INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Detection of total syllables and canonical syllables in infant vocalizations

Anne S. Warlaumont[1], *Heather L. Ramsdell-Hudock*[2]

[1]Cognitive and Information Sciences, University of California, Merced, Merced, CA, USA
[2]Communication Sciences & Disorders, Idaho State University, Pocatello, ID, USA

awarlaumont2@ucmerced.edu, ramsdell@isu.edu

Abstract

During the first two years of life, human infants produce increasing numbers of speech-like (canonical) syllables. Both basic research on child speech development and clinical work assessing a child's pre-speech capabilities stand to benefit from efficient, accurate, and consistent methods for counting the syllables present in a given infant utterance. To date, there have been only a few attempts to perform syllable counting in infant vocalizations automatically, and thorough comparisons to human listener counts are lacking. We apply four existing, openly available systems for detecting syllabic, consonant, or vowel elements in vocalizations and apply them to a set of infant utterances individually and in combination. With the automated methods, we obtain canonical syllable counts that correlate well enough with trained human listener counts to replicate the pattern of increasing canonical syllable frequency as infants get older. However, agreement between the automated methods and human listener canonical syllable counts is considerably weaker than human listeners' agreement with each other. On the other hand, automatic identification of syllable-like units of any type (canonical and non-canonical both included) match human listeners' judgments quite well. Interestingly, these total syllable counts also increase with infant age.

Index Terms: automatic syllable detection, canonical babbling, infant vocalization

1. Introduction

A canonical syllable has full articulation of at least one consonant and at least one vowel, with swift, adult-like transitions between the two. Infants' production of canonical syllables is of interest to basic scientists because such syllables are foundational elements of human speech. At roughly 7 months of age, infants begin regularly producing canonical syllables as part of their vocal repertoires [1, 2, 3, 4, 5]. Studies tracking infants from 4 months to 18 months of age have indicated that the ratio of canonical syllables to non-canonical syllabic units increases steadily over this time period [6, 7]. Canonical syllable production has been shown to relate to later speech-language abilities [8, 9, 10]. Further, canonical syllable production rates have been shown to differ in certain at-risk groups, particularly children later diagnosed with autism spectrum disorder (ASD) [11] and children with severe hearing impairment [12, 13, 14, 15, 16].

Automated methods have the potential to greatly speed up the assessment of how many canonical syllables are present within infant vocalizations. Methods for automatic detection of canonical babbling skills could potentially become useful in early diagnosis of disorders such as ASD [17] as well as in interventions. Such tools would also be very useful for basic science researchers, who are increasingly relying on daylong audio recordings combined with automated analysis methods. In addition to providing a more efficient means of coding data than human listeners, a notable advantage of automated methods is their consistency: the same algorithm will always apply the same criteria in estimating the number of syllables in a vocalization; human raters may not. Automated methods hold promise for standardizing the characterization of infant vocalizations across studies, human raters, and labs.

Here we compare four existing tools on how well they match trained human listeners' judgments of the number of syllables and canonical syllables present in infant vocalizations, individually and in combination. We compare human-machine reliability to human-human reliability. Finally, we attempt to use the automated methods to replicate the pattern found in previous studies of more canonical syllables produced with increasing infant age. Note that our current goal is only to count the number of syllables, not to phonetically transcribe those syllables. Although phonetic transcriptions would be highly valuable, they would presumably be extremely challenging, as even trained human transcribers show show relatively low inter-rater reliability for transcription of infant utterances [18].

2. Method

2.1. Recording and infant utterance identification

Our data came from 531 recordings of 16 English-learning infants (9 female) who participated in a longitudinal study from 3 to 20 months of age. Recordings were made in a laboratory designed to mimic a home nursery setting. Caregivers and infants engaged in free play sessions, sometimes with lab staff involved, or the caregiver was engaged in an interview with the lab member while the infant was present.

Human listeners labeled the onsets and offsets of each infant vocalization. Vocalizations perceived as taking place during the same breath out were grouped together. Infant utterances included protophones (babbling, squealing, growling, quasivowels, yells, whispers, ingressive vocalizations, etc.), reflexive sounds (cries and laughs), and vegetative sounds (e.g., burps, sneezes, coughs, etc.). In total, 57,629 utterances were identified. Example utterances can be found in the multimedia files accompanying this paper.

2.2. Human identification of canonical syllables

Three human listeners (Listener 1 was the first author and Listeners 2 and 3 were undergraduate research assistants) judged randomly selected infant utterances. Listeners 2 and 3 were given training on the definition of a canonical syllable, through a combination of in-person training with the first author, reading a chapter that provides definitions for canonical syllables, marginal syllables, and other protophone categories [5], and

http://dx.doi.org/10.21437/Interspeech.2016-1518

undergoing example-based training through the Infant Vocalization Interactive Coder Training (IVICT) program [19].

Two classes of syllables were counted: (1) "infant-produced syllables (either canonical or noncanonical)" (minimum value of 1) and (2) "infant-produced canonical syllables (adult-like syllables containing at least one consonant other than "h" and at least one vowel)" (no minimum). "h" was excluded based on the reasoning that glottal stops and glottal fricatives do not require supraglottal articulatory control [6]. Thus, total syllable count gives the number of general amplitude nuclei whereas canonical syllables are the subset of the total syllables where there are both consonant and vowel components having features of well-formed adult speech. Canonical syllables are more advanced from a speech production standpoint.

Listeners also indicated when the utterance appeared to be a mis-labeling (i.e. it didn't actually contain an infant utterance), when it appeared to be a cry, laugh, or vegetative sound (e.g. a cough or a burp), and when there was overlap from another human or from a non-human sound source. After excluding sounds where at least one listener indicated any of these issues, there were 85 utterances labeled by Listener 1, 2,515 labeled by Listener 2, and 687 labeled by Listener 3. Listener 1 did not make enough judgments to afford inter-rater comparisons, so those are based on comparisons between Listener 2's and Listener 3's judgments only.

2.3. Syllable Detection methods

Our syllable detection and data analysis code is available at `https://github.com/AnneSWarlaumont/CountInfantSyllables` (v0.1.1).

We applied four different freely available algorithms for detecting syllables, salient events, or phones in speech or babble. Note that Oller et al. have reported good results using another algorithm [17], but as that method is not yet openly available, it is not represented among those tried here.

2.3.1. Speechmark syllable detection

The first existing tool was the syllable detection method from the freely available SpeechMark software. This is the only tool of the four that was specifically designed to process infant vocalizations (it also can process adult vocalizations depending on the user's settings). [20, 21]. SpeechMark takes a "landmark" [22] analysis approach. It looks for regions of the audio recording where there are abrupt changes in the signal, either in its amplitude or in its spectral properties, and classifies these changes into different acoustic-phonetic types. SpeechMark includes a syllabic analysis that finds regions where there are sets of landmarks indicating likely onset or offset of a syllable unit [23, 21]. We used the count of syllables provided by this feature of the SpeechMark software as a feature to try to predict the total number of syllables and the number of canonical syllables in our infant vocalization data.

SpeechMark also allows for more specific analyses based on the particular types of landmarks involved. These ought to be useful for differentiating canonical from non-canonical syllables (see [24] for a similar approach with promising results). This would require some additional processing steps so it is left as a future direction.

2.3.2. de Jong & Wempe syllable detection

The second tool we investigated was the syllable detection method of de Jong & Wempe, implemented in Praat [25]. The method uses amplitude information and pitch estimates to estimate the locations of syllabic nuclei within a waveform. Nuclei are assumed to be located during voiced portions of the sound where there are amplitude peaks. The algorithm was developed and tested on adult speech data. One of us has previously applied it to analyze synthesized vocalizations [26]. It was unknown how the method would perform at identifying syllabic nuclei in infant vocalizations.

2.3.3. Coath & Denham salient event detection

The third tool was developed by Coath, Denham, and colleagues to model the auditory salience of a stream of input to the auditory system [27, 28, 29]. This method attempts to model the processing performed by both peripheral and cortical nervous system regions. In essence, the system detects "edges" in the sound stimulus, either marked by changes in the activation of spectral components or by onsets or offsets of "cortical filters" (spectro-temporal patterns learned by machine learning over an adult American English speech corpus). The model has previously been shown to be able to track beats in sung music [28, 29]. We used the program described in [28], modified to use a lower threshold salience for event detection (thresh0 = .3 instead of 1) and a smaller divisor for the threshold adaptation (div1 = 1 instead of 2, so there was no adaptation over time).

Table 1: *Spearman's rank correlation coefficients (ρ) between the individual syllable detection methods and human syllable judgments. All correlations are statistically significant, $p <$ .001. 95% confidence intervals are in parentheses.*

Method	Canonical syllables	Total syllables
SpeechMark	.22 (.19,.25)	.50 (.47,.53)
de Jong & Wempe	.24 (.20,.27)	.65 (.62,.67)
Coath & Denham	.15 (.12,.19)	.41 (.37,.44)
Sphinx consonants	.30 (.26,.33)	.52 (.49,.55)
Sphinx vowels	.27 (.24,.30)	.57 (.55,.60)

2.3.4. Sphinx phone recognition

The fourth tool was the phone recognition tool from the open source Sphinx speech recognition software [30]. We used PocketSphinx, which has a mode that provides broad phonetic transcriptions of audio input without performing word recognition. The procedure is given at `http://cmusphinx.sourceforge.net/wiki/phonemerecognition`. We ran this in default configuration on each infant sound. We then added up the instances of all consonant phones, excluding HH, to match as closely as possible the instructions give to human listeners not to consider syllables to be canonical when "h" was the only consonant. We also added up all the instances of vowel phones. Thus, the Sphinx phone recognition system yielded two output counts, number of consonants and number of vowels, which we used to try to predict human listener syllable counts.

The Sphinx phone recognition method relies on acoustic models trained on adult American English speech. A previous study applied the tool to child speech (canonical speech-like utterances only) and compared performance to human transcriber consonant and vowel counts, with good correlations between the human transcriber counts and the Sphinx counts at the recording level [31]. The work also showed a relationship between the automatically obtained counts and child age, with differences across different clinical groups. However, that work

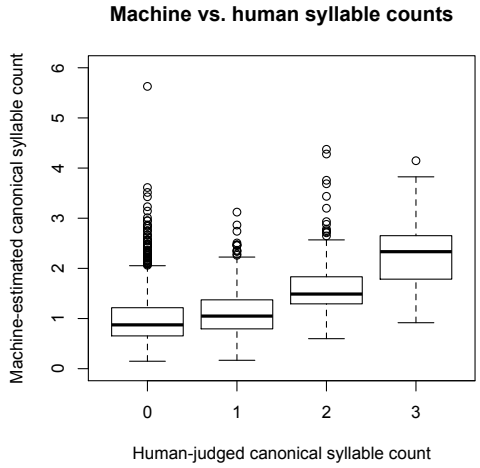

Figure 1: *Correspondence between machine-estimated canonical syllable counts and human listener counts.*

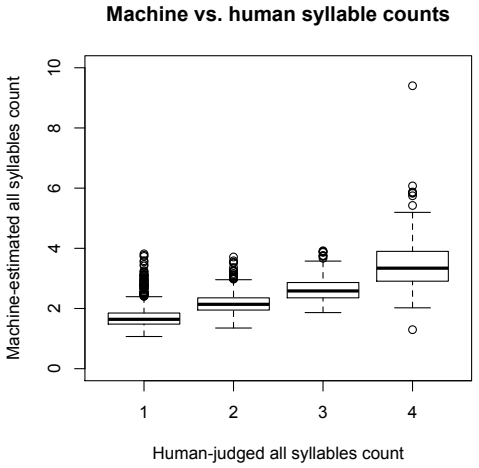

Figure 2: *Correspondence between machine-estimated total syllable counts and human listener counts.*

did not assess how well the Sphinx consonant and/or vowel counts would perform at identifying syllables or how it would do when applied to all child vocalizations, not only the clearly transcribable ones. It also focused on older children (approximately 13 to 47 months of age and older) than are the focus of the present study.

2.4. Model training and evaluation

We trained a generalized additive model with four inputs: number of syllables per utterance estimated by each of the four methods above plus utterance duration. The model thus had six input variables: duration in ms, number of salience onsets, number of syllables based on the de Jong & Wempe algorithm, number of syllables according to the SpeechMark software, number of consonants estimated by Sphinx, and number of vowels estimated by Sphinx. One model was trained to predict the number of canonical syllables in an utterance; a separate model was trained to predict the number of syllables of any type. Leave-one-child-out cross-validation was used to divide data into training and test sets.

Several pre-processing steps were performed. First, when more than one human listener judged a given utterance, the average number of syllables across the listeners was used. Mean syllable count averages were rounded to the nearest integer, and canonical syllable counts of three or more were grouped into a single category and total syllable counts of four or more were grouped into a single category. This created four ordinal levels for each syllable type count (0, 1, 2, and 3 for canonical syllables and 1, 2, 3, and 4 for total syllables of any type). Because there were unequal numbers of utterances falling into each category, the data from all count categories except for the most frequent were resampled, repeating as many utterances as were needed to create category sample sizes that were equal across all categories and matched to the most frequent category. This prevented the generalized additive model from becoming dominated by a general bias toward the most frequent count category. All five input variables were then converted to z-scores based on the training set data prior to building the model. Principal components analysis based on the scaled training set data was used to pre-process the model inputs; based on pilot explorations, the top three principal components were used. The resampling and scaling were done separately for each prediction type and for each training dataset in the leave-one-child-out cross-validation procedure.

3. Results

3.1. Human syllable count inter-rater reliability

Human Listener 2 and Human Listener 3's judgments were strongly correlated, $\rho = .74$, $p < .001$ for canonical syllable counts, and $\rho = .70$, $p < .001$ for total syllable counts, indicating reasonable but not perfect inter-rater reliability on these tasks. This provides a baseline against which the automated methods can be compared.

3.2. Individual performance of existing syllable detection methods

We tested separately how well each of the five syllable detection methods correlated with the average human listener total syllable and canonical syllable counts. Results are given in Table 1. All correlations were positive and statistically significant. Across all the syllable/phone detection methods, correlations with human listener total syllable counts were stronger than correlations with human listener canonical syllable counts. All correlations were weaker than human interrater correlations.

3.3. Reliability of syllable counts

It was possible that a weighted combination of the five syllable/phone detection methods would yield a better fit to human judgments than any of the individual methods on their own. We also thought that duration might play a useful role in predicting human syllable counts. We therefore turn to the results of the model that takes predicts human listener counts based on a combination of all the individual methods plus duration. Recall that these results are based on leave-one-child-out cross-validation test set performance. The combined machine-based canonical syllable estimates were significantly but weakly correlated with the average human listener judgments,$\rho = .29$,

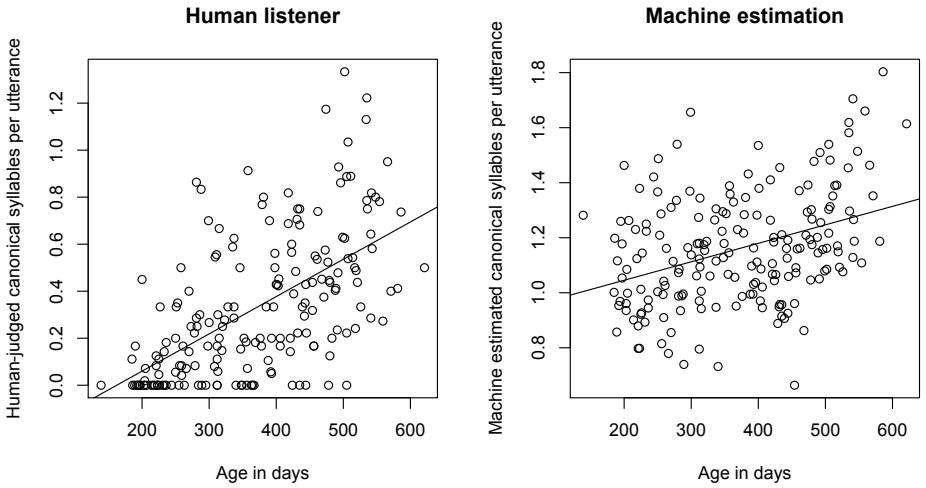

Figure 3: *The left plot shows the positive relationship between age and number of canonical syllables per utterance as judged by the human listener. Each data point represents the average number of canonical syllables per utterance for a single child on a single recording day. The right plot shows the relationship between age and number of canonical syllables per utterance at the day level as estimated by the automated method. The automated method shows a pattern of increase in canonical syllables with age, replicating what has been found with human listener judgments, both in the present dataset, as shown on the left, and in previous studies.*

$p < .001$ (Fig. 1. Machine and human counts of syllables of any type (canonical and non-canonical both included) were strongly correlated, $\rho = .70$, $p < .001$ (Fig. 2). The figures show increasing machine syllable counts as the human listener syllable counts increase, with a bias on both tasks toward over-counting in low-syllable utterances. The stronger ability to predict humans' total syllable counts compared to prediction of specifically canonical syllable counts can be seen in the smaller error bars in Fig. 2 compared to those in Fig. 1.

3.4. Correlation between syllable counts and age

We tested for a correlation between human syllable counts and age, to ensure our data replicate the pattern found in previous studies. Indeed, we found that the average human-judged canonical syllable count per utterance for a given child on a given recording day strongly predicted that child's age in days, $r = .59$, $p < .001$ (left side of Fig. 3). The average human-judged canonical syllable count per utterance divided by the total syllable count per utterance also shows a pattern of increase with age, as expected from prior published studies, $r = .54$, $p < .001$. Interestingly, although previous studies have not focused on increase in total number of syllables as a function of age, we found this to also be a pattern in our data, albeit weaker than the canonical rate trends, $r = .36$, $p < .001$.

We then asked whether the same pattern held when automated syllable counts (combined method) were used in place of human counts. A pattern of increase in canonical syllables per utterance with age was found, $r = .37$, $p < .001$ (right side of Fig. 3). This is weaker than what we found with the human judgments, but in the same direction and still very statistically significant. The correlation between age and automated canonical to total syllable ratio yielded very similar results, $r = .38$, $p < .001$. This suggests that the automated methods may be able to discern some differences between canonical and non-canonical syllable types. Interestingly, in keeping with the automated method's better performance matching human total syllable counts, its ability to predict age based on total syllables per utterance was similar to that obtained using human judgments and similar to the combined automated method's ability to predict age based on canonical syllable measures, $r = .32$, $p < .001$.

4. Discussion

We found a set of existing syllable detection algorithms to be capable of producing both total syllable counts and canonical syllable counts that reliably correlated with human listener judgments. When all methods were combined in a generalized additive model, total syllable count reliability was comparable to human inter-rater reliability. Reliability for canonical syllable counts was lower and not comparable to human-human reliability. Performance of the combined model on canonical syllable counting was not superior to performance of the Sphinx consonant count alone (Table 1).

Both our human listener judgments and our automated method showed an increase in syllable counts with age. This replicates the prior finding that infant canonical syllable production increases in relative frequency from 4–18 months of age (Oller et al., 1997). It demonstrates the validity of the automated method as applied to the study of infant speech development.

Future work should investigate using the richer sets of features available as part of the existing syllable detection methods. Another exciting future direction would be to train machine-learning-based (e.g. neural networks-based) ASR methods on the infant sounds and human judgments provided here. These may lead to better performance counting canonical syllables.

5. Acknowledgements

We thank Alessandra Fontana, Gabriela Macedo, Cecilia Valdovinos, Jessica Ross, and Gina Pretzer for help organizing the recordings and with the human listener judgments. ASW's work on the project was supported by NSF BCS-1529127.

6. References

[1] D. K. Oller, "The emergence of the sounds of speech in infancy," in *Child Phonology: Volume 1: Production*, G. H. Yeni-Komshian, J. F. Kavanagh, and C. A. Ferguson, Eds. New York: Academic Press, 1980.

[2] R. E. Stark, "Stages of speech development in the first year of life," in *Child Phonology: Volume 1: Production*, G. H. Yeni-Komshian, J. F. Kavanagh, and C. A. Ferguson, Eds. New York: Academic Press, 1980.

[3] F. J. Koopmans-van Beinum and J. M. van der Stelt, "Early stages in the development of speech movements," in *Precursors of Early Speech*, B. Lindblom and R. Zetterström, Eds. New York: Stockton Press, 1986.

[4] L. Roug, I. Landberg, and L.-J. Lundberg, "Phonetic development in early infancy: A study of four swedish children during the first eighteen months of life," *Journal of Child Language*, vol. 16, no. 1, pp. 19–40, 1989.

[5] E. H. Buder, A. S. Warlaumont, and D. K. Oller, "An acoustic phonetic catalog of prespeech vocalizations from a developmental perspective," in *Comprehensive Perspectives on Child Speech Development and Disorders: Pathways from Linguistic Theory to Clinical Practice*, B. Peter and A. M. MacLeod, Eds. New York: Nova Science Publishers, 2013.

[6] D. K. Oller, R. E. Eilers, M. L. Steffans, M. P. Lynch, and R. Urbano, "Speech-like vocalizations in infancy: An evaluation of potential risk factors," *Journal of Child Language*, vol. 21, no. 1, pp. 33–58, 1994.

[7] D. K. Oller, R. E. Eilers, R. Urbano, and A. B. Cobo-Lewis, "Development of precursors to speech in infants exposed to two languages," *Journal of Child Language*, vol. 24, no. 2, pp. 407–425, 1997.

[8] P. Menyuk, J. Liebergott, and M. Schultz, "Predicting phonological development," in *Precursors of Early Speech*, B. Lindblom and R. Zetterström, Eds. New York: Stockton Press, 1986.

[9] J. L. Locke and D. M. Pearson, "Linguistic significance of babbling: Evidence from a tracheostomized infant," *Journal of Child Language*, vol. 17, no. 1, pp. 1–16, 1990.

[10] K. M. Bleile, R. E. Stark, and J. S. McGowan, "Speech development in a child after decannulation: Further evidence that babbling facilitates later speech development," *Clinical Linguistics & Phonetics*, vol. 7, no. 4, pp. 319–337, 1993.

[11] E. Patten, K. Belardi, G. T. Baranek, L. R. Watson, J. D. Labban, and D. K. Oller, "Vocal patterns in infants with autism spectrum disorder: Canonical babbling status and vocalization frequency," *Journal of Autism and Developmental Disorders*, vol. 44, no. 10, pp. 2413–2428, 2014.

[12] C. Stoel-Gammon and K. Otomo, "Babbling development of hearing-impaired and normally hearing subjects," *Journal of Speech and Hearing Disorders*, vol. 51, no. 1, pp. 33–41, 1986.

[13] D. K. Oller and R. E. Eilers, "The role of audition in infant babbling," *Child Development*, vol. 59, no. 2, pp. 441–449, 1988.

[14] B. L. Davis, H. M. Morrison, D. von Hapsburg, and A. D. Werner, "Early vocal patterns in infants with varied hearing levels," *The Volta Review*, vol. 105, no. 1, pp. 7–27, 2005.

[15] S. Nathani Iyer and D. K. Oller, "Prelinguistic vocal development in infants with typical hearing and infants with severe-to-profound hearing loss," *The Volta Review*, vol. 108, no. 2, pp. 115–138, 2008.

[16] D. J. Ertmer and S. Nathani Iyer, "Prelinguistic vocalizations in infants and toddlers with hearing loss: Identifying and stimulating auditory-guided speech development," in *The Oxford Handbook of Deaf Studies, Language, and Education*, M. Marschark and P. E. Spencer, Eds. Oxford: Oxford University Press, 2010.

[17] D. K. Oller, P. Niyogi, S. Gray, J. A. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. F. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 30, pp. 13 354–13 359, 2010.

[18] D. K. Oller and H. L. Ramsdell, "A weighted reliability measure for phonetic transcription," *Journal of Speech, Language, and Hearing Research*, vol. 49, no. 6, pp. 1391–1411, 2006.

[19] "IVICT (Infant Vocalization Interactive Coding Trainer)," http://www.babyvoc.org/IVICT.html, 2015.

[20] S. Boyce, H. Fell, L. Wilde, and J. MacAuslan, "Automated tools for identifying syllabic landmark clusters that reflect changes in articulation," in *Models and Analysis of Vocal Emissions for Biomedical Applications*, C. Manifredi, Ed. Firenze: Firenze University Press, 2011.

[21] S. Boyce, H. Fell, and J. MacAuslan, "Speechmark: Landmark detection tool for speech analysis," in *Interspeech*, 2012.

[22] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, 2002.

[23] H. J. Fell and J. MacAuslen, "Vocalization analysis tools," in *Proceedings of the 7th Annual Workshop for Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, C. Manifredi, Ed. Florence: Firenze University Press, 2005.

[24] H. J. Fell, J. MacAuslen, L. J. Ferrier, S. G. Worst, and K. Chenausky, "Vocalization age as a clinical tool," in *Proceedings of the 7th International Conference on Spoken Language Process (ICSLP 02)*, Denver, 2002.

[25] N. H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, 2005.

[26] A. S. Warlaumont and M. K. Finnegan, "Learning to produce syllabic speech sounds via reward-modulated neural plasticity," *PLOS ONE*, vol. 11, no. 1, p. e0145096, 2016.

[27] M. Coath and S. L. Denham, "Robust sound classification through the representation of similarity using response fields derived from stimuli during early experience," *Biological Cybernetics*, vol. 93, no. 1, pp. 22–30.

[28] M. Coath, S. L. Denham, L. Smith, H. Honing, A. Hazan, P. Holonwicz, and H. Purwins, "An auditory model for the detection of perceptual onsets and beat tracking in singing," in *Neural Information Processing Systems, Workshop on Music Processing in the Brain*, Vancouver, 2007.

[29] M. Coath, S. L. Denham, L. M. Smith, H. Honing, A. Hazan, P. Holonowicz, and H. Purwins, "Model cortical responses for the detection of perceptual onsets and beat tracking in singing," *Connection Science*, vol. 21, no. 2-3, pp. 193–205.

[30] "CMU Sphinx," http://cmusphinx.sourceforge.net/, 2015.

[31] D. Xu, J. A. Richards, and J. Gilkerson, "Automated analysis of child phonetic production using naturalistic recordings," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 5, pp. 1638–1650, 2014.