VOCALIZATION AGE AS A CLINICAL TOOL

Harriet J. Fell College of Computer Science, Northeastern University
Joel MacAuslan Speech Technology and Applied Research, Lexington, Massachusetts
Linda J. Ferrier Speech-Language Pathology and Audiology, Northeastern University
Susan G. Worst College of Computer Science, Northeastern University
Karen Chenausky Speech Technology and Applied Research, Lexington, Massachusetts
eva@harrietfell.com JoelM@S-T-A-R-corp.com

ABSTRACT

The Early Vocalization Analyzer (EVA) is a computer program that automatically analyses digitized acoustic recordings of infant vocalizations. Using the landmark detection theory of Stevens et al for the recognition of phonetic features in speech, EVA detects syllables in vocalizations produced by infants. Landmarks are grouped into standard syllable patterns and syllables are grouped into utterances. Statistics derived from these groups and the underlying features are used to derive a "vocalization age" and two specific screening rules that can clinically distinguish infants who may be at risk for later communication or other developmental problems from typically developing infants in the six to fifteen month age range.

1. BACKGROUND

Considerable research supports the position that infant vocalizations effectively predict later articulation and language [1, 2, 5, 10, 11, 13]. Intervention to encourage babbling activity in at-risk infants is frequently recommended. However, research and clinical diagnosis of delayed or reduced babbling have so far depended on time-consuming and often unreliable perceptual analyses of tape-recorded infant sounds. The acoustic analysis of infant cry has been examined as a diagnostic index of the child's neurological development [7, 12]. But non-cry vocalizations, which may be a more sensitive index of development because of their roots in different developmental domains, are not yet widely used clinically as predictors of future communication performance. Kent and Murray [6, p.412] state, "Whereas a considerable amount has been written about the clinical implications of acoustic analysis of infant cry, . . . relatively little has been written about similar implications for comfort-state vocalizations during the cooing, and babbling stages." This lack of research information arises presumably because traditional methods of analysis are time-consuming or unreliable. While acoustically analyzing infant sounds has provided important information on the characteristics of infant vocalizations, use of this information in automatic analysis is relatively new [3, 4].

2. OVERVIEW

We have developed the Early Vocalization Analyzer (EVA), a computer program to automatically analyze recorded samples of infant vocalizations. We see EVA as a tool to build a normative data base of pre-speech infant vocalizations and to provide the basis of a screening test, for use in hospitals and clinics, to evaluate which infants are at risk for later communication or other developmental problems. EVA is successful at detecting syllable boundaries in pre-speech vocalizations [4]. It can also reliably count utterances and classify them as to high, medium, or low F_0; and to short, medium, or long duration [3]. In this paper, we show that EVA can also provide a clinical tool to distinguish infants at risk for communication problems from typically developing infants.

3. SUBJECTS

Nine typically developing subjects (four male, five female) as well as five at-risk infants (four male, one female) were recorded on digital audio tape. In the latter group, one subject was diagnosed with apraxia, one with Down syndrome, one with hydrocephaly, and three (one of whom was premature) showed motor delay. Of the fourteen infants, two are African-American and one is Hispanic. All but the Hispanic infant have American-English-speaking parents. Each infant was recorded four to eight times for 40 minutes, at approximately monthly intervals, from six to thirteen months.

Subject	**Sex**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
T01	boy	v		v	v		v	v			
T02	boy		v	v		v	v	v			
T03	boy				v	v	v	v	v		
T04	girl		v	v		v	v				
T05	boy	c	c	c	c	c	c	c	c		
T06	boy		c	c		c	c	c	c	c	c
T07	boy	c	c	c		c	c	c	c	c	
T08	girl	c	c	c	c	x	c	c	c		
T09	girl	x	c	c	c	c	c	c	c		
T10	boy	v	v		v	v	v	v	v	v	
T11	girl	c	c	c	c		c	c	c		c
T12	girl	c	c	c	c	c	x	x	x		
T13	girl	v		c	c	c	c	c	c	c	
T14	boy	c	c	c	c	c	x	x	x		
Total		10	11	12	10	12	14	12	10	4	2

Table 1: Typically Developing Subjects and Months Recorded, c - calibration set v - validation set x - not yet processed

Subject	Sex	6	7	8	9	10	11	12	13	14	15	16	17
A2 motor delay	boy					x	x	x	x	x	x	x	x
A3 Down syndrome	boy			x		x	x	x	x	x			
A4 premature	girl			x	x	x	x	x	x	x	x		
A5 hydrocephaly	boy	x		x	x		x						
total		**1**	**0**	**3**	**2**	**3**	**4**	**3**	**3**	**3**	**2**	**1**	**1**

Subject	Sex	20	21	22	23	24	25	26	27	28	29
A1 apraxia	boy	x	x	x	x		x	x		x	x

Table II: At-Risk Subjects, Diagnosis, and Months Recorded

4. THE EVA SOFTWARE

The EVA software consists of three parts: the Landmark Detector, the Syllable and Utterance Analyzer, and the Vocalization Age Evaluator.

4.1. Landmark Detector

The landmark detector is built on the Liu-Stevens Landmark Detection program for adult speech founded on Stevens' acoustic model of speech production [8, 9, 15, 16]. Central to this theory are landmarks, points in an utterance around which listeners extract information about the underlying distinctive features. They mark perceptual foci and articulatory targets.
The program detects three types of landmarks:

glottis: marks the time when the vocal folds transition from freely vibrating to not freely vibrating (-g) or vice versa (+g)

sonorant: marks sonorant consonantal closures (-s) and releases (+s) (e.g., nasals)

burst: designates stop/affricate bursts (+b)and points where aspiration/frication ends (-b) due to stop closure

4.2. Syllable and Utterance Analyzer

Uses landmark types and times output by the Landmark Detection program to:

1. Remove landmarks from areas of the recording that have been corrupted by noise, as well as landmarks produced as artifacts of the process
2. Group landmarks into "syllables" of certain specific types such as "+g/-s/-g", using information about their order and spacing. For each recording session, creates a profile using the number of syllables, number of syllables types, and duration of each syllable type.
3. Group syllables into "utterances"--series of syllables occurring closely together--based on timing considerations. Describes the average number of syllables per utterance, as well as the number of utterances comprised of 1, 2, 3, and more syllables, for each subject and recording session.

4.2.1. Finding Syllables

The program identifies sequences of landmarks as syllables based primarily on their order. Thirty-eight possible syllables are recognized, plus a catchall category of "other". Eleven recognized syllables begin with +g:

+g/-g, +g/-g/-b
+g/+s, +g/+s/-g, +g/+s/-g/-b, +g/+s/-s, +g/+s/-s/-g,
+g/+s/-s/-g/-b
+g/-s, +g/-s/-g, +g/-s/-g/-b

Each of these eleven syllables may have a prefix of +b (yielding an additional eleven syllables) or by +b/-b (providing eleven more). Five syllables begin with +s:
+s/-g, +s/-g/-b, +s/-s, +s/-s/-g, +s/-s/-g/-b

If a landmark sequence does not precisely match any of the syllables listed above, then the syllable is classified as "other."

4.2.2. Finding Utterances

The program computes the number of utterances in each file, and in the group of files as a whole, as well as the average number of syllables per utterance. An utterance is a sequence of syllables in which gaps between syllables are no more than 200 milliseconds long.

4.3. Vocalization Age Evaluator

There are ninety-three syllable and utterance measurements that we consider in forming the vocalization age:

4 - Number of syllables per utterance
- 1, 2, 3, or (4 or more, grouped together)
38 - Number of syllables of each syllable type
3 - Number of syllables starting with +b, +g, +s
4 - Number of syllables ending with -b, -g, -s, +s
6 - Number of syllables with n landmarks, n =2 through 7
38 - Mean duration for each syllable type.

We also measure the standard deviations of the syllable durations (if any) for each syllable type.

In this initial form, these measures cannot be easily compared to each other: some are discrete and others continuous, some cover wide ranges and others narrow, some may be subject to substantial measurement error or day-to-day variability and others quite robust. Therefore all are converted to continuous measures of approximately equal uncertainty or measurement error. The result is a list of 93 values whose estimated measurement errors have approximately unit standard deviation.

For calibration, we compute the mean of the 93 values across all 63 sessions in the calibration set. This list of 93 mean values is subtracted from every session's own 93-element list and the set of differences is subjected to principal components (PC) analysis or, equivalently, singular-value decomposition [14]. Because of the previous attention to scaling by the measurement errors, singular values to be ignored (with their corresponding PCs) are just those of order unity and smaller. By suppressing PCs that merely describe small and unreliable features in the calibration data set, we are left with 34 of the possible 62 PCs for any subsequent fits.

The fit to chronological age produces residuals with a standard deviation of $\sigma = 1.56$ months. (The mean of the residuals is necessarily zero.) For comparison, this fit (subtracting the mean of the *calibration* set, of course) yields differences with a mean of -0.6–0.1 month (i.e., a mean delay of 0.6 month) and standard deviation of 2.02 months when applied to the 28-session validation set. As Figure 1 shows, this is virtually indistinguishable from the results for the calibration set itself and, because the mean difference is much smaller than 1.56 months, clinically undetectable.

5. DISCUSSION

Figure 1 shows the Vocalization Age versus the Chronological Age of all the infants in our study with different symbols for the three populations: the calibration set, the validation set, and the atypicals. In comparing the atypicals with the calibration and validation sets, we feel it is not appropriate to include the child with severe apraxia (sessions denoted by dots in Figure 3) as his chronological age at the recording sessions is outside the range of the other populations. The atypicals without this child show a delay in Vocalization Age of 3.7–0.6 months. The standard deviation of delays for this group is 2.6 months, showing (as might be expected) that this group is less homogeneous than the two sets of typically developing infants. Obviously, including the apraxic infant would only magnify the between-group differences (5.4 months delay, 3.5 months standard deviation) and strengthen the conclusions.

These differences are large enough and systematic enough that they provide the basis for clinical rules (see tables *III* and *IV*). Two simple versions, stated in a form suitable for a one- or two-session screening test, may be formulated:

An infant is (or is not) in the atypical group if any session (respectively, no session) shows a delay of at least 2 σ, or 3.1 months.

An infant is (is not) in the atypical group if any (respectively, no) two consecutive sessions both show delays of at least 1.5 σ, or 2.3 months.

	sessions	sessions with delay > 2σ	infants identified by this rule
Typical	91	4	4 of 15
Atypical	22	12	4 of 4

Table III: Rule 1

	pairs of consecutive sessions	consecutive sessions - delays both > 1.5σ	infants identified by this rule
Typical	76	0	0 of 15
Atypical	18	8	3 of 4

Table IV: Rule 2

In these Tables, "Typical" consists of the calibration and validation sets. We have kept the calibration that was derived from the calibration set. Because the validation set shows a small overall delay, this produces a slightly higher number of false alarms than if all Typical data were combined and used for a new calibration.

The data also support an "inverse" screen: An infant is *not* (is) in the atypical group if any (respectively, no) session shows *no* delay, i.e., shows performance at least at the Chronological Age level.

	number of sessions	number of sessions with vocalization age above chronological age	Number of infants identified (as *not* atypical) by this rule
Typicals	91	43	15 of 15
Atypicals	22	3	1 of 4

Table V: Inverse Rule

6. CONCLUSIONS

The vocalization age computed by the EVA software can clinically distinguish infants, six to fifteen months old, who may be at risk for later communication or other developmental problems from typically developing infants.

7. FUTURE WORK

Infants seem to go through periods of experimenting with pitch variation. They might gain little additional control of the oral articulators during these times and hence the computed vocalization age might show little increase during these intervals. EVA extracts information from the digitized recordings about the variation of fundamental frequency, an indicator of pitch. It may be appropriate to integrate this information into the vocalization age predictor.

We expect that attributes other than chronological age relate to the features we extract from the digitized acoustic recordings. For example, the presence of certain features might depend strongly on fine-motor control, allowing those features to be used in a measure of motor development. Others might correspond closely to later linguistic abilities, allowing the features to predict deficits in such abilities before the abilities themselves had emerged.

For updates on this work, see
<http://www.ccs.neu.edu/home/fell/index.html>.

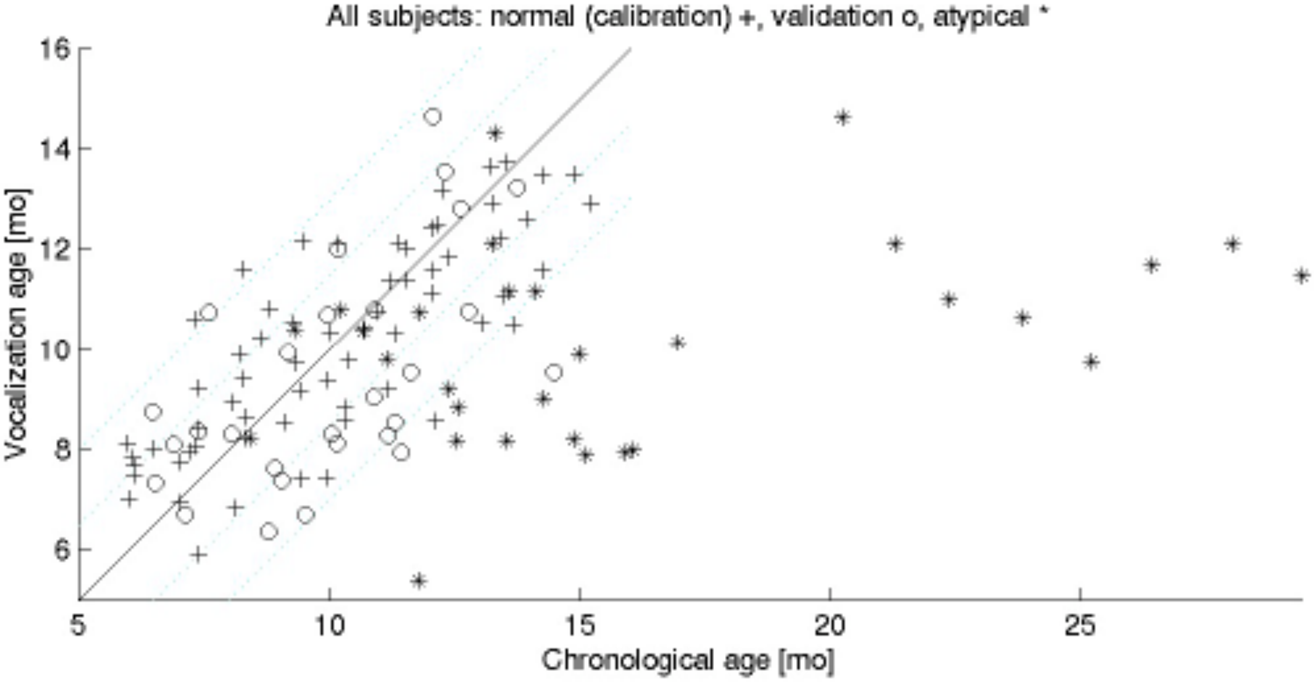

Figure 1. All Subjects

8. References

[1] Bayley, N. (1969). *Manual for the Bayley scales of infant development.* The Psychological Corporation, New York

[2] Capute, A. J., Palmer, F. B., Shapiro, B. K., Wachtel, R. C., & Accardo, P. J. (1981). Early language development: Clinical application of the language and auditory milestone scale. In *Language Behavior in Infancy and Early Childhood* (R.E. Stark ed.), *d.* Elsevier, New York.

[3] Fell, H.J., Ferrier, L.J., Schneider, D., & Mooraj, Z. (1996). EVA, An early vocalization analyzer: an empirical validity study. Proceedings of Assets □96, 57-61.

[4] Fell, H.J., MacAuslan, J., Ferrier, L.J. (1998, 1999) Automatic Babble Recognition for Early Detection of Speech Related Disorders. Behaviour & Information Technology, **18,** No. 1, 56-63.

[5] Jensen, T. S., Boggild-Andersen, B., Schmidt, J., Ankerhus, J., & Hansen, E. (1988). Perinatal risk factors and first-year vocalizations: Influence on preschool language and motor performance. Developmental Medicine and Child Neurology, **30**, 153-161.

[6] Kent, R. D. & Murray, A. D. 1991. Acoustic features of infant vocalic utterances at 3, 6 and 9 months. In R. J. Baken, & R. G. Daniloff (Eds.), *Readings in Clinical Spectrography of Speech* (pp. 402-414). New Jersey: Singular Publishing Group and Kay Elemetrics.

[7] Lester, B.M., and Zeskind, P.S. 1978. Brazelton scale and physical size correlates of neonatal cry features. *Infant Behavior and Development*, **1**, 393-402.

[8] Liu, S. (1994). Landmark detection of distinctive feature-based speech recognition, Journal of the Acoustical Society of America, **96**, 5, Part 2, 3227.

[9] Liu, S. (1995). *Landmark Detection for Distinctive Feature-based Speech Recognition*. Ph.D. Thesis. M.I.T. Cambridge, Massachusetts.

[10] Locke, J. L. (1993). *The child's path to spoken language*. Harvard University Press. Cambridge. Massachusetts.

[11] Menyuk, P., Liebergott, J., Shultz, M., Chesnick, M, & Ferrier, L.J. (1991). Patterns of Early Language Development in Premature and Full Term Infants. Journal of Speech and Hearing Research. **34**, 1.

[12] Murry, T. & Murry, J. (Eds.). (1980*). Infant communication: Cry and early speech.* Houston, Texas: College-Hill Press.

[13] Oller, D.K. & Lynch, M.P. (1992). Infant utterances and innovations in infraphonology: Toward a broader theory of development and disorders. In *Phonological Development: Models research implications* (Ferguson, C.A., Menn, L., & Stoel-Gammon, C. eds.), pp. 509-536. York Press. Timonium, Maryland.

[14] Press, W., S. Teukolsky, W. Vetterling, & B. Flannery. (1992). Numerical Recipes in C, 59-70. New York: Cambridge University Press.

[15] Stevens, K.N. (1992). Lexical access from features. *Speech Communication Group Working Papers, Volume VIII*, pp. 119-144, Research Laboratory of Electronics, M.I.T, Cambridge, Massachusetts.

[16] Stevens, K.N., Manuel, S., Shattuck-Hufnagel and Liu, S. (1992). Implementation of a model for lexical access based on features. *Proceedings of the International Conference on Spoken Language Processing*, **1**, 499-502.