**Abstract:** 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 fundamental frequency (F0) contours.

![Figure 1: Syllable Analysis of "two" spoken by an adult female.](image1)

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. 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.](image2)

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 syllable onset (+s) that is not closely preceded by a voicing onset (+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 six to 15 month-old, typically-developing infants. The tool thus derives a "vocalization age".

In an application of the vocAge, we found two specific screening rules [7] that clinically distinguish infants who may be at risk for later communication or other developmental problems from typically developing infants:
- An infant is (or is not) in the atypical group according as any session (respectively, no session) shows a "delay", i.e., difference between chronological and vocalization age, 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 (widths of 400-2000 Hz) selected to detect important speech features such as the second formant. 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, i.e., presence of strong harmonic content, (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).

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.

<table>
<thead>
<tr>
<th>SyllableType</th>
<th>Count</th>
<th>Mean</th>
<th>StdDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>+g-g</td>
<td>3</td>
<td>704.000</td>
<td>532.626</td>
</tr>
<tr>
<td>+s-g</td>
<td>1</td>
<td>408.000</td>
<td>0.000</td>
</tr>
<tr>
<td>+g-g-b</td>
<td>3</td>
<td>96.000</td>
<td>36.368</td>
</tr>
<tr>
<td>+g+s-g</td>
<td>1</td>
<td>48.000</td>
<td>0.000</td>
</tr>
<tr>
<td>+b+g-g-b</td>
<td>1</td>
<td>344.000</td>
<td>0.000</td>
</tr>
<tr>
<td>+g+s-g-b</td>
<td>1</td>
<td>64.000</td>
<td>0.000</td>
</tr>
<tr>
<td>+g+s-s-g</td>
<td>1</td>
<td>160.000</td>
<td>0.000</td>
</tr>
<tr>
<td>+g+s+s</td>
<td>1</td>
<td>696.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>343.333</td>
<td>382.559</td>
</tr>
<tr>
<td>2 lm/syl</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 lm/syl</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 lm/syl</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 lm/syl</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 lm/syl</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 lm/syl</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg</td>
<td>2.917</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DiffSyl</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Utts</td>
<td>Count</td>
<td>AvDur</td>
<td>StDev</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>541.000</td>
<td>386.982</td>
</tr>
</tbody>
</table>

REFERENCES