VOCALIZATION ANALYSIS TOOLS

H. J. Fell[1] and J. MacAuslan [2]
[1]College of Computer and Information Science, Northeastern University, MA, USA
[2]Speech Technology and Applied Research, Bedford, MA USA

We offer two tools for automated vocalization analysis. The Syllable tool uses the Stevens landmark theory to find landmarks in vocalizations digitized as "wav" files. The landmarks are grouped to identify syllable-like productions in these vocalizations and the results are summarized. The Vocalization-Age tool is intended for pre-speech vocalizations. It uses the landmark and syllable information to yield a vocalization age that has been shown to clinically distinguish typically-developing children from children who are at-risk for later speech impairment.

I. INTRODUCTION

Many speech-related studies result in voluminous acoustic data and our projects are no exception. We have therefore developed two tools for automated vocalization analysis. One tool extracts and summarizes features from acoustic waveforms. The other tool computes a performance level of pre-speech productions. Beta-test versions of our software are now available for Matlab users.

Examples:

We have applied these tools to recordings collected in several studies.

- In the Early Vocalization Analysis (EVA) project [1], typically and atypically developing infants were recorded for 45 minutes at a time, for a total of more than 100 sessions.
- In a study of emotional stress in voice [2], we analyzed about 400 single-word, pre-existing audio recordings [SUSAS] of several subjects speaking many tokens in sometimes noisy environments.
- In the visiBabble project [3], 30 ten-minute in-home audio recordings of several children with severe speech delays were processed in real-time in a single-case-study design.
- In the UCARE project [4], 40 hours of pre-existing [5] video-taped sessions of children with physical or neurological impairments were analyzed with these tools.

Most of these recordings were made in less-than-ideal environments. Babies crawled on the floor and played with toys. Mothers, siblings, and graduate students were present and sometimes talked. The recordings also contain other environmental sounds such as air-conditioners or vacuum cleaners. Because our tools use knowledge-based speech-processing, they are robust to many of these contaminating sounds. For more subtle cases, e.g. a sibling talking nearby, the researcher can specify recording sections to ignore.

II. THE SYLLABLE TOOL

The Syllable tool uses the Stevens landmark theory [6] to find acoustically abrupt events, consonantal *landmarks*, in vocalizations digitized as "wav" files. The tool also determines voicing intensity and pitch contours.

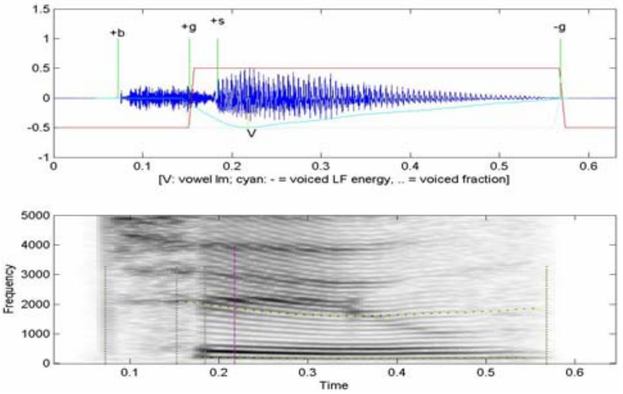


Figure 1: Syllable Analysis of "two" spoken by an adult female. The waveform (top) is shown with labeled landmarks (onset/offset of: b, bursts/frication; g, voicing; s, syllabicity) and strength of voicing; V denotes the nominal vocalic center of the syllable. The spectrogram (bottom) is shown with the pitch contour and its 10th harmonic (dashed line). Voice onset time VOT is measured by the interval between start of burst +b and onset of voicing +g.

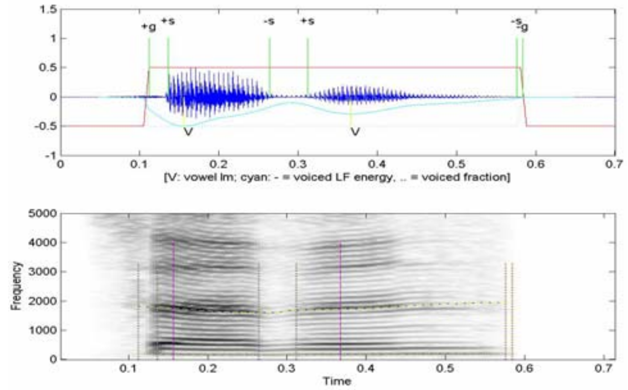


Figure 2: Syllable Analysis of "seven" spoken by an adult female. In the waveform (*top*), V denotes the nominal vocalic center of each syllable. Notice that voicing persists without a complete oral closure between the syllables. The second syllable is identified by a landmark-based rule, i.e., a +s that is not closely preceded by a +g.

The landmarks are grouped to identify syllable-like productions in these vocalizations and the results are summarized.

III. The Vocalization Age Tool

The Vocalization Age (or vocAge) tool is specifically intended for pre-speech vocalizations. The digitized recording of a single session with a child is first analyzed by the syllable tool. The resulting information is then summarized and compared against data collected in ~100 sessions with typically-developing infants ranging in age from six to 15 months. The tool thus derives a "vocalization age".

In an application of the Vocalization Age, we found two specific screening rules [7] 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 15 month age range:

- An infant is (or is not) in the atypical group according as any session (respectively, no session) shows a delay of at least 3.1 months.
- An infant is (is not) in the atypical group according as any (respectively, no) two consecutive sessions both show delays of at least 2.3 months.

IV. How the Tools Work

A. Landmarks

Landmark processing begins by analyzing the signal into several broad frequency bands. Because of the different vocal-tract dimensions, the appropriate frequencies for the bands are different for adults and infants; however, the procedure itself does not vary. First, an energy waveform is constructed in each of the bands. Then the rate of rise (or fall) of the energy is computed, and peaks in the rate are detected. These peaks therefore represent times of abrupt spectral change in the bands. Simultaneous peaks in several bands identify consonantal landmarks.

B. Syllables and Utterances

The program identifies sequences of landmarks, e.g., +g-g or +s-g-b, as syllables based on the landmark order and inter-landmark timing. Among other constraints, syllables must contain a voiced segment of sufficient length. Figure 4 shows an example of this rule.

An utterance is a sequence of syllables in which gaps between syllables are no more than (nominally) 200 milliseconds long.

Both syllables and utterances may have properties of their own, such as a pitch template (rise/fall/rise) or a peak zero-crossing rate.

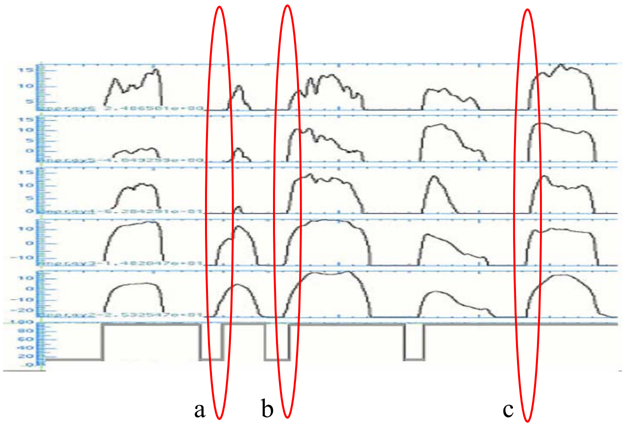


Figure 3: Initial spectral analysis of an infant utterance: voicing (*bottom*) and five frequency bands' energy waveforms. Landmarks are identified by large, abrupt energy increases or decreases that are simultaneous in several bands. (*a*) Too few bands show large, simultaneous changes in energy. (*b*) All bands show large, simultaneous energy increases immediately before the onset of voicing, identifying a +b (burst) landmark. (*c*) All bands show large, simultaneous energy increases during ongoing voicing, identifying a +s (syllabic) landmark.

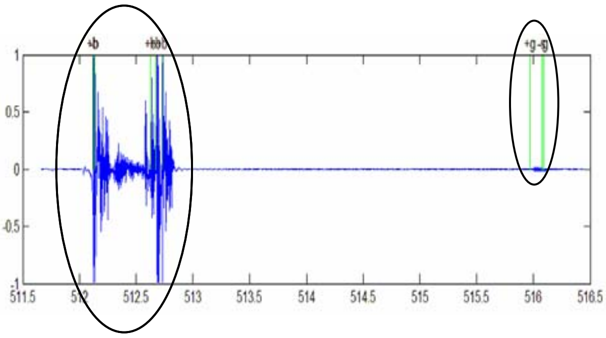


Figure 4: Ignored noise vs. recognized syllable. (*Left segment*) Noise marked by only +b and -b landmarks; (*right segment*) a faint babble marked by +g-s-g. Because any syllable must contain a voiced segment, the loud, noise segment is automatically ignored in subsequent processing. The babble, in contrast, has well defined voicing and sufficient duration and is hence retained.

C. Vocalization Age

There are many syllable and utterance measurements that the tool uses in forming the vocalization age:

- Number of syllables per utterance
- Number of occurrences and mean duration for each syllable type
- Number of syllables starting with a given onset landmark: +g, +s, etc.
- Number of syllables ending with a given offset landmark: -b, -g, etc.
- Number of syllables with n landmarks, n = 2 to 7.
- Standard deviations of related quantities, when they apply.

The Syllable tool can be set to extract and summarize exactly those measurements that are needed to compute the vocalization age.

V. Using the Tools

We have applied these tools to recordings collected in several studies and we hope that other researchers will use the tools on their own data. Also, as we improve the feature collection capabilities of our tools, we, and perhaps others, will want to use them repeatedly on previously collected recordings. These tools can be run on all the recordings for a single "wav" file, a complete session, all the sessions of a subject, or an entire study with a single invocation. (See Figure 5.)

A. System Requirements

Currently, our software requires Matlab and the Matlab Signal Processing Toolbox. We run it under the Windows XP operating system.

B. Preparing Data

All sound files must be in "wav" format. Because much of our own data was collected using Entropic's ESPS-WAVES and saved in "sd" files, we also provide software to convert Entropic "sd" files to "wav" format. A user can run the Syllable and Vocalization Age tools on a single recording or on a directory tree containing recordings of (see Figures 5 and 6):

- a single session with a subject,
- all sessions of a subject,
- an entire study.

A simple text file may be included for any recording to indicate sections that should not be analyzed.

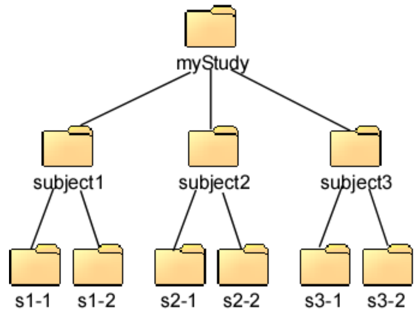


Figure 5: Arrangement of session data in a study. This figure represents a study with two sessions for each of three subjects.

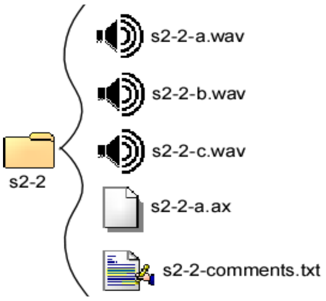


Figure 6: A session directory. This figure shows a folder containing three audio files from a single session, a text ("ax") file that marks segments to be ignored in one of the "wav" files, and a text file with the researcher's comments about the session.

C. Single-Subject Experiments

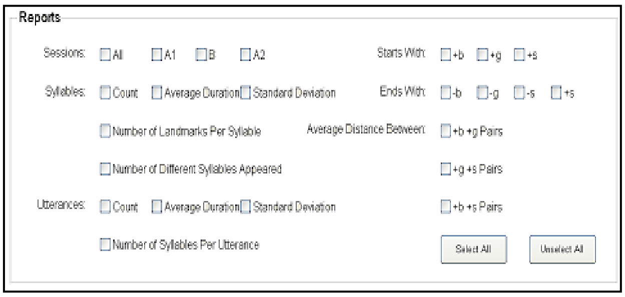
Single case study designs [8] are particularly suited to studies on a small heterogeneous group of subjects. For example, in our preliminary tests of visiBabble, a real-time visual-feedback system, we ran sessions in a variety of formats:

1) Baseline (recording, no graphic display);
2) Response (graphic display is always present, while recording);
3) A-B-A (display off-on-off).

Data was collected during all phases of all formats to allow a comparison of behavior during the baseline and feedback phases. The Syllable tool can analyze the landmarks and syllables for the A and B phases combined or separately.

D. Reports

A dialog box allows the user to select particular features to be summarized by the Syllable tool (see Figure 7). The tool will then generate, in the data directory, a tab-delimited report that can be easily copied into a spreadsheet for future analysis (see Table 1).


Reports

Sessions: ☐ All ☐ A1 ☐ B ☐ A2
Starts With: ☐ +b ☐ +g ☐ +s
Syllables: ☐ Count ☐ Average Duration ☐ Standard Deviation
Ends With: ☐ -b ☐ -g ☐ -s ☐ +s
☐ Number of Landmarks Per Syllable
Average Distance Between: ☐ +b +g Pairs
☐ Number of Different Syllables Appeared
☐ +g +s Pairs
Utterances: ☐ Count ☐ Average Duration ☐ Standard Deviation
☐ +b +s Pairs
☐ Number of Syllables Per Utterance

Figure 7: Dialog box for selecting report features.

VI. Further Development

We anticipate adding capabilities over the next two years while our visiBabble project is under development. We encourage other researchers to try a beta-test version and to suggest enhancements that would be useful to them.

Acknowledgements

The authors appreciate the encouragement and testing by Prof. Cynthia Cress, Univ. of Nebraska-Lincoln. We also thank Jun Gong for his programming contributions. This work was funded in part by National Institutes of Health STTR grant R42 DC005534.

Table 1: Summary of a Short Sample Session

The report summarizes the landmark, syllable, and utterance statistics of a sample, 30-second session. These statistics are among those used in the vocAge tool.

Total			
SyllableType	**Count**	**Mean**	**StdDev**
+g-g	3	704.000	532.626
+s-g	1	408.000	0.000
+g-g-b	3	96.000	36.368
+g+s-g	1	48.000	0.000
+b+g-g-b	1	344.000	0.000
+g+s-g-b	1	64.000	0.000
+g+s-s-g	1	160.000	0.000
+g+s+s	1	696.000	0.000
Total	12	343.333	382.559
2 lm/syl	4		
3 lm/syl	5		
4 lm/syl	3		
5 lm/syl	0		
6 lm/syl	0		
7 lm/syl	0		
Avg	2.917		
DiffSyl	8		
Utts	**Count**	**AvDur**	**StdDev**
	8	531.000	486.982

mean duration of +g-g syllables

mean duration of all syllables

number of syllables with 2 landmarks

average number of landmarks/syllable

number of syllable types occurring

total number of utterances

References

[1] H.J. Fell, J. MacAuslan, L.J. Ferrier, K. Chenausky, "Automatic Babble Recognition for Early Detection of Speech Related Disorders," *Journal of Behaviour and Information Technology*, 1999, **18**, no. 1, 56-63.

[2] H.J. Fell, J. MacAuslan, "Automatic Detection of Stress in Speech," *Proceedings of MAVEBA 2003*, Florence, Italy, pp. 9-12.

[3] H.J. Fell, J. MacAuslan , C. J. Cress , L. J. Ferrier, "Using Early Vocalization Analysis for visual feedback," *Proceedings of MAVEBA 2003*, Florence, Italy.

[3] H.J. Fell, J. MacAuslan, C. Cress, L.J. Ferrier, "visiBabble for Reinforcement of Early Vocalization," *Proceedings of ASSETS 2004*, Atlanta, GA., pp. 161-168.

[4] C.J. Cress, S. Unrein, A. Weber, S. Krings, H. Fell, J. MacAuslan, J. Gong, "Vocal Development Patterns in Children at Risk for Being Nonspeaking," submitted to *ASHA 2005.*

[5] C.J. Cress, *Communicative and symbolic precursors of AAC*. Unpublished NIH CIDA Grant: University of Nebraska-Lincoln, 1995.

[6] K.N. Stevens, S. Manuel, S. Shattuck-Hufnegel, and S. Liu, “Implementation of a model for lexical access based on features,” *Proc. ICSLP (Int. Conf. on Speech & Language Processing)*, Banff, Alberta, **1**, 499-502, 1992.

[7] 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.

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