USING EARLY VOCALIZATION ANALYSIS FOR VISUAL FEEDBACK

H. J. Fell[1], J. MacAuslan[2], C. J. Cress[3], L. J. Ferrier[4]
[1]College of Computer and Information Science, Northeastern University, MA, USA
[2]Speech Technology and Applied Research, Lexington, MA USA
[3]Communication Disorders Department, University of Nebraska – Lincoln, NE, USA
[4]Department of Speech Language Pathology and Audiology, Northeastern University, MA, USA

Abstract: **The Early Vocalization System (EVA) applies the Stevens landmark theory to infant vocalizations (babbles). The landmarks are grouped to identify syllable-like productions in these vocalizations. The visiBabble system processes vocalizations in real-time. It responds to the infant's syllable-like productions with brightly colored animations and records the landmark analysis. The system reinforces the production of syllabic utterances that are associated with later language and cognitive development. We report here on the development of the visiBabble prototype and our initial field-testing.**

Keywords : **acoustic analysis, babbles, landmarks**

I. INTRODUCTION

Communication skills are vital to educational and vocational success. Cerebral palsy, developmental apraxia (DAS), neurological insult/injury (e.g. head injury, encephalitis, meningitis), oral/motor dysfunction, cognitive impairments, tracheotomy, and deafness can all cause a child to be at risk for being non-speaking. A child having any of these or other syndromes may not be able to produce a sound when he or she wants to, may produce a limited range of sounds (often vowels and 1-2 consonants), or may not have learned to associate his or her sounds with meaningful referents [2]. During an intervention to promote speech-like vocalizations, non-speaking children tended to have difficulty initiating sounds and participating in vocal imitation play. They produced atypical sounds such as elongated vowels, distorted consonants, and non-speech sounds.

Because of the atypical sound production of infants in this population [8], traditional intervention strategies to prompt or respond to infant vocalizations may not be sufficient to promote change. Children at risk for being nonspeaking may produce a higher percentage of vowel-like sounds (*vocants*) and consonant-like sounds (*closants*) during later development than would be expected for typically developing children. Without strategies to detect and respond appropriately to these sound approximations, listeners may not be able to tailor their activities and responses appropriately to children's sound productions.

There is considerable research to support the position that infant vocalizations are effective predictors of later articulation and language abilities [7, 10, 12]. These studies have been carried out on normally developing children and on children with a variety of early diagnosed problems. These research studies emphasize the importance of early speech intervention for children at risk for being non-speaking. They also point out the difficulty of providing sufficient speech practice and feedback for children with such atypical speech patterns through traditional forms of intervention and interaction.

Closants and oral-cavity openings can be detected in the sound waveform from acoustic evidence of discontinuities in the spectrum of sound. These discontinuities have been called landmarks by some researchers of adult speech [9, 13]. Landmarks that result from the creation or release of a narrow constriction or closure along the vocal tract are also found in pre-linguistic vocalizations. We can hypothesize that the development of the ability to produce sounds exhibiting landmarks is a necessary skill underlying the production of syllables.

Vocants appear early in the vocalizations of infants and are characterized by slowly time-varying spectral patterns. These sounds result from movements of the tongue body, the jaw, and the lips, and are usually produced with the vocal folds positioned to vibrate. A variety of vowel-like sounds appear as the infant learns to control the positioning of these articulators. [1].

As babbling develops, the infant begins to coordinate control of the vocal folds and the velopharyngeal opening with control of the tongue blade and the lips, and the true consonants appear. In the landmark model, the larynx and the velum are considered secondary articulators, and they are "bound" to control by the primary articulators, in that implementation of the laryngeal and nasal features depends, in some ways, on the implementation of the primary articulator. This landmark model has proved useful in various applications concerning adult speech and has been successfully applied to analysis of infant vocalizations [3, 4, 5]. This analysis has, in turn, been used to formulate a "vocalization age" that clinically distinguishes between typically developing infants and infants at risk for later speech difficulties [6]. A vocalization age is a normative age-equivalence estimate

of the range of speech sounds (landmark sequences) expected for typically developing children.

The visiBabble system processes vocalizations in real-time. It responds to the child's syllable-like productions with brightly colored animations and records the landmark analysis. The system reinforces the production of syllabic utterances that are associated with later language and cognitive development. As a child interacts with visiBabble, the program collects and analyzes the infant's utterances so that it can be used by a child as a toy/trainer or as a clinical or research implement.

II. METHODOLOGY

A. The visiBabble System

The visiBabble system includes a modern notebook computer (Dell Inspiron, 2.4 GHz Pentium 4 running Windows XP), a microphone, a 15" flat-panel display, and software, which carries out the following functions:

- Landmark detection – detects landmarks in a child's vocalizations in real-time.
- Graphic feedback -- provides real-time visual response to sound input;
- Data collection – records each session and saves the result as a wav file, collects data on the types and duration of vocalizations produced;
- Experimental formats -- allows the system to run and data to be collected in single-case study formats.

B. Finding Landmarks

Our landmark detector is based on Stevens' acoustic model of speech production [13]. 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 start (+g) and stop (-g) vibrating;

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

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

The visiBabble system can track simple aspects of the acoustic signal in real time, based on a low-resolution spectrogram. That is, the signal is sampled at 16 kHz and analyzed into a small number, nominally 64, of separate, frequency intervals of ~256 Hz each. A 16 kHz rate provides information up to 8 kHz, sufficiently high to include at least 3-4 formants for an infant and to show the distinction between voicing and other speech sounds: fricatives, stop releases, bursts, etc. (These parameters are suitable for using the FFT and impose no delay of their own beyond 4 ms, i.e., 1/256-th of one second.) The visiBabble system uses only one-half of these intervals because the others differ only in phase.

The spectral intervals are grouped into six broad bands. An energy waveform is constructed in each of the six bands, the time derivative of the energy is computed, and peaks in the derivative are detected. These peaks represent times of abrupt spectral change in the six bands. Energy in bands 2 (1200 - 2500 Hz.) and 3 (1800 - 3500 Hz), e.g., provides evidence of voicing or, in some cases, of bursts. The distinction between these is readily made in the time domain (voicing persists much longer than bursts) as well as by appeal to information in the other spectral bands: voicing provides a power spectrum that decays with frequency approximately as $1/\text{frequency}^2$, whereas most other speech sounds have flatter spectra.

For the poorly formed or unstable closants and vocants typical of infants, wide frequency bands are well suited to recognition: Higher frequency resolution would require averaging over bands anyway. It would require spending more time computing and – worse – more time sampling the signal for the initially higher resolution.

C. Graphic Feedback

The visiBabble prototype responds to the child's utterances with five different brightly colored animations that cycle to avoid habituation: (a train, a bird, a frog and two cartoon creatures that move across the screen). It responds to the start of each syllable it detects by advancing the current animation one step.

It determines that a syllable has started either by voicing onset or by a voiced closant that occurs at least 100 ms after start of the previous syllable. Admittedly, a syllable might start with a burst before the voicing onset but, to avoid responding to noise, visiBabble waits for the onset of voicing. The system responds in no more than 0.1 second of the corresponding acoustic event.

C. Data Collection

As visiBabble runs, it makes a digital recording of the session in wav format. It also saves a record of the times and types of landmarks it found during the session. A second program uses this landmark data to produce a syllable and utterance summary as shown in Table 1.

D. Experimental Formats

Single case study designs [11] are particularly suited to our preliminary tests of visiBabble since they provide the freedom to conduct a study on a small heterogeneous group of subjects. The prototype program can be run in a variety of "formats":

1) Baseline (recording, no graphic display);

2) Response (graphic display is always present, while recording);

3) A-B-A (no display, display on, no display). The length of A or B phases can be changed.

Data is collected during all phases of all formats to allow a comparison of behavior during the baseline and active phases. The analyses of landmarks and syllables are conducted and recorded separately for the B phase and two A phases.

E. Field Testing

As part of the software development, a prototype of the system, visiSyl 1.2, was beta-tested by a typically-developing one-year-old and is currently being evaluated in trials with four at-risk children, ranging in age from 28 months to 7.5 years, and three premature but typically developing infants with ages, corrected for prematurity, from 8 to 11 months. The system will be iteratively modified in response to the results of this field-testing.

Preliminary questions on the use of the visiBabble include:

1) What features of infant vocalization can the system respond to in real time?

2) What graphic feedback do infants find appealing?

3) What changes have to be made in the graphic feedback to avoid habituation?

4) Do the infants show increased babbling during the treatment (B) phases?

5) Do infants adjust the amplitude of their utterances in response to the visual reinforcement?

6) Do infants adjust the pitch of their utterances in response to visual reinforcement?

7) Do infants increase the variety of syllable types and complexity of their utterances?

8) Is there any change in the distribution of utterances as an infant matures?

9) Do parents perceive changes in their infants' vocalizations in response to the visiBabble program?

The ABA design allows direct comparisons of the child's productions (items 4 to 8) with and without the system's visual feedback. Both the rate and the variety of syllables may be tested for the stimulating effect of the system by several techniques.

III. Results

Our beta-testing with a typically developing one-year old showed that our system was responsive to a child of that age. On days when he wasn't cranky, as reported by his parents, he showed an interest in the visual response screens. These sessions were run by the child's parents in a particularly noisy environment. Noise from the heating system, a vacuum cleaner, parents talking, and the computer itself were often louder than the child and clearly affected the output. The child was also very interested in the buttons on the display.

As a result of these sessions, we now ask that the computer be placed behind the microphone and that observers, if they must speak, do so as quietly as possible and also behind the microphone. We have also placed black tape over the display buttons.

Our current tests are being run by trained speech pathology students. The system rarely responds to noise and whispering that can be heard in the background. The exception to this is when such sounds overlap with the child's utterances. The results of a sample session are shown in Table 1. Landmarks that were clearly caused by noise or adults were removed before the syllable analysis.

The subject of this session was a 6 year old male child with cerebral palsy and cortical visual impairments (but who focuses intently on book pictures and loves TV). He is a symbolic communicator with signs and word approximations, limited range of vowel and consonant sounds (about 4 consonants in repertoire).

IV. Further Development

There are several features we plan to add to the visiBabble system. We have observed that some young infants are not always interested in our visual feedback. They may not be focused on the part of the screen where the bird is flying or the frog is hopping. We will add feedback that occupies more of the screen, e.g. fireworks or large faces that wink or smile. We may add sound or tactile feedback to the responses.

Though our prototype system just responds to the detected start of syllables, it is also capable of responding to other aspects of the child's vocalizations, e.g. variation in pitch or energy, the duration of syllables or utterances, or the complexity of syllables in terms of landmark structure. We plan further tests with infants and children on these aspects of the system. We envision a system where a speech pathologist, for example, might choose to work with a child on producing longer utterances and set the visiBabble system accordingly.

For research purposes, we plan to add to the information saved by the visiBabble system. We currently save a digital audio recording of each session and the landmark analysis as it was computed in real time. From this, we are able to compute the syllables that visiBabble found and hence responded too. In future systems, we will likewise record which response was displayed so that we might determine whether certain responses are particularly effective. We will also save the pitch information as it was computed during the session. Our summary program will then be augmented to classify syllables according to pitch contours as well as landmark content.

We hope to see visiBabble become a product that is useful as a clinical and research tool for work with at-risk infants or older non-speaking children. We also intend to produce a version that can be used as a training toy for these infants and children.

Table 1: Sample Summary of Data Collected During a 10 minute A-B-A visiBabble Session

A1 - 2.5 minutes with no display; B - 5 minutes with responsive display; A2 - 2.5 minutes with no display

Syllables	**Entire Session**		**A1**		**B**		**A2**	
Type	number	average duration	number	average duration	number	average duration	number	average duration
+g-g	7	0.164			6	0.167	1	0.144
+g-s	1	0.048			1	0.048		
+s-g	1	0.120			1	0.120		
+s-s	3	0.048			3	0.048		
+b+g-s	1	0.016			1	0.016		
+g+s-g	5	0.199			5	0.199		
+g+s-s	3	0.109			3	0.109		
+g-s-g	3	0.131			2	0.165	1	0.064
+s-s-g	4	1.211			4	1.211		
+g+s-g-b	1	0.112			1	0.112		
+g+s-s-g	3	0.230			2	0.265	1	0.161
+g-s-g-b	1	0.707			1	0.707		
+s-s-g-b	1	0.273			1	0.273		
+s+s	2	3.962			2	3.962		
+g+s+s	1	3.318			1	3.318		
+g+s-s-s-g	2	0.591			1	0.490	1	0.691
+g+s-s+s-s-g	2	0.972			2	0.972		
Totals	**41**	**0.590**	**0**	**NaN**	**35**	**0.562**	**6**	**0.750**
Average Number of Landmarks per Syllable		3.049		NaN		3.029		3.167
Utterance Summary:								
	number	avdur	number	avdur	number	avdur	number	avdur
	33	0.756	0	NaN	28	0.728	5	0.911
Average Number of Syllables per Utterance:		1.242		NaN		1.250		1.200

REFERENCES

[1] C. Bickley, "Acoustic evidence for phonological development of vowels in young children," *MIT Speech Communication Working Papers **IV***, 111-124, 1984.

[2] C.J. Cress and L. Ball, "Strategies for promoting vocal development in young children relying on AAC: Three case illustrations," *Proc. Rehab. Eng. & Assist. Tech. Soc. North Am.*, RESNA Press, pp.44-46, 1998.

[3] H.J. Fell, L.J. Ferrier, D. Sneider, and Z. Mooraj, "EVA, An early vocalization analyzer: an empirical validity study of computer categorization," *Assets '96*, pp. 57-61, 1996.

[4] H.J. Fell, J. MacAuslan, L.J. Ferrier, Chenausky, "Automatic Babble Recognition for Early Detection of Speech Related Disorders," *Assets* '98, pp. , 1998.

[5] H.J. Fell, L.J. Ferrier, C. Espy-Wilson, S.G. Worst, E.A. Craft, K. Chenausky, J. MacAuslan, G. Hennessey, "Automatic Analysis of Infant Babbling in EVA, the Early Vocalization Analyzer", *ASHA Proc.*, 2000.

[6] H.J. Fell, J. MacAuslan, L.J. Ferrier, S.G. Worst, and K. Chenausky, "Vocalization Age as a Clinical Tool," *Proc. ICSLP (Int. Conf. on Speech & Language Processing)*, Denver, September 2002.

[7] T.S. Jensen, B. Boggild-Andersen, J. Schmidt, J. Ankerhus, and E. Hansen, E. "Perinatal risk factors and first-year vocalizations: Influence on preschool language and motor performance ," *Develop. Med. & Child Neur.*, **30**, pp. 153-161, 1988.

[8] K. Levin, "Babbling in infants with cerebral palsy" *Clinical Ling. and Phonetics*, **13** (4), pp. 249-267, 1999.

[9] S. Liu, "Landmark detection of distinctive feature-based speech recognition," *JASA,* **96**, 5, Part 2, p. 3227, 1994.

[10] J.L. Locke, "Babbling and early speech: Continuity and individual differences," *First Language*, **9**, pp. 191-206, 1989.

[11] L.V. McReynolds, K.P. Kearns, *Single Subject Experimental Designs in Communication Disorders*, Baltimore: University Park Press, 1983.

[12] P. Menyuk, J. Liebergott, M. Shultz, M. Chesnick, and L.J. Ferrier, "Patterns of Early Language Development in Premature and Full Term Infants," *JSHR* **34**, p. 1**,** 1991.

[13] K.N. Stevens, S. Manuel, S. Shattuck-Hufnegel, and S. Liu, "Implementation of a model for lexical access based on features," *Proc. ICSLP*, Banff, Alberta, **1**, 499-502, 1992.