Automatic estimation of infant syllable production in naturalistic recordings

Anne S. Warlaumont, University of California, Merced
Heather L. Ramsdell-Hudock, Idaho State University

UCLA BEC Seminar, May 9, 2016

Definition of canonical babbling

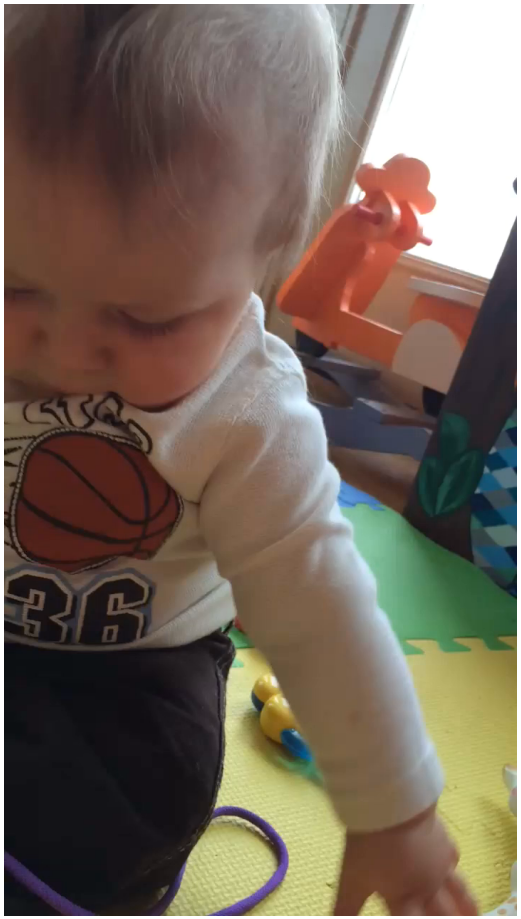
- Babbling that contains at least one canonical syllable
  - at least one consonant
    - Not “h”
  - at least one vowel
    - sometimes required to be a “full” vowel
  - swift, adult-like transitions between the two

Oller (1980, 2000); Stark (1980); Koopmans-van Beinum & van der Stelt (1986); Roug, Landberg, & Lundberg (1989); Buder, Warlaumont, & Oller (2013)

Development of canonical babbling (consonant-containing syllabic sounds)

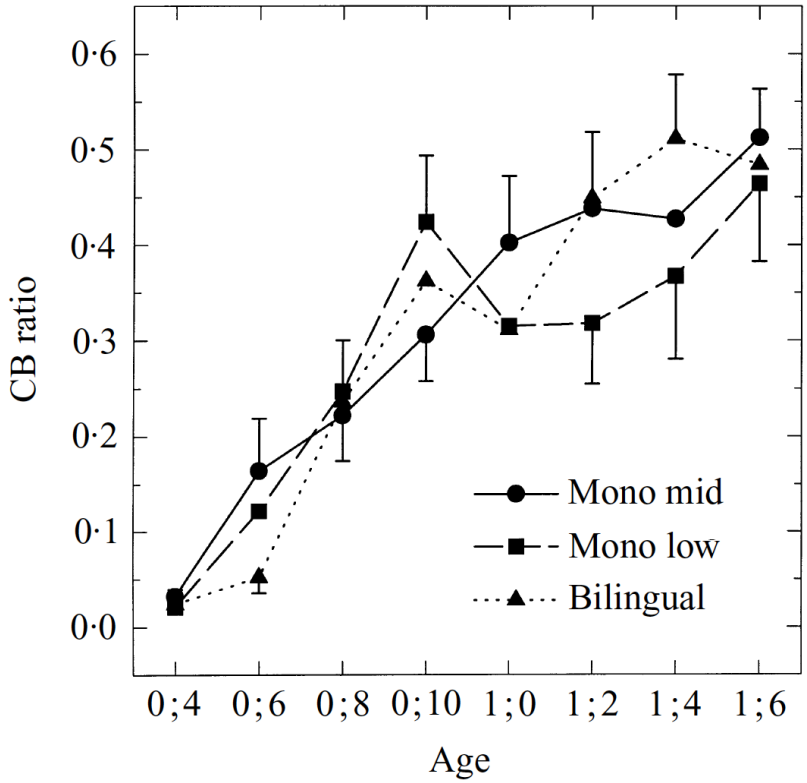
2 weeks 6 days old

~4.5 months old



~10 months old

Consonant frequency increases with age



Oller et al. (1997)

Canonical babbling is a foundational precursor to meaningful speech

- The sounds that infants produce during their babbling tend to be the ones that appear in first words.
- Babbling appears to essentially always precede word production.
  - Tracheostomized infants provide the most compelling evidence for this.

Why do we need automated methods for detecting canonical babbling?

- Efficiency
- Standardization across labs
- Basic science
  - Including computational modeling projects
- Application to intervention and education

Existing methods

- Landmarks-based syllable detection
  - Intended for babble!
  - Could use more validation on an independent dataset
- Sphinx consonant and vowel detection
  - Xu et al. (older children, canonical sounds only)
- de Jong & Wempe syllable detection
  - Built for adult speech
- Coath et al. salience peaks detection
  - Built for adults speech, previously tested on synthesized speech
- Oller et al. (2013)
  - Code not publicly available

The present study

- How do existing & readily available methods do at automatically counting syllables in infant vocalizations?

Audio recording

- 521 recordings of 16 English-learning infants between 3 and 20 months old
- Laboratory designed to mimic a nursery

Human coding

- 57,629 utterances identified by human listeners
- 2 human listeners were trained to identify canonical syllables
  - IVICT (Infant Vocalization Interactive Coding Trainer), available at babyvoc.org
  - Read Buder et al. (2013)
  - Discussed the definition with me
  - Each prompt contained the basic definition ("adult-like syllables containing at least one consonant other than 'h' and at least one vowel")

Human coding

- 57,629 utterances identified by human listeners
- 2 human listeners were trained to identify canonical syllables
  - IVICT (Infant Vocalization Interactive Coding Trainer), available at babyvoc.org
  - Read Buder et al. (2013)
  - Discussed the definition with me
  - Each prompt contained the basic definition ("adult-like syllables containing at least one consonant other than 'h' and at least one vowel")

Human reliability

- rho = .74, $p < .001$ for canonical
- rho = .7, $p < .001$ for total

Acoustic analyses

- Landmarks (MATLAB)
- de Jong & Wempe (Praat/shell)
- Sphinx (shell/Python)
  - consonant count (excluding ‘h’)
  - vowel count
- Coath et al. salience (MATLAB)
- 5 counts per infant utterance

https://github.com/AnneSWarlaumont/CountInfantSyllables

Performance of existing methods

Method	Canonical syllables	Total syllables
SpeechMark	.22	.50
de Jong & Wempe	.24	.65
Coath & Denham	.15	.41
Sphinx consonants	.30	.52
Sphinx vowels	.27	.57

rho (Spearman's rank correlation coefficients) w/ human judgments. All p's < .001

Statistical prediction

- Leave-one-child-out method of dividing data (Oller et al., 2010)
- Equalization of number of items per category in training datasets
- Conversion to z-scores
- PCA on the training datasets
- Generalized additive models w/ 5 event counts + duration as inputs, either # of canonical syllables or total syllables as outputs
- Machine estimate based on the model's predicted syllable counts for test datasets (data from a child the model had never seen before)

Examples

1 syllable, 0 canonical

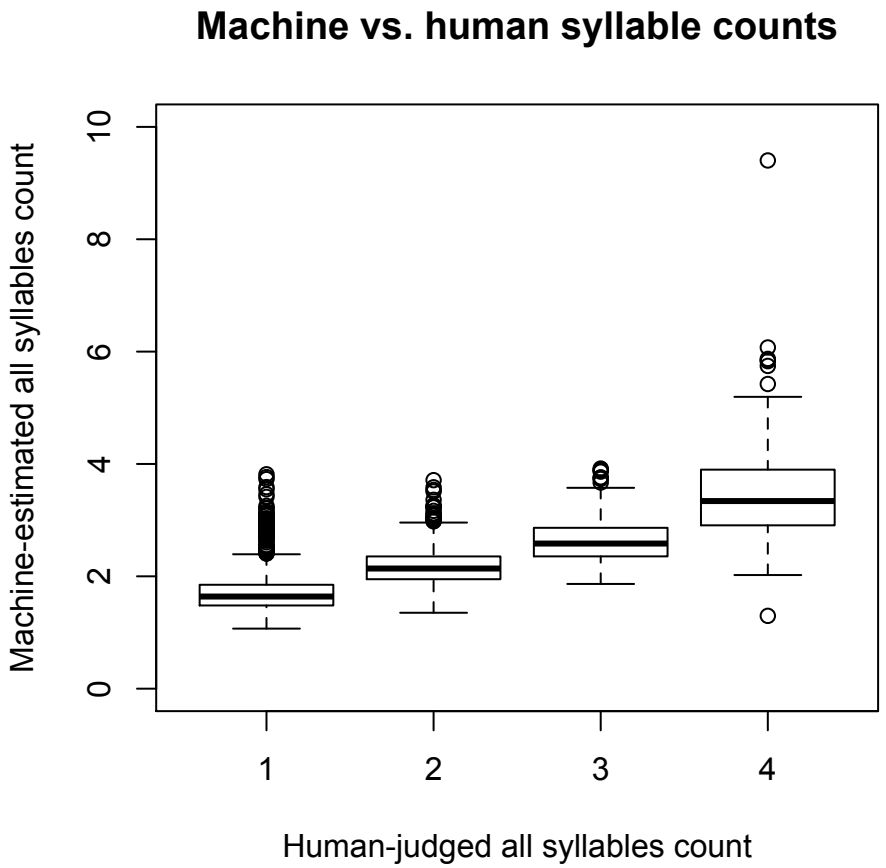
1 syllable, 0 canonical

1 syllable, 1 canonical

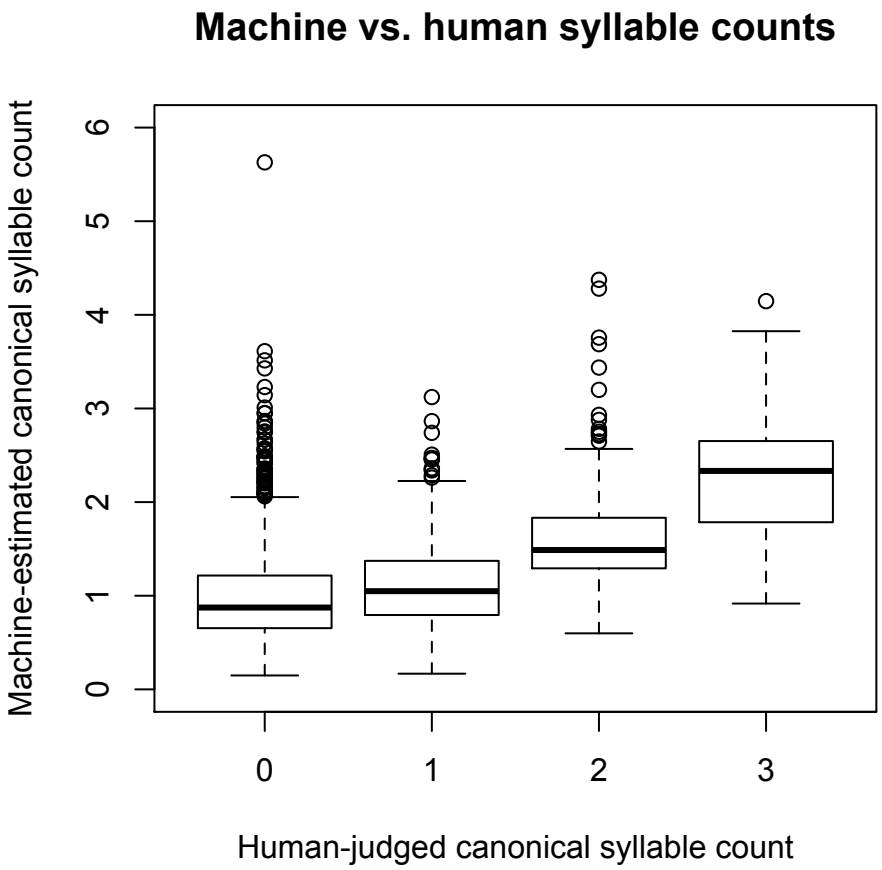
3 syllables, 3 canonical

4 syllables, 2 canonical

Human-machine reliability



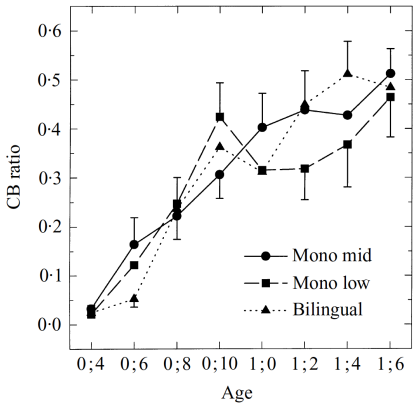
rho = .70, p < .001

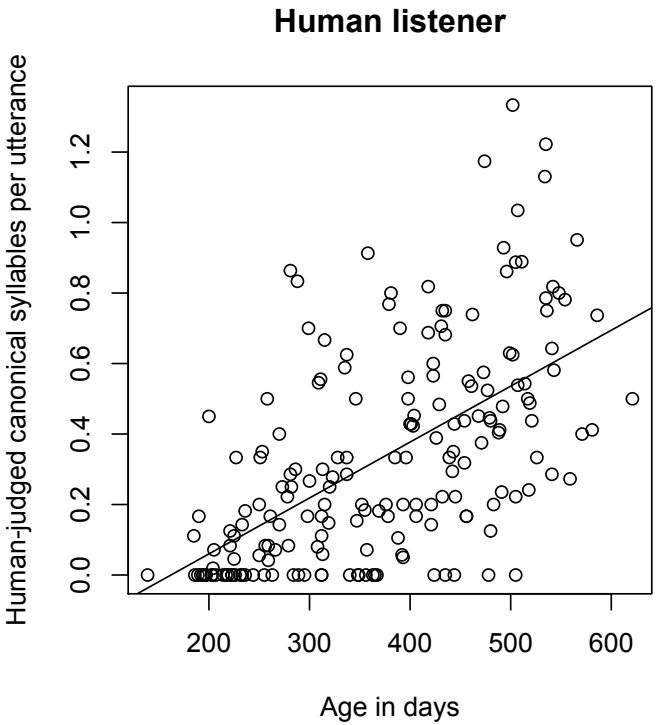


rho = .29, p < .001

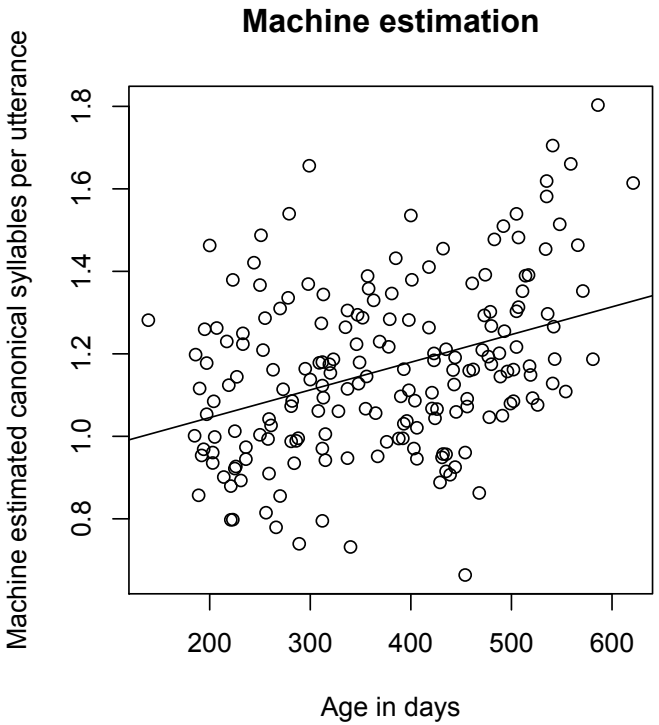
Good enough to replicate general developmental trends

Oller et al. (1997)



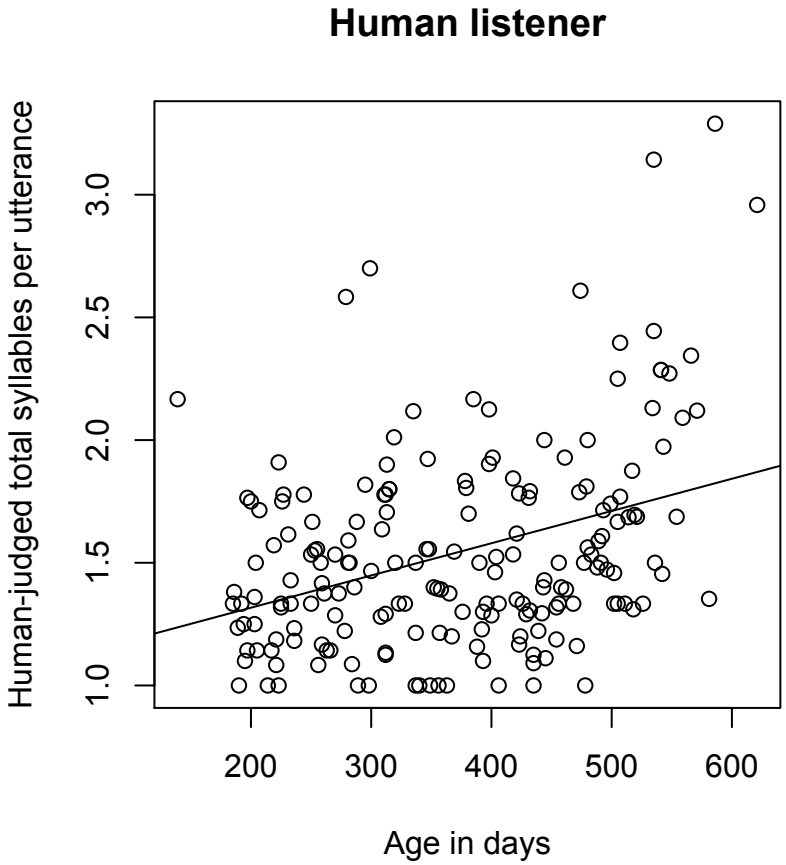


r = .59, p < .001
ratio: r = .54, p < .001

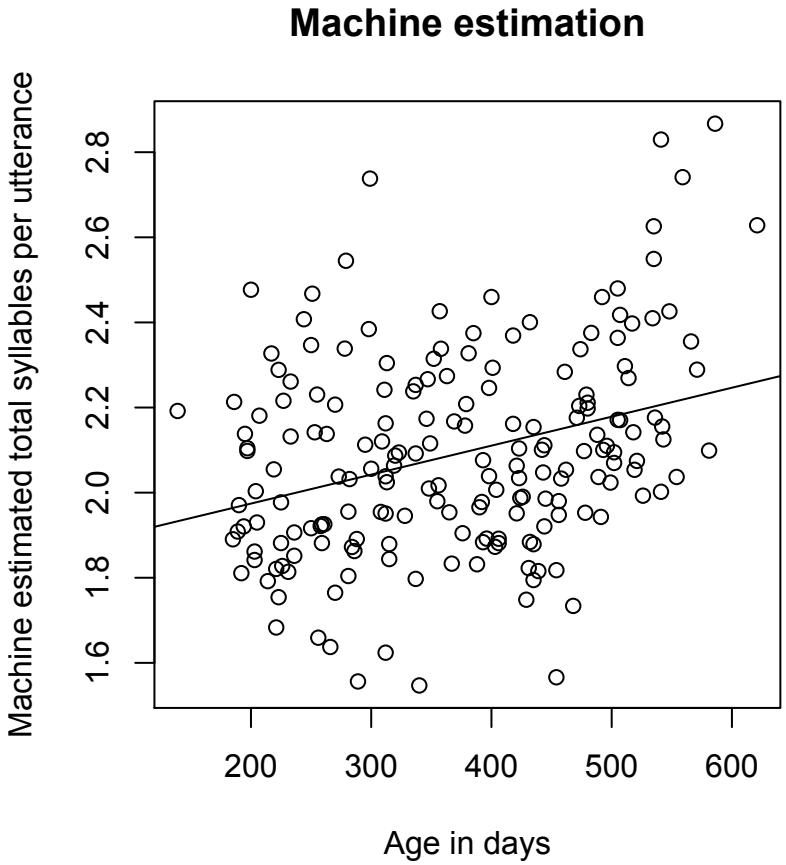


r = .36, p < .001
ratio: r = .38, p < .001

Total syllables also increase with age



r = .37, p < .001



r = .32, p < .001

Point of possible interest: Because accuracy of machine estimates is better for total syllables, the match to human performance is higher, and the age trend is better replicated. Can use total *or* canonical syllables if you just want to capture age trends.

Conclusions

- Promising results counting canonical and non-canonical syllables in infant data
- Performance counting syllables of any type is similar to human performance!
  - And combined method is better than any of the other methods alone
- Separation of canonical from non-canonical syllables still needs some work
  - Combined method performs similarly to the Sphinx consonant count, so may as well just use Sphinx's consonant count
- Sufficient to identify one trend of interest: increase in syllables w/ increasing age
  - New finding: total syllables (not just canonical) increase with age
- Code is available at https://github.com/AnneSWarlaumont/CountInfantSyllables

Future directions

- Apply the methods here to LENA recordings
  - Noisier contexts
  - Incorrect speaker labels
  - More data, so possibly limited accuracy will be ok
  - How will they do on the adult vocalizations? Sphinx results are promising...
- Improve on the core acoustic methods to try to achieve better accuracy and precision
- Apply to study temporal dynamics of vocal learning
  - E.g., do canonical infant vocalizations get more adult responses?
  - Are infant canonical productions influenced by contingent adult responses (and to content + contingency of responses, a la Goldstein)?
  - Are canonical vocalizations produced in clumps/bursts?
  - Does variation across families predict outcomes?
- Making it easier to install things etc. using Virtual Speech Kitchens

Thanks!

https://github.com/AnneSWarlaumont/CountInfantSyllables