LENA™ FOUNDATION

Reliability of the LENA™ Language Environment Analysis System in Young Children's Natural Home Environment

Dongxin Xu, Umit Yapanel, & Sharmi Gray
LENA Foundation, Boulder, CO

LTR-05-2

February 2009

Software Version: V3.1.0

ABSTRACT

The LENA language environment analysis system was designed to provide information about the language environment of infants and toddlers. In this technical report, we describe the reliability of the LENA System in terms of segmentation, adult word counts, and child vocalizations. We also describe unique sources of variability associated with data collection in the natural home environment.

Keywords

Accuracy, adult word count, child vocalizations, LENA language environment analysis system, reliability, segmentation, transcription, variability.

1.0 INTRODUCTION

The LENA language environment analysis software V3.1.0 was developed to process and selectively filter audio and interference signals resulting from a natural data collection environment, that is, audio data recorded directly in a live, not already recorded, environment.[1] The primary goals of the audio data processing are to estimate Adult Word Counts (AWC), Child Vocalizations (CV), and Conversational Turns (CT) between the adult and key child. Here, we establish the reliability and accuracy of the audio processing system as well as possible sources of variability that were derived from the natural language environment in which data were collected.

2.0 ACCURACY OF THE LENA SYSTEM

The spontaneous speech environment in which data for this study were collected is by design natural and without restriction. Traditionally, speaker recognition software is intended for controlled environments; external sounds and events are in general greatly minimized. By contrast, speech recorded by the LENA System is spontaneous, real, unrehearsed, and representative of a child's typical daily language environment.

The LENA software V3.1.0 selectively segments audio into meaningful speech and non-speech then filters out interfering signals present in the naturalistic environment. These exclusions derive primarily from sound segments that are not likely to contribute meaningfully to the child's language environment, such as those described above. Table 1 summarizes the sound categories in a naturalistic environment.

Table 1: Categories of Natural Environment Audio Data.

Live Human Sounds	Background Sounds
Adult Male	Overlapping Speech
Adult Female	Electronic Media (e.g. TV/Radio)
Key Child	Noise
Other Child	Silence

1 Although it is possible to use the LENA system to process pre-recorded audio signals by re-recording the playback of a pre-existing recording, this use is neither recommended nor supported. Acoustic features of a recorded audio signal are markedly different from those of one recorded directly in a live environment. The LENA system was not designed to provide accurate information for pre-recorded audio data.

Segmentation accuracy may be displayed using confusion matrices, in which human transcription and machine-generated classifications are compared visually (Table 2). Data evaluation in matrix format highlights percentage of instances of false positive and false negative classifications. The goal here was to lower the incidence of false positive classifications that inflate the final AWC and CT estimates. False negative classifications were a less serious error, as these were simply excluded from the final estimates, and the quantity of data collected mitigates the impact of such exclusions. In the following sections, we provide a general description of the accuracy of the LENA System audio processing in terms of the segmentation, adult word count and child vocalization identification.

Table 2: Interpretation of data from a 2x2 confusion matrix.

		LENA System	
		Target	**Non-Target**
Human Transcribers	**Target**	**Agreement**	False Negative
	Non-Target	False Positive	**Agreement**

2.1 SEGMENTATION

For the purposes of speaker recognition, audio data recorded by the LENA DLP must be accurately segmented into subcategories (Table 1). LENA software V3.1.0 initially segments the audio file into eight categories: adult male; adult female; key child; other child; overlapping speech; noise (e.g., bumps, rattles); electronic media (e.g., radio or television); and silence categories of speech. The seven categories other than silence are further divided into two types, based on how "near" or "far" each segment is from the statistical model for that category (i.e., how well the model fits for that segment). The segmentation process selectively eliminates noise, unclear (e.g., distant or faint) speech, overlapping speech, electronic media sounds, and child non-speech sounds such as laughing or crying. A language-dependent statistical model was used to estimate the number of words spoken in each clear adult segment without recognizing either the content or meaning of the speech.

As mentioned above, to assess the daily language environment of the child, the processing software must accurately and reliably distinguish between the adult speaker and the key child speaker, as well as distinguish the key child from other children. It is also essential that factors interfering with AWC estimates (e.g., overlapping or unclear speech, distant or faint voices, transient noise such as bumps, and electronic noise such as TV or radio) are identified and selectively eliminated during the audio processing. To assess the accuracy of LENA-based segmentation, segments identified by professional human transcribers were compared to segments identified by the LENA software. For visualization purposes, the accuracy of the LENA System classification is displayed as a confusion matrix (Table 3). Percent agreement between human-transcribed segmentation and LENA-based segmentation are shown along the diagonal. Deviations in agreement are shown in the off-diagonal regions.

An algorithm developed by software engineers from the LENA Foundation Research and Development team was used to select six ten-minute segments from each of 70 audio recording files used in our test set. The algorithm was designed to automatically detect high levels of speech activity between the key child and an adult. Each of the six audio sections were concatenated to form one hour-long audio file. Thus, a total of 70 hours of data were transcribed from the 70 test set files. Transcriber-determined segmentation information and AWC estimates were acquired from these data.

Table 3: LENA System Sensitivity: Segmentation agreement between human transcribers and LENA software V3.1.0.

		LENA System			
		Adult	**Child**	**TV**	**Other**
Human Transcribers	**Adult**	**82%**	2%	4%	12%
	Child	7%	**76%**	0%	17%
	TV	8%	0%	**71%**	21%
	Other	14%	4%	6%	**76%**

The data displayed in Table 3 shows a high degree of agreement between human and LENA-based segmentations. The LENA System and human-transcribed near-field adult segments were identified with agreement 82% of the time. Similarly, LENA and human-transcribed near-field child and television were identified with agreement 76% and 71% of the time, respectively. Differences in agreement among these categories were minor, ranging from 0.1% to 8%. In general, the false negative misclassifications outweighed the false positive misclassifications. Larger variations in identification agreement in the category "Other" are primarily due to the presence of overlapping speech segments.

The LENA software V3.1.0 segmentation algorithm was designed specifically to minimize categorical misclassification. As a result, segments that contained overlapping speech were not classified as either adult or child speech and the LENA processing software excluded these regions. The human auditory system has an innate ability to identify speakers in overlapping speech environments; thus, when professional human transcribers segmented these regions, they could distinguish and appropriately categorize the primary speaker. However, even though an astute professional human transcriber may be able to identify the speaker and process the speech in these overlapping segments, it is not known whether an infant or toddler would be able to distinguish similarly noisy language input. In fact, research indicates that an environment in which overlapping speech is present is not as beneficial for language learning as quieter environments (Poag, Goodnight, & Cohen, 1985; Wachs, 1982). Thus, the exclusion of these segments from categorization could provide a more accurate representation of the child's meaningful language environment.

2.2 ADULT WORD COUNT (AWC)

Adult word count (AWC) is an estimate of the number of adult words spoken near a child per hour or per day. The research and development teams at the LENA Foundation have developed novel instrumentation to estimate AWC in natural, spontaneous speech environments. Data for the reliability assessment were selected from the LENA Natural Language Corpus. As described above, 70 independent 12-hour long audio files were selected using a block randomization scheme to ensure a test sample representative of the entire data set. Test set files were selected on the basis of age (2-36 months) as well as maternal socioeconomic status (SES). Two children were selected per age group, one from a relatively higher SES bracket and the other from a comparatively lower SES bracket. Please refer to Technical Report LTR-06-2 for the demographic distribution of the 70 test set files. One hour (i.e., six 10-minute segments) of each file was transcribed. Thus, the test set contains 70 hours of transcription data.

To assess the accuracy of the AWC estimates, we compared LENA System-detected AWC to the AWC reported by the human transcribers. Results are shown in Figure 1.

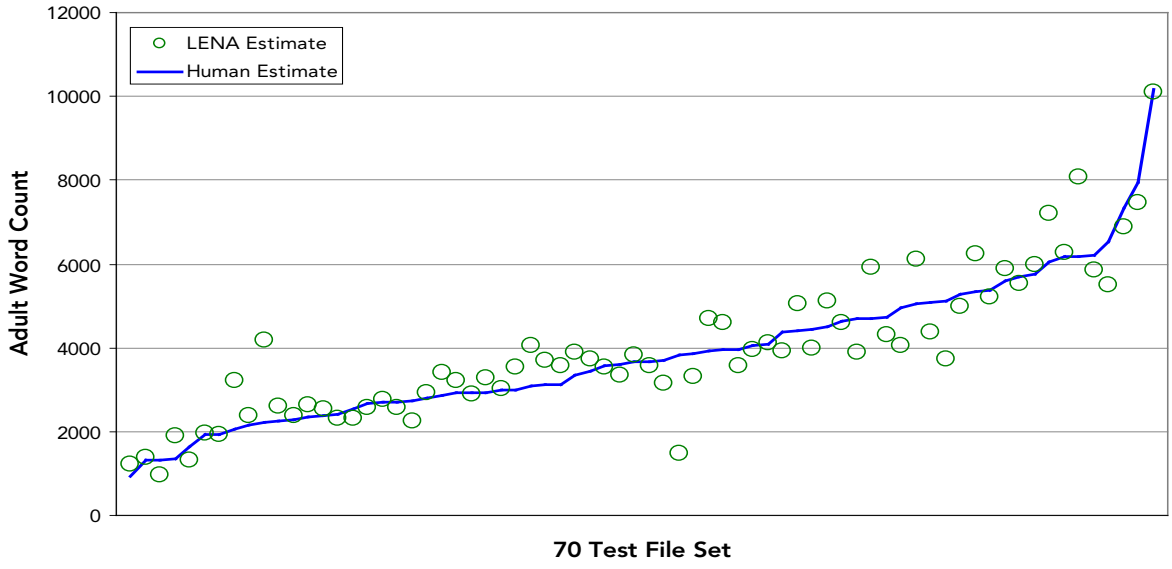

Figure 1. Human and LENA-based AWC estimates for 70 test files.

The data in Figure 1 suggest a significant linear correlation between human and LENA-based AWC estimates. The Pearson product-moment correlation between the two variables was $r = .92$, $p<.01$. The LENA mean word count was on average 2% lower than the word count reported by the transcribers, primarily a result of underestimation (i.e., false negative misclassification). LENA-based AWC estimates are generally lower because LENA System segmentation processes eliminate overlapping speech, possibly providing a more accurate representation of a child's language environment. In addition, the estimated reliability of LENA analyses may be higher than reported due to inherent measurement error present within the human-based transcriptions. Please refer to Technical Report LTR-06-2 for further information on transcription reliability.

To further assess the accuracy of the LENA software V3.1.0, LENA-based and transcribed AWC were compared through complete transcriptional analysis of two twelve-hour sessions randomly selected from the LENA Natural Language corpus. In one file, "File 1 – Typical Quiet Day," the language environment was generally quieter than the language environment on the second file, "File 2 – Typical Active Day." The two 12-hour files were analyzed in their entirety to assess the accuracy of the LENA-based AWCs relative to human-based AWC as influenced by environmental intensity.

The child in File 1 was a 10-month-old male assessed at an average pre-language skill level by the LENA Foundation's speech language pathologist. As shown in Figure 2, LENA-based AWC estimates were similar to human-based AWC. The LENA mean word count estimate was less than 1% lower than the transcriber-reported word count, similar to the difference observed for the 70 test set analysis described above. Notably, human and LENA-based values were consistently similar for the entire duration of the analysis, even in the presence of an alternative dialect of English near the start of the timeline.

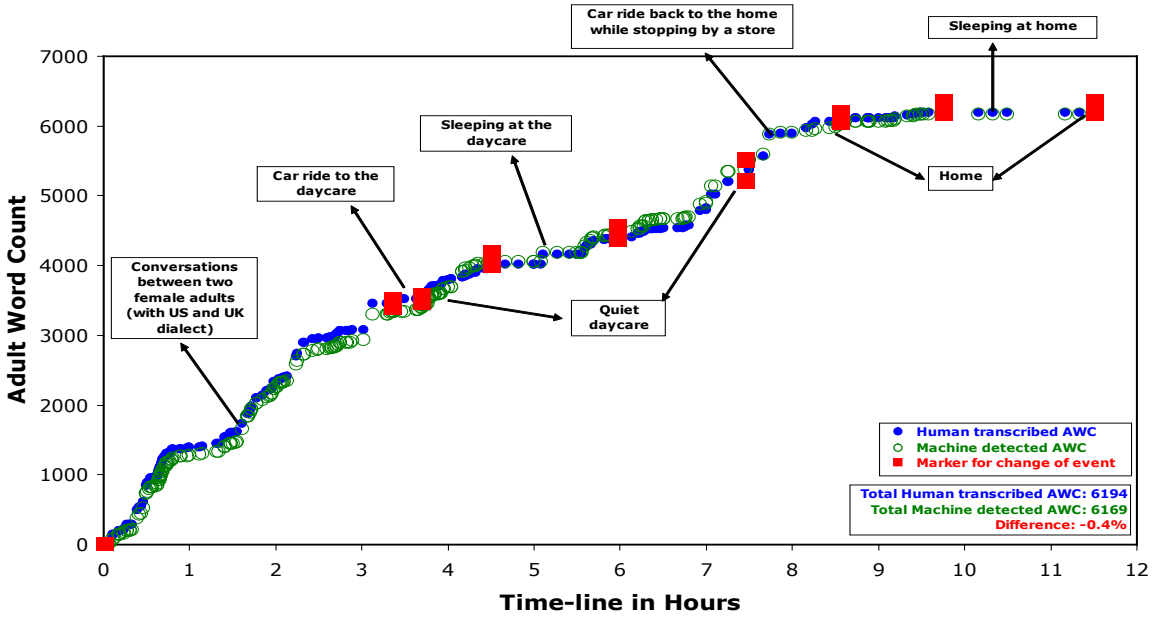

Figure 2. LENA and human-identified AWC spanning a continuous 12-hour recording session in a predominantly quiet environment (File 1 – Typical Quiet Day).

The key child in File 2 was a 31-month old female who was rated at an outstanding language skill level by the LENA Foundation Speech Language Pathologist. The AWC estimate for File 2 revealed an interesting phenomenon in the LENA-based modelling. As shown in Figure 3, the human-transcribed and LENA-based AWC estimates were less highly correlated during the child's participation in outdoor and other less-quiescent activities. As a result, human and LENA-based AWC estimates deviated during these time periods. However, the AWC began to follow human-based estimates once quiescent activity was again resumed. This deviation resulted from the segmentation process. During the strident activities such as those that occur outdoors, variations in human and LENA-derived AWC was due primarily to the presence of multiple overlapping speech segments. Ultimately, LENA-based AWC estimates for the 12-hour period deviated from the human AWC estimates by 27%. Recall that the LENA System was designed to assess the language environment a child is exposed to. As suggested earlier, noisy environments may be less beneficial to the language development of the child; if so, the LENA may appropriately penalize the situation (Poag, et.al., 1985; Wachs, 1982). As a result, lower LENA AWC estimates may better reflect what the child is able to absorb.

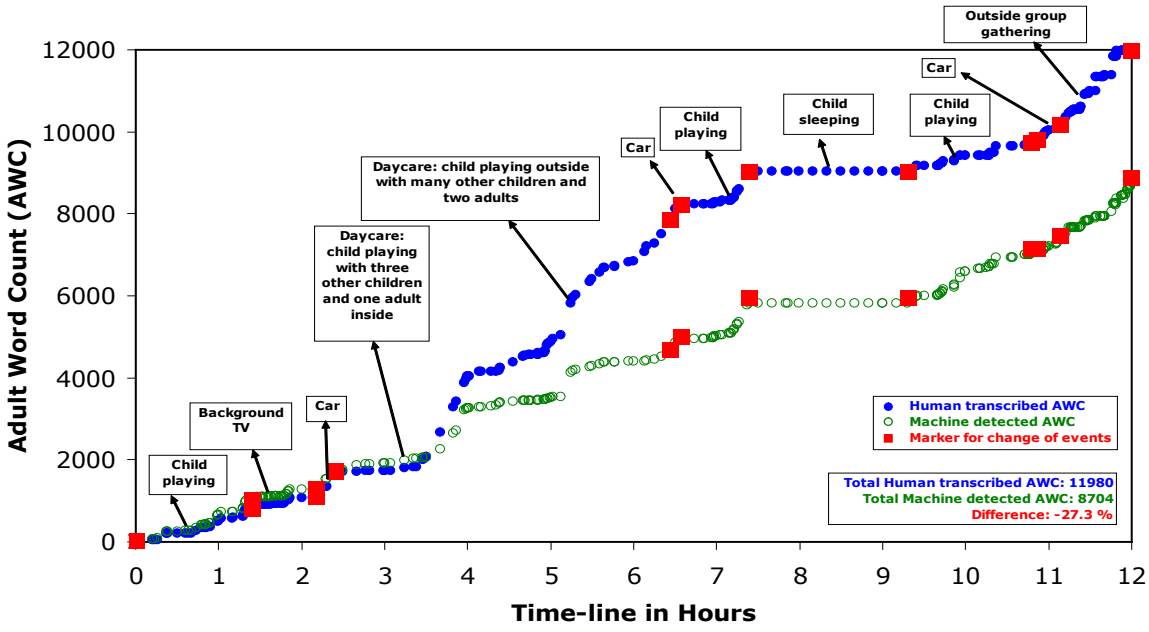

Figure 3. LENA and human-identified AWC spanning a continuous 12-hour recording session in a predominantly noisy environment (File 2 – Typical Active Day).

2.3 CHILD VOCALIZATIONS

Accurate and reliable processing of the audio data is critical to the accurate assessment of the language environment a child is exposed to in a natural state. Of equal importance is the ability of LENA-based algorithms to distinguish between key child speech and key child non-speech. Although this was a challenging task, it was clear that we first had to define what we considered speech and what we considered non-speech. Sounds that were considered speech included words, babbles, and pre-speech communicative sounds or "protophones" such as squeals, growls, or raspberries (see Oller 2000 for more detail on protophones). Sounds considered non-speech were further subdivided into either fixed signals or vegetative sounds. Fixed signals contain sounds that are instinctive emotional reactions to the environment (e.g. cry, scream, laugh). Vegetative sounds are those sounds resulting from respiration (e.g. breathing) or digestion (e.g. burping). Please refer to Technical Paper LTR-06-2 for further information on the child vocalization classifications.

To assess the accuracy of the LENA System's detection of child vocalizations versus non-vocalizations (i.e., crying/vegetative sounds/fixed signals), human and LENA-based segment classifications from the 70 test set files were compared. Only those segments that both the human transcribers and the LENA algorithms agreed were produced by the key child were included in this comparison. As Table 4 shows, the LENA algorithms correctly detected 75% of the human-identified child vocalizations and misclassified them as non-vocalizations 25% of the time. Performance for non-vocalizations was somewhat higher; 84% of these were categorized correctly by the LENA algorithms. Importantly, in only 16% of the cases were child non-vocalizations misclassified as vocalizations.[2]

Table 4: Human and LENA-algorithmic based Detection and Classification of Sound as Either Speech or Non-Speech Key Child Vocalizations.

		LENA System	
		Child Vocalizations	**Child Cry/Veg/Fixed**
Human Transcribers	**Child Vocalizations**	75%	25%
	Child Cry/Veg/Fixed	16%	84%

The primary source of LENA-based misclassification is the lack of context present in the algorithmic-based analyses. The professional human transcribers have a distinct advantage over the algorithms resulting from the innate ability of the human auditory system to identify context associated with each sound. For example, a sound without context may sound like a squeal (speech), but when the context of the situation is noted, the same sound may clearly be a scream (non-speech, fixed signal). Nonetheless, the high degree of classification agreement between the transcription and the LENA System is noteworthy. LENA Foundation engineers continue to work to improve the accuracy of the child speech segmentation algorithms.

2 Performance numbers presented in Table 4 have been updated in this revision.

3.0 RELIABILITY OVER TIME

The data described in Section 2 reveal the degree to which the LENA software accurately and reliably assesses the language environment of the child. In this section, we discuss the reliability of the LENA processing over time.

The data in Figure 4 reveal that the accuracy of the LENA-based AWC estimate is a function of recording time. The solid line indicates the percentage of difference between the human transcriber word counts and the LENA Adult Word Count estimates. Initially, human and LENA-based AWC estimates differ by more than 40%, however this variation decreases virtually logarithmically as a function of time. The variability begins to plateau after approximately one continuous hour of recording, and ultimately the error steadies at a rate of variability of roughly 5%.

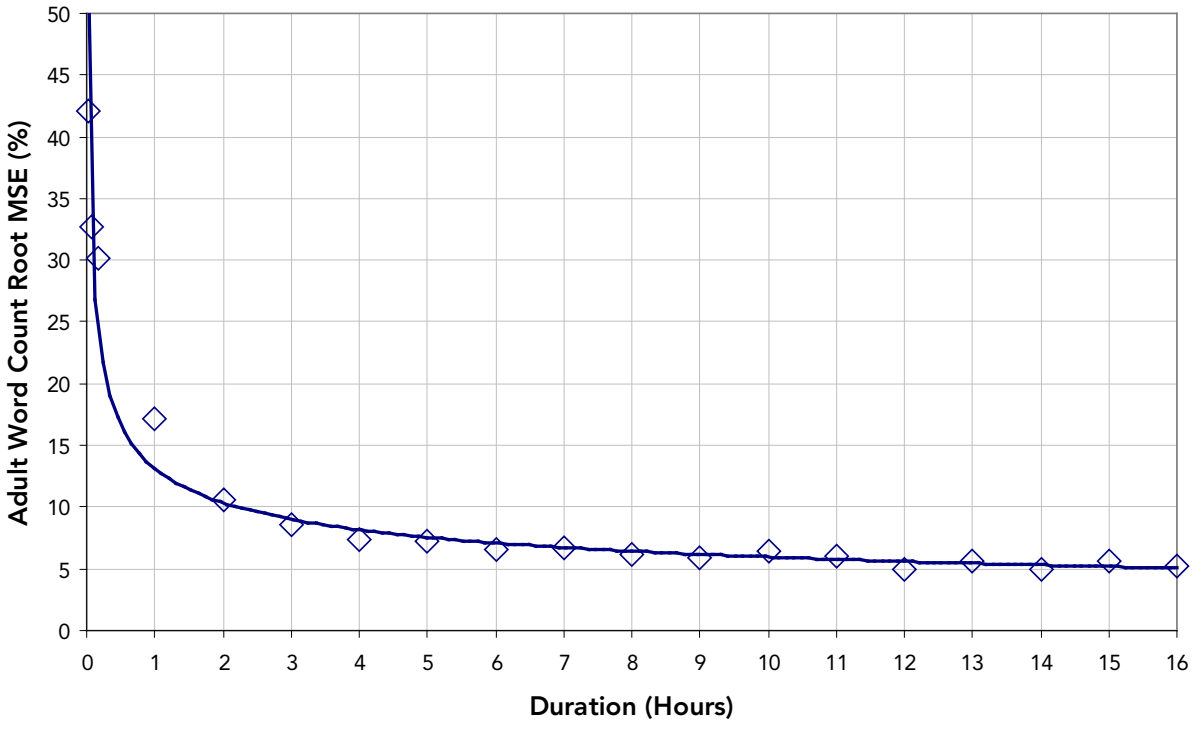

Figure 4. LENA-based AWC error as a function of time.

This time-dependent variability curve is primarily a result of the regression modeling approach used for the analyses. Specifically, over time, false positives and negatives cancel out, thus minimizing the effect of individual misclassifications. This being the case, it is of utmost importance that users of the LENA record for a minimum of one hour (preferably longer) for relatively accurate estimation of AWC. It should also be noted that LENA was designed and normed for use in a 12-hour long spontaneous speech environment, and it is not intended for use in a controlled environment or for use with short excerpts of read speech.

4.0 SOURCES OF VARIABILITY

The LENA System was designed to function as an audio processing unit rather than as a speech recognition device (i.e. the LENA System does not identify individual words, but provides word count estimates based on acoustic information in the audio stream). The algorithms used to process data from the LENA DLP are probabilistic modeling approaches. As a result, any variability in the source signal will affect the system performance, with the affect being dependent on the amount of degradation caused by the interference. Due to the nature of the data collection, we faced a variety of challenges that were unique to the natural environment where data were collected. These challenges are briefly described in Table 5. Here, we classify the sources of variability and elaborate on the effect of these variations on the final system performance. The most significant sources of variability in the LENA System include interferences from environmental factors, speaker differences, effects of clothing, and inter-recorder differences.

Table 5: Sources of variability for naturalistic, real-time data collection.

Variability Source	Natural Environment	Traditional Environment
Environmental	Background noise	No background noise
	External conversations	No external conversations
	Overlapping speech	Speech controlled
	Channel Acoustics	Relevant
Speaker Variations	Speaking style	Controlled
	Rate	Controlled
	Accent or dialect	Controlled
	Pitch	Controlled
	Sickness	Controlled
Clothing Effects	Thickness	Not Applicable
	Sound absorption rate	Not Applicable
Hardware Effects	Inter-processor variability	Relevant
	Hardware and operating system variability	Relevant

Environmental effects are a major source of variability. Channel acoustics were a primary source of error in this category. Echo and reverberation effects resulting from room size, flooring type, environmental location, and far-field effects could negatively impact signal integrity.

The AWC estimates are also influenced by variations of speech quality introduced by different speakers in the audio files. These variations include speaking style, rate, accent or dialect, pitch (including pitch variations resulting from parentese) and voice changes resulting from a variety of health conditions.

The child wears the LENA DLP for a continuous span of 12-16 hours. As a result, the signal quality is affected by the child's clothing. This source of variability was greatly minimized through the production of custom-made clothing tailored specifically to optimize the quality of audio files from the DLP. It should be noted that the LENA clothing has been rigorously tested to ensure that variability associated with the clothing is minimized; the LENA clothing should be worn during audio recording sessions for optimal recording quality.

Finally, hardware-related sources also contributed to the overall variability of the estimates. Error may have been introduced by the inherent variability between different DLP instruments as well as variability within the computer hardware and operating systems.

5.0 CONCLUSION

We have described the accuracy and reliability of LENA language environmental analysis software V3.1.0 in terms of segmentation, AWC estimates, and child vocalization classification. Confusion matrices showed that LENA and human-based segmentations had a high level of agreement. Misclassifications were primarily false negatives resulting from the elimination of overlapping speech, and thus the LENA estimates were likely more representative of the meaningful language environment of the child. Similarly, LENA and human-based estimates of AWC were highly correlated when random concatenated sections were transcribed; this reliability increased as a function of time. Full-day transcriptional analyses of two select files (one representing a typical quiet day, the other representing a typical active day) revealed that LENA System and human-based AWC estimates deviated during segments containing substantial noise and overlapping speech. The lower estimates by the algorithmic models may again be more reflective of what the child is absorbing. In addition, the reliability of LENA analyses may be higher than reported due to inherent measurement error present within the human-based transcriptions. LENA- and human-based child vocalizations classifications were predominantly in agreement, with a slightly greater tendency for the algorithmic models to misclassify child non-speech sounds as speech. Finally, the unique data collection environment introduced sources of variability not generally seen in traditional data collection environments. These sources included primarily environmental, speaker variation, and technical factors.

REFERENCES

Oller, D.K. (2000) The Emergence of the Speech Capacity. Mahwah, New Hersey: Lawrence Erlbaum Associates.

Poag, C.K., Goodnight, J.A., & Cohen, R. (1985). The Environments of Children: From Home to School. In R. Cohen (Ed.), The Development of Spatial Cognition (pp. 71-113). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Wachs, T.D. (1982) Relation of home-noise confusion to infant cognitive development. Annual Meeting of the American Psychological Association.

