Acoustic phonetic classification of primary infant protophones

Eugene H. Buder[1], D. Kimbrough Oller[1]

[1]University of Memphis, Memphis, TN, United States
ehbuder@memphis.edu

ABSTRACT

Infants' utterances are readily classified by researchers and parents into phonatory-based categories such as vocant, squeal, and growl. This classification has proceeded on auditory grounds referencing features such as pitch and voice quality, but there have been no systematic efforts to identify appropriate acoustic phonetic dimensions underlying these vocalization classes: the current study addresses that problem. Fundamental frequency studies have been numerous, and observations regarding non-modal regimes are also longstanding; here those dimensions, reflecting well-recognized features of these speech-like utterances, are integrated. Analysis of a corpus representing utterances produced by 3 infants, in recordings at ~3, 6, and 9 months, reveals a feature set adequately classifying these types relative to human judges: The phonatory protophone classes distinctly cluster in a 3D space defined by f_0 mean, f_0 SD, and a regime-based scale aligned with glottal closure. Classification of infant utterances in these terms reveals distinct development trends.

Keywords: Infant Vocalization, Phonation, Development, Voice Quality

1. INTRODUCTION

Clear phonation-based categories of utterance are so typical in human infancy as likely to be universal, and terms such as squeal, growl, and vocant (vowel-like) have been used in language development literature for at least 50 years [1-3]. While not considered fully phonological and inherently fuzzy, these categories do exhibit distinctive features that can be defined acoustic phonetically, with the greatest elaboration seen in phonatory aspects [4]. Over the same period, acoustic phoneticians have observed evidence in the harmonic structure of infant voices of numerous alternative patterns of vibration [5]. These phonatory patterns have since come to be understood within the framework of non-linear dynamics [6], and dubbed vibratory regimes in this context [7].

The occurrence of vibratory regime shifts during infant vocalization is inherently categorical, yielding categories expected from general non-linear dynamics theory, from harmonic to subharmonic, or even to non-periodicity, e.g. to chaotic vocal fold vibrations or cessation of voicing, all regularly observed in infant phonation [6, 8-10] Systematic analyses of harmonic frequency distributions or amplitudes [11] can be used to fully characterize this aspect of infant voice quality. Furthermore, the arrangement of regimes in such systems is expected to order according to some underlying parameter; in the case of regimes observed in infancy, glottal closure (produced by adduction, aerodynamic force variation, particular configuration of cover/body tissues, etc.) appears to be a prime candidate [12].

In perceptual terms, the most elementary distinction among the earliest protophones is pitch; vocants mid-range, growls lower and squeals higher. Traditionally, many vocal development researchers have been interested in f_0 characteristics [3, 8, 13-15], especially because of anatomical developments of the larynx [16]. Yet virtually none of these studies distinguished amongst protophones, or considered f_0 ambiguities introduced by harmonic regime variations. Another shortcoming of prior reports utilizing f_0 measures is failure to report within utterance variability, overlooking the possible importance of f_0 variability. Finally, intensity is rarely explored in infant vocal development.

In summary, study of the acoustic structure of early protophones is overdue, while acoustic parameter candidates have now been identified: This acoustic phonetic grounding of infant vocal categories helps re-orient basic questions in infant vocal development by operationalizing an acoustically-based utterance typology.

2. METHODS

2.1. Materials

The corpus of 2,312 vocalizations analysed here represents virtually all of the non-vegetative vocalizations heard by trained personnel as either a vocant, growl, or squeal utterance produced by three female infants during pairs of 20-minute lab-based sessions, recorded at approximately 3, 6, and 9 months of age. The infant and caregiver were both fitted with FM wireless microphones with Countryman Associates MEMWF0WNC capsules in vests, configured to minimize friction noise and maintain as consistent mic-mouth distances as possible. Calibration tones were recorded during each

session at known dB levels for use as reference levels for converting RMS voltages in recordings to dB levels comparable across sessions.

2.2. Protophone Coding

Human coding of the targeted vocalization types was conducted in the AACT environment [17] by either the first author or a highly experienced PhD student following procedures outlined in [2], and focusing exclusively on speech-like vocalizations. Segmentations of ongoing vocalizations followed breath-group phrasing principles [18], exclusion of protracted breathy offsets and of vocalizations shorter than 50 ms or effectively at floor amplitudes. Inter-judge Kappa reliability on 20% of the dataset was 0.66, establishing a benchmark for acoustic classification; judge's disagreements in coding were subsequently resolved by consensus coding.

2.3. Acoustic Analyses

2.3.1 fo

The AACT environment implements the ActiveX library of the TF32 program, including implementation of its waveform-correlation based pitch determination algorithm's parameter controls and hand editing tools [19]. Another difficulty with prior f_o literature has been variable results due to different ways of handling the many-octave range of infant phonation.

Here, analysis parameter variation, followed by hand-marking of glottal epochs, was used to yield carefully validated data series ranging from 15 Hz to over 4 kHz. Importantly, in the face of harmonic ambiguities, selection of appropriate f_o was guided by prior regime coding as described below, e.g. subharmonics should not be tracked as they were already coded as such, f_o in chaos and stops is null by definition, etc. At the utterance level, the f_o dataset analysed here ranges from means of 43 Hz (a growl, dominated by pulse register) to 1740 Hz (a squeal, dominated by loft register).

2.3.2 dB

Mean RMR values were extracted at the utterance level and converted to standardized dB with reference to the recorded calibration tones.

2.3.3 Regimes

Regime coding, adopting procedures documented in [7], involved inspection of narrowband spectrograms aided by auditory judgments and inspection of waveforms, in order to classify all segments of each protophone as one of 8 types representing distinct vocal fold vibration patterns: modal, pulse, subharmonics, biphonation, chaos, 'c-stop' (overadduction), 'o-stop' (underadduction), and "hi-modal".

In the absence of any prior reporting on a distinct loft (or 'falsetto') register in infancy, the "hi-modal" code was actually a placeholder for high-f_o segments for further consideration of this possibility. Subsequent research, investigating pitch breaks within infant utterances from modal into loft, identified significant differences across those breaks in the relative amplitudes of the first two harmonics (H1-H2), consistent with variations between modal and falsetto registers known in the human adult literature to involve distinctive dynamics of VF vibration [11]. This work provided threshold f_o and H1-H2 values that were implemented in the current dataset to discriminate loft segments (516 Hz and 1.6 dB respectively), and this register distinction is then incorporated under the general rubric of regime type.

2.3.4 Closure Scale

Regime codes subsegment protophone codes; these can be as short as 50 ms, extend throughout the whole vocalization, or break the vocalization into many units including the possibility of regime repeats. To associate regimes with protophones it was therefore necessary to 'score' vocalizations according to the occurrence of regimes within them, and aggregating their percentages of occurrence within each vocalization served this purpose. This did not, however, solve a 'sparsity' issue: Most vocalizations contained just one regime and none contained them all. A full aggregation of the categorical codes onto some relatively continuous metric addresses this issue.

Aggregation of regimes can be motivated by a principle by which they might be expected to occur along some interpretable scale, and models with parametric control of non-linear dynamic systems' driving forces motivate such principles. Principles of phonation point towards respiratory flow and glottal closure as likely parameters. Observing that over- and under-compressed glottal configurations provides natural poles for a 'closure scale' (while not ruling out flow or other additional parameters), this scaling was explored amongst the entire array of regimes via χ^2 analyses. Ultimately regime associations with protophones provided the scaffolding needed for this exploration.

Details of that analysis are provided in [12], with an emphasis on highly significant +/- standardized deviates within tables associating protophone categories with regime occurrence tallies, as seen in

Table 1 here. Most notably, clear associations emerge for squeal (positively with loft and o-stops), and for growls (positively with subharmonics, pulse, chaos, and c-stops). Aggregating the percentages of those groupings helped to mitigate the sparsity issue (see Table 1 note), and observing that all non-modal regimes were negatively associated with the vocant type, scoring modal as a 0 helped to complete the scale. (Only biphonation is neglected, due to its ambiguous associations but also its sparsity.)

Table 1: Chi-square and standardized deviate statistics assessing the associations of individual regime codes with protophone types.

			Standardized Deviates		
Regime	n[1]	χ^2	growl	vocant	squeal
Loft	296	965	-3.1	-11.0	+26.6
Subharmonics	308	349	+14.8	-8.9	+2.3
Pulse	362	297	+13.9	-8.6	ns
C-Stop	157	110	+8.9	-4.7	ns
Biphonation	106	156	+4.6	-6.7	+9.0
Chaos	90	137	+10.2	-5.3	ns
O-Stop	62	124	ns	-4.3	+10
Modal	1599	42.7	ns	ns	-3.2

[1]n = the number of protophones containing an instance of the regime code.

Scaling the current dataset in this manner, a continuum is obtained that ranges -/+1, and 1614 (70%) of the vocalizations receive a non-0 value. Technically, the values might better be dubbed 'opening' rather than compression due to the polarity of the scale, but the underlying concept remains the same.

3. RESULTS

3.1 Parameter Identification

Exploration of the four candidate parameters identified above was initially conducted by logistic regression analyses with protophone type as the 3-way categorical outcome and candidate parameters as predictors (f_o mean and SD values log transformed to correct for skewness). An optimal model was identified retaining f_o mean, f_o SD, and the closure scale as predictors, obtaining 85% overall classification success. While addition of the dB mean value added a marginally significant increment in log-likelihood, this resulted in no improvement to classification, and it improved χ^2 from 2137 only to 2141. Based on these diagnostics, and conveniently for visualization purposes, a 3D space was created—see Figure 1.

Assessed by Cohen's Kappa against the human coding, the 3-parameter classification yielded 0.67; virtually identical to the reliability with which human coders matched one another, thereby affirming the full adequacy of this model. Subsequent appraisal of Hedge's *g* effect sizes in 2-way contrasts amongst protophone pairs also affirmed that each parameter was operating independently: All three yielded large (>1) effects for distinguishing squeal/vocant and squeal/growl, and while effects remained large for all three in the vocant/growl distinction, the closure scale effect size was twice that of the f_o parameters.

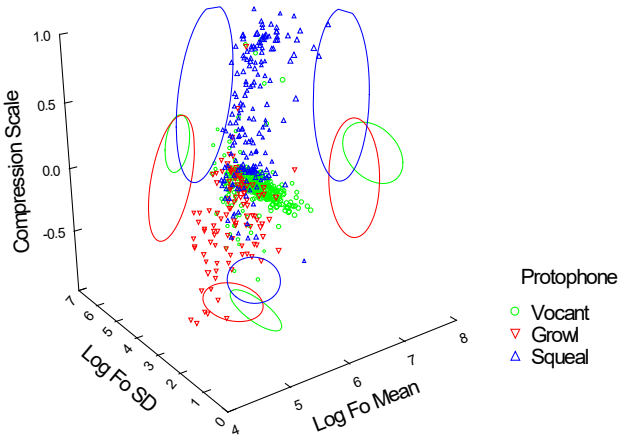

Figure 1: Scatterplot of all vocalizations in the three-dimensional space defined by the logistic regression predictors optimally classifying these vocalizations by protophone type. Ellipses projected onto two-dimensional facets are centered on class means and with axes defined by standard deviations to encompass data with probabilities equal to 0.6827 ("Compression" = "closure").

3.1 Developmental observations

All four parameters were retained for further exploration of developmental trends. Given the small number of infants included, the observations made here are not claimed to generalize. The purpose is rather to demonstrate applicability of the phonetic parameters identified in the proposed protophone classification model, as driven by the sufficient power of this dataset and its sampling strategy. The infants in this dataset were all typically developing and, as among the first to have been recorded, all happened to be female, but should be represented as a random factor in variance

modelling, so a mixed-model analysis was appropriate for examination of age differences in these parameters with protophone codes factored in as well.

Key outcomes from that analysis were: 1) With all other effects accounted for, the three primary f_o- and regime-based parameters significantly accounted for the main effect protophone variation, while dB mean did not; and 2) As a simple main effect, no Age variance was explained; yet, 3) Age × Protophone variance explained was significant, driven by numerous specific interactions in all three primary parameters. Figure 2 displays these effects for two parameters that merit further commentary here.

While, as in prior literature on the first years of life, no age effects were obtained for average vocalization pitch, drops in f_o are observable in *non-squeal* vocalizations, and (as seen by miniscule error bars in vocants) these effects were especially consistent in that protophone type. Examining vocant f_o cross age, small but significant age-difference Cohen's *d* effect sizes on the order of 0.31 are obtained. Furthermore, when vocant values were converted back to Hz the drop in those vocalizations during the 2nd half year of life was 17 Hz (5 %), a period when vocal tract growth would be expected to cause such effects.

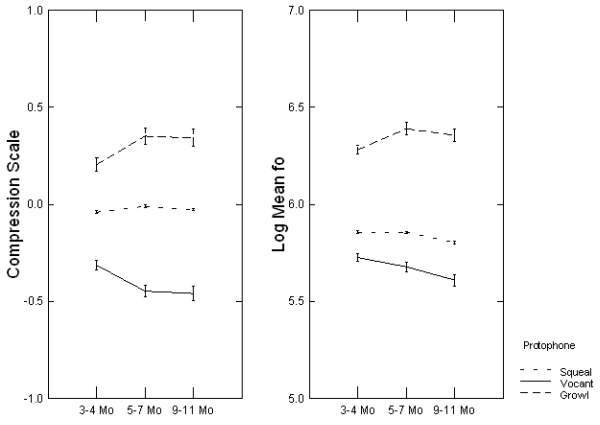

Figure 2: Line charts depicting protophone-specific age effects for two phonetic parameters; see legend for protophone line weights. Error bars are SEs.

The display of closure ("compression") scale values in the left panel of Figure 2 affirm the factors driving Age × Protophone interaction results with this parameter as well: While vocants are steady in this respect, it is interesting to observe the divergence of values occurring from the first to second age periods, as it would appear to be strong affirmation of the awareness among vocal development scientists that infants across the age range considered here appear to engage in vocal play, practicing variations on the phonatory parameters that yield the apparent early protophone categories [1, 20]. It may also evince operation of the respiratory-laryngeal system described recently as among the least well understood "Developmental Functional Modules" underlying emergent stages of early human vocal development [21].

4. DISCUSSION

The present analysis provides a rationale for identifying fairly straightforward acoustic parameters, f_o mean, f_o SD, and the identified "closure scale," as underlying protophone classification but also vocal development itself, with findings that, pending affirmation among more infants, should identify development trends more precisely than prior approaches which specifically neglected protophone or vibratory regime distinctions. Because glottal closure incorporates parameters such as medial compression and aerodynamics, the framework aligns well with other phonatory-based approaches [22, 23].

The outcome that these parameters model the human percept is by design, so the framework they provide for categorizing infant protophones implies that these specific parameters are utilized by the auditory system; this could of course be tested by psychoacoustic research paradigms. And it has always been expected that some acoustically-based model should be able to work well, since human judges have consistently agreed well on the identification of the three phonatory protophone types. What is quite surprising, however, is that, unlike in many cases where human coding has been found to be vastly superior to automated analyses of vocal activity based on acoustic evidence only [2], the present analysis yielded an outcome where the acoustically-based classification approximated its own gold standard (the human coding) precisely.

Having reached that essential benchmark, next steps following this approach should consider automation of spectral harmonic analyses (e.g. [24]) to side-step human coding of regimes, optimally guiding then guiding parameter selection in PDAs for f_o analysis. Such developments would facilitate applications of the model to more infants to affirm the trends observed here, and the resulting larger datasets should then help create improved processing, or at least greater tolerance for error in acoustic input, while still generating significant outcomes. A mapping of f_o and phonatory harmonic structures in infant vocal output should thereby lay groundwork for new approaches to the understanding of vocal development.

Acknowledgment: This work was supported by NIDCD R01001027 to D. K. Oller.

7. REFERENCES

[1] D. K. Oller, *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000, p. 428.

[2] D. K. Oller, E. H. Buder, H. Ramsdell, A. S. Warlaumont, L. B. Chorna, and R. Bakeman, "Functional flexibility of infant vocalization and the emergence of language," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 110, pp. 6318-6323, 2013, doi: 10.1073/pnas.1300337110.

[3] M. Z. Laufer and Y. Horii, "Fundamental frequency characteristics of infant non-distress vocalizations during the first twenty-four weeks," *Journal of Child Language,* vol. 4, pp. 171-184, 1977.

[4] 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 Speech Sound Development and Disorders: Pathways from Linguistic Theory to Clinical Practice*, B. Peter and A. A. N. MacLeod Eds. Hauppage, NY: Nova Science Publishers, Inc., 2013, pp. 103-134.

[5] P. A. Keating and R. Buhr, "Fundamental frequency in the speech of infants and children," *J Acoust Soc Am,* vol. 63, pp. 567-571, 1978.

[6] W. Mende, H. Herzel, and K. Wermke, "Bifurcations and chaos in newborn infant cries," *Physics Letters A,* vol. 145, pp. 418-424, 1990 1990.

[7] E. H. Buder, L. Chorna, D. K. Oller, and R. Robinson, "Vibratory regime classification of infant phonation," *J Voice,* vol. 22, pp. 553-564, 2008.

[8] R. D. Kent and A. D. Murray, "Acoustic features of infant vocalic utterances at 3, 6, and 9 months," *J Acoust Soc Am,* vol. 72, pp. 353-365, 1982.

[9] R. Buhr and P. Keating, "Spectrographic effects of register shifts in speech production," *The Journal of the Acoustical Society of America* vol. 62, p. 25, 1977.

[10] M. P. Robb and J. H. Saxman, "Acoustic observations in young children's non-cry vocalizations," *J Acoust Soc Am,* vol. 83, pp. 1876-1882, 1988.

[11] E. H. Buder, V. F. McDaniel, E. R. Bene, J. Ladmirault, and D. K. Oller, "Registers in infant phonation," *J Voice,* pp. 1-12, 2018.

[12] E. H. Buder and D. K. Oller, "Glottal closure as a parameter underlying phonatory regime occurrence in infancy," submitted for review.

[13] S. Amano, T. Nakatani, and T. Kondo, "Fundamental frequency of infants' and parents' utterances in longitudinal recordings," *J Acoust Soc Am,* vol. 119, pp. 1636-1647, 2006.

[14] M. P. Robb and J. H. Saxman, "Developmental trends in vocal fundamental frequency of young children," *Journal of Speech, Language, and Hearing Research,* vol. 28, no. 3, pp. 421-427, 1985.

[15] E. Scheiner, K. Hammerschmidt, U. Jurgens, and P. Zwirner, "Acoustic analyses of developmental changes and emotional expression in the preverbal vocalizations of infants," *J Voice,* vol. 16, no. 4, p. 509, 2002.

[16] R. D. Kent and H. K. Vorperian, "Development of the craniofacial-oral-laryngeal anatomy: A review," *Journal of Medical Speech-Language Pathology,* vol. 3, no. 3, pp. 145-190, 1995.

[17] *AACT - Action Analysis Coding and Training Software*. (2020). Intelligent Hearing Systems Corp., Miami, FL.

[18] M. P. Lynch, D. K. Oller, M. L. Steffens, and E. H. Buder, "Phrasing in prelinguistic vocalizations," *Developmental Psychobiology,* vol. 1, pp. 3-25, 1995.

[19] *TF32*. (2018). University of Wisconsin-Madison, Madison, WI.

[20] F. J. Koopmans-van Beinum and J. M. van der Stelt, "Early stages in the development of speech movements.," in *Precursors of early speech*, R. Zetterstrom Ed. New York: Stockton Press., 1986, pp. 37-50.

[21] R. D. Kent, "Developmental functional modules in infant vocalizations," *J Speech Lang Hear Res,* vol. 64, no. 5, pp. 1581-1604, May 11 2021, doi: 10.1044/2021_JSLHR-20-00703.

[22] J. H. Esling, S. R. Moisek, A. Benner, and L. Crevier-Buchman, *Voice quality: The laryngeal articulator model*. Cambridge, UK: Cambridge University Press, 2019.

[23] M. P. Robb, F. Yavarzadeh, P. J. Schluter, V. Voit, W. Shehata-Dieler, and K. Wermke, "Laryngeal constriction phenomena in infant vocalizations," *Journal of Speech, Language, and Hearing Research,* vol. 63, no. 1, pp. 49-58, 2020.

[24] G. Ramsay, "Multitaper harmonic analysis of infant vocalizations," vol. 144, no. 3, pp. 1767-1768, 2018, doi: 10.1121/1.5067818.