The LENA™ Vocal Productivity Measure

Shichuan Du, Dongxin Xu, Jeffrey A. Richards,

Stephen M. Hannon, Jill Gilkerson

LENA Foundation, Boulder, CO

LTR-11-1

January 2017

Software Version: V4.6.1

ABSTRACT

This report describes the development of the LENA Research Foundation's vocal productivity measure (VP). The term vocal productivity is used here to refer to the length of children's vocal output in canonical syllables (CS), or well-formed consonant/vowel pairs (Oller, 2000). Canonical syllables have been used as a reliable metric for gauging vocal development in young children for several decades, showing particular clinical utility in identifying children at risk for language delays or deafness (Eilers & Oller, 1994). More recent work by Oller et al. (2010) used LENA technology to demonstrate the feasibility of a new methodology for automatically identifying canonical syllables, reporting good reliability in distinguishing typically developing children from children with autism based on their vocal output. The current research investigates an adaptation of this automated canonical syllable measure, focusing on turn-contingent CS count per conversational turn to assess child vocal development from children 6 to 48 months of age. In this report we describe the rationale of this adaptation, methods to achieve a robust estimation, correlations with age and validity compared to professional transcription.

Keywords

Vocal productivity, canonical syllable, vocal development, MLU, expressive language

1.0 INTRODUCTION

This paper presents a general overview of the development, reliability and validity of the LENA Research Foundation's vocal productivity measure (VP).

1.1 Background and Motivation for VP Development

Mean length of utterance (MLU) is widely used to assess early stages of spoken language development by estimating the number of morphemes or words in a typical sentence or utterance (Brown, 1973). MLU calculation is a labor intensive process requiring professional transcription and judgment about utterance boundaries and morpheme identification, and is thus typically based on very short speech samples (e.g., 30 min). Here we describe an attempt to both avoid laborious transcription and utilize considerably larger speech samples to estimate the average length of child vocalizations, using a purely automated method void of word content information.

Canonical syllables emerge at a stage of development when children begin to produce clearly articulated consonant/vowel (CV) pairs. These well-formed syllables contain a CV transition duration shorter than 120 ms (Oller, 2000), similar to that of adult speech, such as a sharply pronounced "ba" instead of "bbbbaaaaa" with a slow transition from "b" to "a" (e.g., >120ms). The production of well-formed syllables (CV pairs) is a distinct phase of vocal development emerging between 7-9 months of age. Children who do not consistently produce canonical syllables by 10 months are at risk for language delay and other neurological disorders (Oller *et al.*, 1998; Oller *et al.*, 1999; Patten *et al.*, 2014). In toddlerhood and early childhood, as the linguistic complexity of spoken language increases, children add multi-syllabic words to their spoken vocabulary and produce increasingly longer multi-phrasal sentences (Fromkin, Rodman and Hyams, 2013), leading to an increase in the number of syllables per utterance over time. Thus, it is reasonable to presume that a syllables-per-utterance measure could be an indicator of early spoken language development.

The current research was motivated by recent work investigating the reliability of automated vocal output measures for identifying children at-risk for autism spectrum disorders (Oller et al., 2010). That study used 12 acoustic parameters associated with early vocal development and included automatic identification of canonical syllables. The original CS measure was shown to reliably differentiate children with and without autism, and has been adapted herein as a potentially useful measure for automatically estimating the length of vocal output in communicative interactions (conversational turns) in young children.

This paper describes the development of an automated proxy for MLU designed to estimate length of vocal output in canonical syllables rather than words or morphemes, since

morpheme development results in increasing utterance length. Section 2 describes the basic algorithm for CS per turn; reliability for the automated analysis of CS is presented in Section 3; Section 4 shows comparisons with criterion measures; and the last section discusses the experimental nature of the measure and potential clinical implications.

2.0 ALGORITHM

The CS algorithm is built on a series of segmentation steps that identify child speech-related vocal segments and then capture CS within each segment based on a set of well-defined rules. This approach is summarized and illustrated in Figure 1, adapted from Oller *et al.* (2010). Step 1 identifies child vocalization segments in a conversation with turns, e.g., mother-child-mother, then Step 2 identifies periods of high energy within the child segments, distinguishing speech-related vocalizations from vegetative sounds and cries. Lastly, canonical syllables are estimated from the speech-related vocal sounds based on formant change rules.

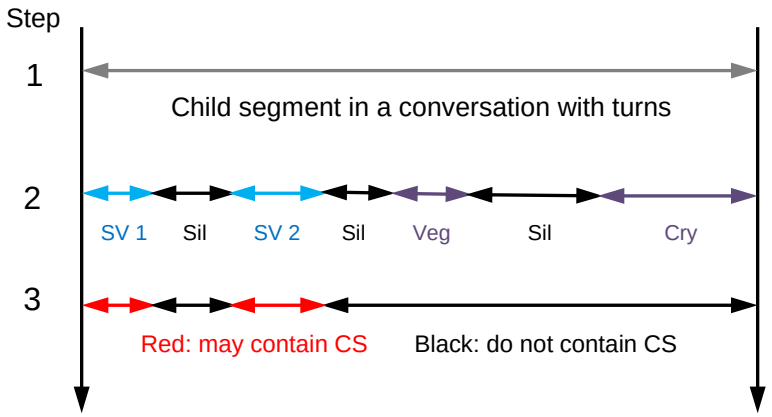

Figure 1. An example of the steps to obtain turn-contingent CS from a recording, adapted from Oller *et al.* (2010). SV: speech-related vocal sounds; Sil: silence; Veg: vegetative sounds.

2.1 Robust Estimation of a Turn-Contingent Canonical Syllable Measure

Examination of CS within turns is of interest to distinguish children's interactive talk from their monologues. In the course of developing such a turn-contingent measure, we have found that normalizing the number of CS produced within turns by the count of turns (CS per turn) rather than the count of utterances produces a more linear age distribution.

Because the length and frequency of conversations vary throughout a day, it was necessary to choose an appropriate time resolution for the computation of CS per turn so that the variance of the measurement could be sufficiently minimized. In other words, one needs to observe a child for a certain amount of time to be able to summarize a behavioral

characteristic or statistic reliably without risking a substantial bias. We also sought an estimator that is robust against errors and outliers. Since in general median values are more robust than mean values when we do not exclude "outliers" (here defined as data points that are two or more standard deviations from the mean), we chose a block-wise paradigm in which conversational turns are grouped into blocks. The CS count is divided by the turn count within each block, and then the median of these block average values defines the CS per turn value for the recording.

To elaborate, the total number of turns in a daylong recording (M) are grouped sequentially into blocks of N=16 turns to generate $\left[\frac{M}{N}\right]$ samples of CS per turn (hereafter, *CS.pt*). Because conversation boundaries are well-defined by LENA system algorithms, conversations with turns are used as basic building units to construct the blocks. For each conversation, we can compute $\mathrm{CS.pt.conv} = \frac{\text{CS count}}{\text{turn count}}$. If the number of turns in a conversation is greater than N, it is split into sub-conversations and the CS count for each sub-conversation generate as [CS.pt.conv × the number of turns in the sub-conversation]. After CS.pt has been computed for each block, vocal productivity VP for the recording is computed as the median CS.pt block value. Figure 2 illustrates how conversations are divided into N-turn blocks.

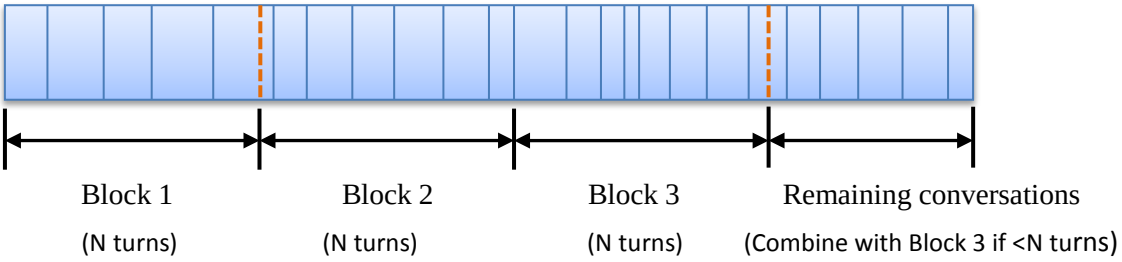

Figure 2. Illustration of breaking conversations into blocks of N turns. The vertical solid lines mark the boundaries between two conversations. The dashed lines indicate where the conversation is separated by two blocks. If the conversations at the end of the recording have less than N turns, they will be merged into Block 3.

2.2 Confidence Interval, Z scores and Percentiles

The $\left[\frac{M}{N}\right]$ CS.pt values for a recording are used to generate a confidence interval (CI) for VP, which requires the probability density function of the median. If the CS.pt values are sorted in ascending order, finding the median is analogous to counting the number of times heads occurs in *n* coin tosses, and thus the median follows a binomial distribution (Conover, 1980). The CI can be obtained using this distribution.

However, $\left[\frac{M}{N}\right]$ must be large enough to calculate a valid CI. When $\left[\frac{M}{N}\right]\geq 6$, the 2.5% of cases in the 95% CI lower bound is greater than or equal to the smallest CS.pt estimated within the recording. When $\left[\frac{M}{N}\right]<6$, the 2.5% bound will be less than any CS.pt estimated within the recording and the 97.5% bound will be greater than any measured CS.pt. Although we know the lower bound cannot be less than 0, we do not know where the higher bound is. Therefore, 6×N was chosen to be the required minimum number of turns a recording must have to generate a median and confidence interval estimate. Recordings that do not meet this requirement are marked "insufficient data" in the ITS (Interpreted Time Segments) file.

We also generate age-adjusted Z scores and percentiles for the VP estimate using an age-normalization method that essentially mimics that utilized in our existing measures for child vocalizations and conversational turns (See Gilkerson & Richards, 2008).

3.0 RELIABILITY

The original CS identification algorithm has been validated by comparison with the judgments of professional transcribers and reported in Oller *et. al* (2010). These reliability results are summarized below.

3.1 Comparison with Human Listeners

Using phonetically trained human listeners' judgments as the gold standard, the CS measure in Oller *et. al* (2010) demonstrated a 65% accuracy rate across 16 five-minute samples (8 typically developing, 4 autistic, 4 language-delayed), as shown in Table 1. Session-level agreement is the correlation between the number of positive judgments made for each of the 16 samples by the human listener and the number of positive judgments made by the algorithm for the same 16 samples (Oller *et. al*, 2010). This Pearson correlation is 0.52 and the Cohen's kappa is 0.21.

Table 1: Reliability results (Oller *et al.*, 2010).

Measure	Session level correlation between human listener and machine classifications	Proportion correct machine classifications with human listener as gold standard	Cohen's kappa for human listener vs. machine classification	Chi-square probability of Cohen's kappa
CS	**0.52**	**0.65**	**0.21**	**<0.0001**

4.0 COMPARISON WITH OTHER MEASURES

Having established a significantly positive correlation with human listeners' judgment, we next compared VP with our existing automated vocalization assessment and other criterion measures.

4.1 Comparison with Age and AVA

Since children's vocalizations get longer as they acquire more vocabulary and use longer sentences (Fromkin, Rodman and Hyams, 2013) we expect the raw count of VP to increase with age. To examine the relationship to age, we obtained VP on our normative reference dataset of 3384 recordings from 294 children, most of whom contributed 4 or more longitudinal recordings. Fitting a line to the raw counts in Figure 3, we found that the slope of the line was significantly positive ($p < .001$) with a Pearson correlation coefficient of $r = 0.64$. Age-standardized VP Z scores were uncorrelated with age, $r = -0.002$.

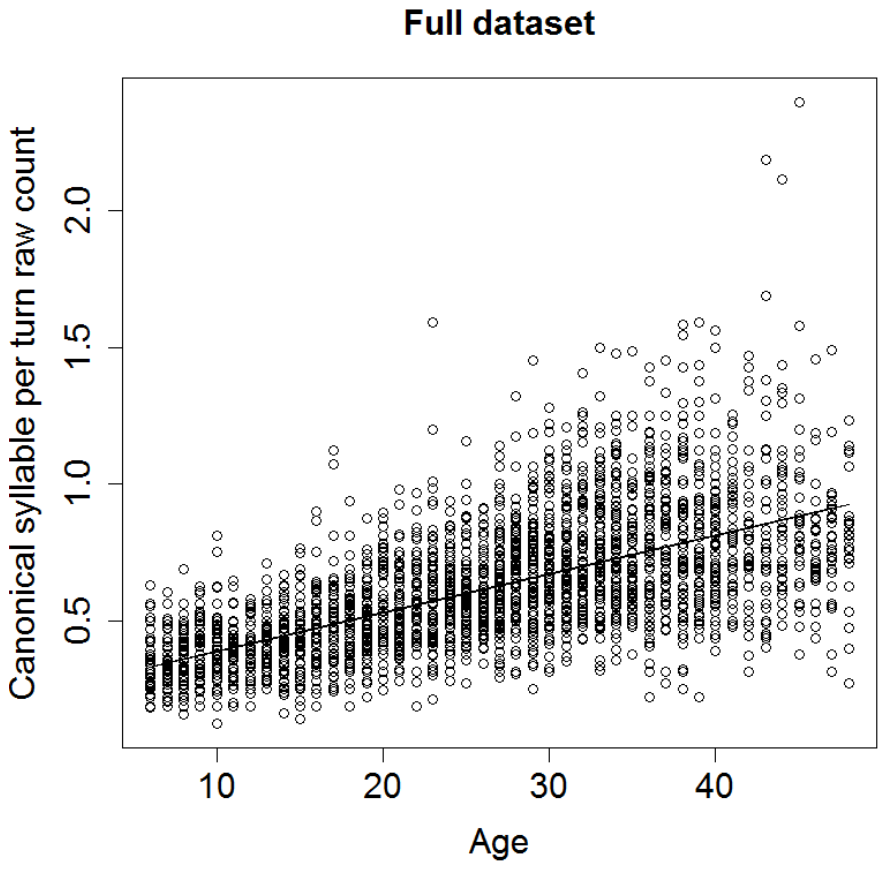

Figure 3. CS.pt raw count from age 6 to 48 months of 292 children. The raw counts increase with age: the slope of the fitted line is 0.014 (p < .001). The Pearson correlation is r = 0.64.

Since VP increases with age, it is reasonable to predict that it should correlate with other measures of vocal complexity. The LENA Automatic Vocalization Assessment (AVA) is an

automatically derived quantitative measure of the acoustic complexity in the child's vocal output on a recording day (Richards *et al*., 2008). It uses the distribution of phones (roughly vowels and consonants) to estimate vocal complexity compared to age-matched peers. The Pearson correlation between VP and AVA score Z scores is r = 0.37 (Figure 4). Lower correlations between VP and AVA results suggest that although VP and AVA are both indicators of growth over time, they likely measure different aspects of development.

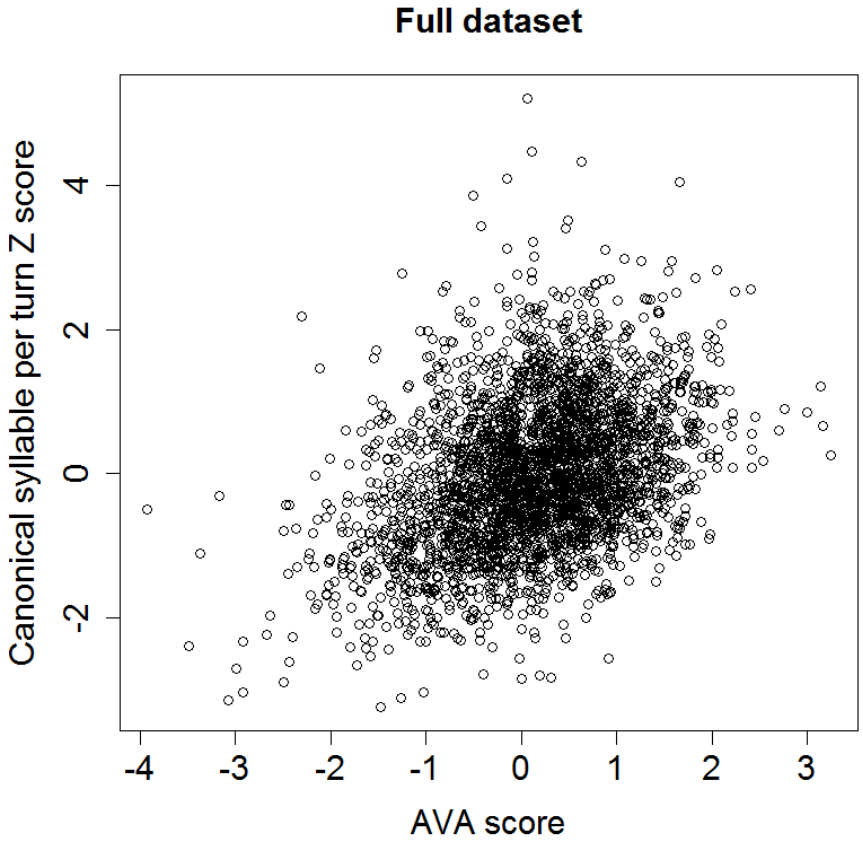

Figure 4. CS.pt Z score vs. AVA Z score of 292 children. The correlation is r = 0.37.

4.2 Comparison with MLU

As previously described, mean length of utterance (MLU) has been widely used to evaluate stages of language development in typically developing children as well as children with language delays or disabilities (Eisenberg, Fersko, and Lundgren, 2001; Rice *et al.*, 2010). To assess the relationship between VP and MLU, professional transcribers using morpheme counting rules from Miller and Chapman (1981) generated MLU using the first 100 utterances from recordings of 67 children evenly distributed across ages 14-48 months.

The Pearson correlation between MLU and VP raw count for these files was r = 0.57 as shown in Figure 5. After removing two outliers (red solid circles), the correlation increases to r = 0.64. Transcription generated MLU has a higher correlation with age (0.77) than the

automatically derived VP raw count (0.64).

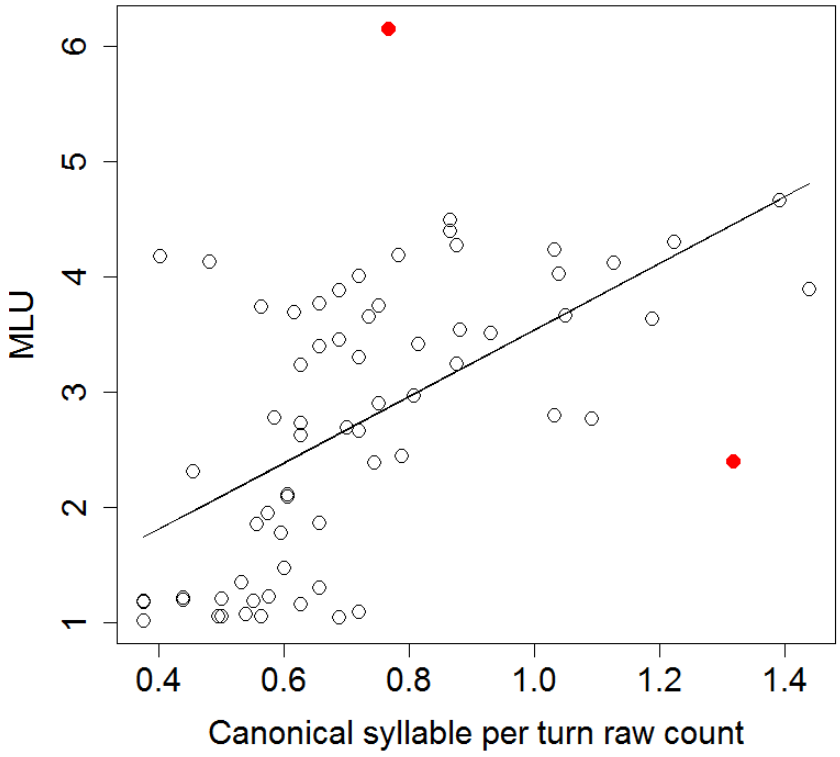

Figure 5. CS.pt raw count vs. MLU for each recording. The slope of the line is 2.89 (p=3.64e-07) and the correlation between the two measures is 0.57. Without the two red solid circles, the correlation becomes 0.64.

5.0 CONCLUSIONS

This paper has described the development of an automated vocal productivity measure VP that generates an estimate of the length of a child's vocal output in terms of canonical syllable production. The vocal productivity measure was shown to have a significant correlation with age and mean length of utterance MLU, whereas it was partially correlated with AVA scores. Future research will focus on comparisons with larger datasets using a variety of demographic subsets and languages. As well, means by which the base estimator of CS can be improved should be explored, for example by investigating the usefulness of incorporating other automated parameters (e.g., consonant voicing, syllable duration) into the algorithm

A primary advantage of the vocal productivity measure is that it obviates the need for labor intensive transcription and coding and provides an analysis of child vocal output throughout the day rather than relying on short speech samples. As such, it has the potential for

demonstrating clinical utility as a supplement to an overall professional language evaluation. Additionally, since it does not consider semantic content it could prove to be a useful language-universal measure of vocalization length, rather than depending on morpheme identification which can vary according to language and morpheme counting rules (Miller and Chapman, 1981). More research is needed to investigate its reliability for identifying children with language delays, but the strong correlations between VP and age suggest that it could provide useful information as a component to a level 1 screen. Given the novelty of this approach, and the important work needed to extend its validation, the vocal productivity estimate should be considered an experimental measure. The LENA Research Foundation welcomes feedback on its research and clinical utility.

REFERENCES

Eilers, R.E. and Oller, D. K. (1994). Infant vocalizations and the early diagnosis of severe hearing impairment. Journal of Pediatrics. 1994, 124, 2, 199-203.

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., Yapanel, U. and Warren, S. F. (2010). Automated Vocal Analysis of Naturalistic Recordings from Children with Autism, Language Delay, and Typical Development. The Proceedings of National Sciences Academy, 2010, 107, 30, 13354-13359.

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

Brown, R. (1973). A First Language: The Early Stages. London: George Allen & Unwin.

Patten, E., Belardi, K., Baranek G. T., Watson L. R., Labban J. D., and Oller, D. K. (2014). Vocal Patterns in Infants with Autism Spectrum Disorder: Canonical Babbling Status and Vocalization Frequency. Journal of Autism and Developmental Disorders. 2014, 44, 10, 2413–2428.

Fromkin V., Rodman R. and Hyams N. (2013). An Introduction to Language. California: Wadsworth Publishing.

Oller, D. K., Eilers, R. E., Neal, A. R., and Cobo-Lewis, A. B. (1998). Late Onset Canonical Babbling: A Possible Early Marker of Abnormal Development. American Journal of Mental Retardation. 1998, 103, 3, 249-63.

Oller, D. K., Eilers, R. E., Neal, A. R., and Schwartz, H. K. (1999). Precursors to Speech in Infancy: The Prediction of Speech and Language Disorders. Journal of Communication Disorders. 1999, 32, 4, 223-45.

Conover, W. J. (1980). Practical Nonparametric Statistics. New York: John Wiley and Sons.

Gilkerson, J. and Richards, J. A. (2008). The LENA natural language study (LTR-02-2). Boulder, Colorado: LENA Foundation.

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N. and Blossom, M. (2010). Mean Length of Utterance Levels in 6-month Intervals for Children 3 to 9 Years with and without Language Impairments. Journal of Speech, Language and Hearing Research. 2010, 53, 2, 333–349.

Eisenberg, S. L., Fersko, T. M., and Lundgren, C. (2001). The Use of MLU for Identifying Language Impairment in Preschool Children: A Review. American Journal of Speech-Language Pathology. 2001, 10, 323–342.

Miller, J. and Chapman, R. (1981). The Relation between Age and Mean Length of Utterance in Morphemes. Journal of speech and hearing research, 1981, 24, 2, 154-61.

Richards, J. A., Gilkerson, J., Paul, T. and Xu, D. (2008). The LENA Automatic Vocalization Assessment.

