracy expressed as an odds ratio ranged from 1.8:1 to 33.1:1.

3.5. Individual variability across participant recordings

In order to examine the variability in frame-based accuracy across families and recordings, the odds ratios for all recordings in the binary speech vs non-speech (Figure 1) and adult speech vs not adult speech (Figure 2) were rank ordered. These plots highlight the degree of variability across families, and across classification types.

Table 1: *Confusion matrix comparing classifications by human listeners (rows) and classifications by LENA (columns). Cells are counts of 50 ms frames, aggregating across all selected portions of recordings for all subjects. Bold-face cells are accurate classifications, given the four-category human coding scheme that was employed. See text.*

		LENA Classifications								
		FAN	**MAN**	**CHN *or* CXN**	**NON**	**OLN**	**TVN**	**FUZ**	**SIL or "faint"**	**Totals**
Human classifications	**Female adult speech**	**50270**	5276	10935	212	6294	1852	6813	5458	87110
	Male adult speech	9422	**26869**	1862	51	1637	870	2966	2464	46141
	Child vocalization	7871	634	**58880**	249	7367	757	5049	10288	91095
	Other	18937	14270	38133	**3032**	**24333**	**8561**	**51737**	**156953**	315956
	Totals	86500	47049	109810	3544	39631	12040	66565	175163	540302

Table 2: *Mean percentages and standard deviations (in parentheses) across participants of 50 ms frames classified by human listeners as one of four categories (rows) and the corresponding classification by LENA (columns). Bold-face entries on the diagonal are correct classifications.*

	FAN	MAN	CHN/CXN	Other
Female adult	**60 (11)**	5 (8)	11 (9)	24 (9)
Male adult	16 (16)	**64 (21)**	4 (6)	16 (9)
Child	8 (5)	1 (1)	**63 (12)**	29 (11)
Other	6 (5)	5 (9)	11 (8)	**78 (10)**

Table 3: *Mean Percentages and standard deviations (in parentheses) across participants of 50 ms frames for the binary distinction of speech (i.e. adult male, adult female, and child vocalizations) vs non-speech for human listeners (rows) and LENA (columns). Odds ratios are also presented for the mean percentages with standard deviations in parentheses. The minimum and maximum odds ratios across participants are also presented.*

	Speech	Non-Speech	Odds Ratio (SD)	Odds Ratio (Min,Max)
Speech	**76 (6)**	24 (6)	3.1 (1.2)	1.8, 6.5
Non-Speech	22 (10)	**78 (10)**	3.5 (2.6)	1.3, 10.1

Table 4: *Mean percentages and standard deviations (in parentheses) across participants of 50 ms frames for the binary distinction of adult speech (i.e., adult female or adult male speech) vs. not adult speech, for human listeners (rows) and for LENA (columns). Odds ratios are also presented for the mean percentages with standard deviations in parentheses. The minimum and maximum odds ratios across participants are also presented.*

	Adult	Not Adult	Odds Ratio (SD)	Odds Ratio (Min,Max)
Adult	**68 (9)**	32 (9)	2.1 (1)	0.9 / 4.8
Not-Adult	11 (9)	**89 (9)**	8.1 (9.6)	1.8 / 33.1

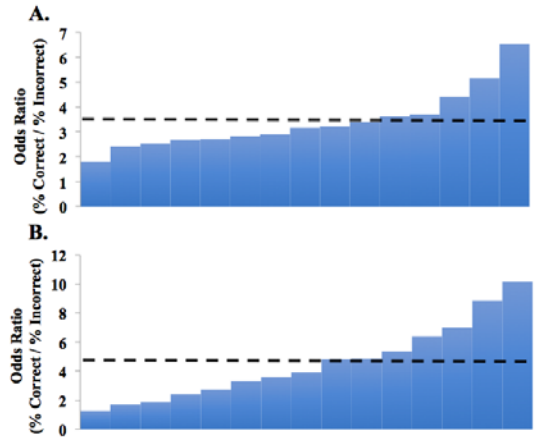

Figure 1: *Odds ratios for speech for each recording rank ordered. A) Odds ratio of detecting frames of speech correctly. B) Odds ratio of correctly identifying frames as non-speech. The dashed lines represent the average odds ratio across participants.*

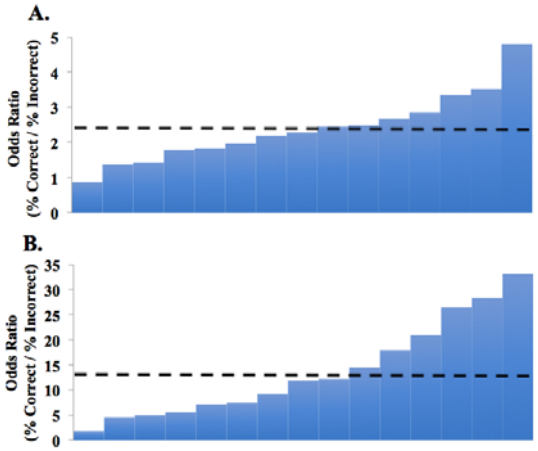

Figure 2: *Odds ratios (% correct / % incorrect) for adult speech for each recording rank ordered. A) The odds ratio of detecting 50 ms of adult speech correctly. B) The odds ratio of correctly identifying a 50 ms frame as not adult speech. The dashed lines represent the average odds ratio across participants.*

4. Discussion

Accurately quantifying the speech environment of young children could assist in maximizing their language outcomes. However, results presented here show that LENA accurately classifies speech vs. non-speech only 76-78% of the time in these samples. For every 3 to 4 frames correctly classified as speech, one was misclassified. LENA's performance was worse when classifying adult speech. LENA correctly classified adult speech (as *FAN* or *MAN*) only 68% of the time. Thus, for every two frames of adult speech correctly classified as adult speech, one was incorrectly classified as not adult speech. This misclassification negatively influences the ability of LENA to generate accurate summary statistics, creating uncertainty in its reliability as a measurement device.

LENA's performance in classification significantly varied across participants. This may indicate that classification by LENA depends upon environmental factors specific to each recording. These findings about variability in LENA's classification accuracy of speech across participants suggests caution in interpreting its automatic language measures.

5. Conclusions

The present work provided evidence that the reliability of LENA software is not consistent across environments. Our findings suggest that LENA is useful for automatic monitoring of speech and language environments but the speech classification abilities need to be tailored to the specific purposes of usage. It remains to be seen what environmental qualities influence reliability, and how classification reliability is related to adult word count and conversational turn counts provided by LENA.

6. Acknowledgements

Research reported in this paper was supported by the National Institutes on Deafness and other Communication Disorders (NIDCD) Grant R01-DC008581 to D. Houston and L. Dilley.

7. References

[1] N. Hurtado, V. A. Marchman, and A. Fernald, "Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in Spanish-learning children," *Developmental Science*, vol. 11, no. 6, pp. F31-F39, 2008.

[2] J. Huttenlocher, W. Haight, A. Bryk, M. Seltzer, and T. Lyons, "Early vocabulary growth: Relation to language input and gender," *Developmental Psychology*, vol. 27, no. 2, pp. 236-248, 1991.

[3] F. J. Zimmerman *et al.*, "Teaching by listening: The importance of adult-child conversations to language development," *Pediatrics*, vol. 124, no. 1, pp. 342-349, 2009.

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

[5] E. Hoff-Ginsberg, "Mother-child conversation in different social classes and communicative settings," *Child Development*, vol. 62, no. 4, pp. 782-796, 1991.

[6] S. Ambrose, E. Walker, L. Unflat-Berry, J. Oleson, and M. P. Moeller, "Quantity and quality of caregivers' linguistic input to 18-month and 3-year-old children who are hard of hearing," vol. 36, no. 1, pp. 48S-59S, 2015.

[7] D. M. Houston, J. Stewart, A. Moberly, G. Hollich, and R. T. Miyamoto, "Word learning in deaf children with cochlear implants: effects of early auditory experience," *Developmental Science*, vol. 15, no. 3, pp. 448-461, 2012.

[8] R. T. Miyamoto, B. Colson, S. Henning, and D. B. Pisoni, "Cochlear implantation in infants below 12 months of age," *World Journal of Otorhinolaryngology-Head and Neck Surgery*, 2018.

[9] J. K. Niparko *et al.*, "Spoken language development in children following cochlear implantation," *Journal of the American Medical Association*, vol. 303 no. 15, pp. 1498-1506, 2010.

[10] M. E. Beckman, A. R. Plummer, B. Munson, and P. F. Reidy, "Methods for eliciting, annotating, and analyzing databases for child speech development," *Computer Speech and Language*, vol. 45, pp. 278-299, 2017.

[11] A. S. Warlaumont, J. A. Richards, J. Gilkerson, and D. K. Oller, "A social feedback loop for speech development and its reduction in autism," *Psychological Science*, vol. 25, no. 7, pp. 1314-1324, 2014.

[12] A. Weisleder and A. Fernald, "Talking to children matters: early language experience strengthens processing and builds vocabulary," *Psychological Science*, vol. 24, pp. 2143-2152, 2013.

[13] M. VanDam *et al.*, "HomeBank: An online repository of daylong child-centered audio recordings. In Seminars in speech and language (Vol. 37, No. 2, p. 128). NIH Public Access.," *Seminars in Speech and Language*, vol. 37, no. 2, pp. 128-142, 2016.

[14] M. VanDam *et al.*, "Automated vocal analysis of children with hearing loss and their typical and atypical peers," *Ear and Hearing*, vol. 36, no. 4, pp. e146-e152, 2015.

[15] M. Caskey, B. Stephens, R. Tucker, and B. Vohr, "Importance of parent talk on the development of preterm infant vocalizations," *Pediatrics*, vol. 128, no. 5, pp. 910-916, 2011.

[16] M. Soderstrom and K. Wittebolle, "When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments," *PLoS One*, vol. 8, no. 11, p. e80646, 2013.

[17] D. K. Oller *et al.*, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13354-13359, 2010.

[18] J. B. Oetting, L. R. Hartfield, and S. L. Pruitt, "Exploring LENA as a tool for researchers and clinicians," *The ASHA Leader*, vol. 14, no. 6, pp. 20-22, 2009.

[19] D. Xu, U. Yapanel, and S. Gray, "Reliability of the LENA(TM) Language Environment Analysis System in young children's natural home environment (LENA Technical Report LTR-05-2)," in "LENA Foundation," Boulder, CO2009, Available: http://lena.org/wp-content/uploads/2016/07/LTR-05-2_Reliability.pdf.

[20] M. Canault, M. T. Le Normand, S. Foudil, N. Loundon, and H. Thai-Van, "Reliability of the Language ENvironment Analysis system (LENA™) in European French," *Behavior Research Methods*, vol. 48, no. 3, pp. 1109-1124, 2016.

[21] T. Busch, A. Sangen, F. Vanpoucke, and A. van Wieringen, "Correlation and agreement between Language ENvironment Analysis (LENA™) and manual transcription for Dutch natural language recordings," *Behavior Research Methods*, 2017.

[22] M. VanDam and N. H. Silbert, " (2016). Fidelity of automatic speech processing for adult and child talker classifications. PloS one, 11(8), e0160588.," *PLoS One*, vol. 11, no. 8, p. e0160588, 2016.

[23] D. C. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 6.0.29 ed, 2017.