THE DESIGN OF A MICROCOMPUTER-CONTROLLED VOICE ONSET TIME ANALYZER

SEYED M. R. MOEINI, JAMES E. FLEGE, PhD,
MARTIN J. McCUTCHEON, PhD

Reprinted from BIOMEDICAL INSTRUMENTATION & TECHNOLOGY
Published by HANLEY & BELFUS, Inc., Philadelphia, PA
The Design of a Microcomputer-Controlled Voice Onset Time Analyzer

SEYED M. R. MOEINI, JAMES E. FLEGE, PhD, MARTIN J. McCUTCHEON, PhD

In spoken syllables such as “ta,” the interval between the release of the tongue constriction for the stop consonant /t/ and the onset of the vowel is called voice onset time, or VOT. Voice onset time is an important determinant of whether the initial consonant will be heard as a /t/ (values of 60–90 ms) or as a /d/ (values of 0–30 ms). VOT information, immediately following a spoken syllable, can provide a speaker with feedback for modifying speech production. Such information can help the hearing-impaired learn to speak. It may also help people who learn English as a second language, since they often produce /b,d,g/ and /p,t,k/ with inappropriate VOT values. A prototype portable device measures VOT for initial voiced and voiceless stop consonants (e.g., “da” and “ta”). A dual-microphone method is used for acoustic measurement. A microphone in front of the mouth picks up the radiated acoustic signal; another over the larynx transduces vocal vibrations that mark the beginning of a vowel. Analog circuits process the transducer signals and provide gain and filtering. Filters were designed on the basis of the acoustic properties of stop consonants. The output from each analog circuit is fed to a comparator that compares the signal level with a fixed threshold voltage reference. A digital timer starts when the amplitude of the oral signal voltage exceeds a threshold and stops when the laryngeal signal voltage exceeds a threshold. VOT values obtained by the device were compared with those made from digital waveforms of words spoken by five talkers. The average difference between the two sets of measurements was 3.5 ms (n = 255).

Software interrupts of the microprocessor timer during counting and differences between hardware processor timers may account for some of the differences. Regression analysis indicated the VOT analyzer to be a valid measurement technique for VOT. The offset in timing (intercept = 2.4 ms) was due to the method used to detect vowel onset. (BIOMEDICAL INSTRUMENTATION & TECHNOLOGY 1990;24:357–362)

Stop consonants are produced through the buildup of air pressure in the oral cavity behind a constriction, and the sudden release of the constriction. For example, the alveolar stop /t/ and its voiced counterpart /d/ are produced by moving the tip of the tongue forward and up to contact the hard palate. When the constriction is released, an audible transient (release burst) can be heard. The release burst is a wide-spectrum noise whose frequency depends upon, among other things, the place of the constriction in the vocal tract. For the voiced (unaspirated) stops in English (/b,d,g/), voicing may continue without interruption through the period of constriction. The aspirated stops in English (/p,t,k/) are quite different, for there is no voicing during the constriction interval, that is, before the stop release. Nor are vocal cords adjusted to a configuration appropriate for vibration at the time of release. As the constriction is being released, the vocal cords are being abducted. As they come together, the size of the glottis decreases and noise is generated at the glottis. This noise source in turn excites the vocal tract and creates an aspirated sound. When the vocal folds are sufficiently closed, the aspiration noise ceases and voicing for the following vowel begins.

The relative timing of glottal and supraglottal events distinguishes “voiced” and “voiceless” stops such as /d/ and /t/. Thus, due to differences in the timing of vocal fold adduction, voicing for the following vowel begins approximately 50–100 ms after the release burst for /t/, but only about 10–20 ms after the release burst for /d/. For the voiceless aspirated stops, in English, the release interval is further divided into frication and aspiration intervals. This relative timing of stop release and onset of voicing is called voice onset time (VOT) by Lisker and Abramson.
PURPOSE

The purpose of this project was to develop a prototype device for measuring the VOTs of voiced (/b,d,g/) and voiceless (/p,t,k/) stop consonants in syllable-initial position. The VOT values measured by the device were compared with values measured directly from digital waveforms, as shown in Figure 1. This is the most commonly used method for measuring VOT, but it requires expensive analog and digital computer equipment, and cannot be used to provide near-instantaneous measurement of VOT.

Voice onset time feedback presented as a visual speech display may be useful for training deaf children to produce the voiced–voiceless distinction for initial stop consonants. Such a distinction is usually very difficult for hearing-impaired children and adults to acquire. Very few visual cues that distinguish between the voiced and voiceless consonants are available to the deaf child. Hearing-impaired children make use of visual cues along with auditory stimulation, but it is difficult for them to learn to produce distinctions that involve articulatory gestures that may not be readily perceived. The stop consonants are extremely difficult to lip-read due to the lack of visual cues. Consequently, problems of vocalization occur when both oral and laryngeal gestures are involved. This frequently causes substitution and omission errors as deaf individuals attempt to pronounce /b,d,g/ and /p,t,k/.7

Voice onset time feedback may also prove useful in the remediation of speech disorders associated with verbal apraxia. These speakers have an impairment in the auditory–cortical–vocal loop. Even if auditory perception of the stop consonant in a word occurs, the information may be processed inappropriately at higher levels.

The VOT analyzer may be a great help to patients suffering from the above-described impairments. Such a measurement of VOT would also be useful in helping non-native English speakers overcome difficulties in producing English stop consonants. For example, the presentation of rapid and accurate VOT feedback could be used to help native speakers of Spanish to produce English /t/ sounds with VOT values of about 70–80 ms, rather than Spanish-like values of 10–20 ms. If such feedback is given immediately after each syllable, the speaker may be able to correct pronunciation errors that he or she may not hear. It may provide the opportunity for the learners to correct inappropriate vocal gestures before they become habitual.

No device is currently available commercially that can be used to provide an accurate and rapid VOT feedback with digital readout. Two laboratory methods are now commonly used for measuring VOT. One is to use a sound spectrograph to display and analyze a signal in terms of frequency, duration, and intensity of components. The other, as mentioned above, is to measure VOT directly from acoustic waveforms (Figure 1).

MATERIALS AND METHODS

The VOT analyzer developed used two near-field microphones. One was positioned 2 to 3 cm from the mouth to pick up the acoustic signal radiated from the mouth. The other was attached by means of a strap to the throat to transduce the tissue-conducted vibrations from the vocal folds (Figure 2). Both microphones (Shure model SM12) were unidirectional, with a -3-dB frequency response of 50–15,000 Hz.
Each microphone output signal was amplified, filtered, and passed to an analog comparator. Each signal was compared continuously with a fixed threshold voltage. When the amplitude of the mouth signal crossed a predetermined threshold, its comparator output sent a logic level signal to a microcomputer (MOSTEK, MK38P73) that started a software counter. When the laryngeal signal amplitude reached its threshold voltage, a second comparator sent a signal that stopped the timer. The measured timing was then converted to appropriate units for an LED display, as shown in Figure 3.

These circuits were designed to reflect the temporal and spectral properties of stop consonants as determined by previous researchers.\textsuperscript{2,12,13} and by our own measurements. The microphones provided signals in the millivolt range, and the preamplifiers provided a voltage gain of 100. A total gain of 2,000–16,000 was available for the oral signal, and 500–5,000, for the laryngeal signal.

The filter for the oral signal had a 6-kHz bandwidth whereas the filter for the laryngeal had a 3.6-kHz bandwidth. The final amplifier stage provided variable gain with eight different, software-selectable levels.

To allow the device to accommodate variations between subjects, stops, and vowels, a two-step operation was provided. The first step in operation is a gain-adjustment process that is carried out as the subject says a word such as “table” repetitively. The oral signal threshold is fixed at 2.5 volts and the laryngeal threshold is set to 2.0 volts. The amplifier gains for the signals are then adjusted by an interactive software program. The lowest gain is set initially. With each subsequent production, the amplitude for the vowel (e.g., /a/) is compared with the corresponding threshold value. When the amplitude reaches the reference value, the comparator output sends an active high transistor–transistor logic (TTL) signal to activate a D-type flip-flop. The output of the flip-flop invokes the external interrupt line on the microcomputer; acknowledgment of it by software fixes the gain selected for the last speech token. Otherwise, the software sets the next higher gain for another try. The gain adjustment process is performed for each source signal (oral and laryngeal) separately.

**Figure 2.** The voice onset time analyzer.

**Figure 3.** Block diagram of voice onset time analyzer circuit design.
The VOT analyzer was tested by examining nine consonant–vowel–consonant words produced by five adult native speakers of American English (3 male, 2 female). The aim of the test was to determine whether VOT measurements made by the device were accurate. To do so, the measurements made by the VOT analyzer were directly compared with measurements made from digital waveforms. The gains were adjusted separately for three places of articulation and each following vowel. Testing was conducted first for the labial stop (/p/), then for the alveolar stop (/t/), and finally for the velar stop (/k/). For the VOT measurements, each subject said each word five times by reading from a randomized list of 45 words. On each trial, the experimenter said the number of the word (1–45), pushed the enable button on the VOT analyzer, and recorded the displayed VOT value. The 45 words spoken by each subject were also tape-recorded simultaneously (Technics model M235X) for later analysis with a waveform editor (AUDED).14

For waveform analysis, the recorded signals were low-pass filtered (4 kHz) and digitized (at 10 kHz, 12-bit resolution). Figure 1 illustrates the segmentation procedures used to measure VOT with the editor. Cursors were placed at the beginning of the release burst and at the first upward-
going zero crossing in the periodic portion of the word (i.e., vowel onset). The experimenter did not have access to measurements obtained previously with the VOT analyzer when making the waveform measurements.

Previous research\textsuperscript{4,15,16} has shown that the VOT values associated with English, /p/, /t/, and /k/, range from about 30 ms to 120 ms. A few VOT values returned by the VOT analyzer fell outside this range. Several experiments were done with different subjects and words to investigate this problem prior to formal testing. It was observed that at the time of the release burst, a component of the burst was picked up by the laryngeal microphone, due to crosstalk between the two channels. Two productions of the word “cot” were measured by the oral and laryngeal microphones. Onset of the release burst, which should have been seen only in the oral microphone signal, could also be seen in the laryngeal signal. This caused the “vowel” onset to be detected prematurely. A few spurious VOT values were generated by this phenomenon.

The average VOT values obtained from the digital waveforms and the VOT analyzer are presented in Figure 4. The average values obtained using the new VOT analyzer and those obtained spectrographically by Lisker and Abramson\textsuperscript{3} differed little for /p/ (62 ms vs. 58 ms), for /t/ (74 ms vs. 70 ms), and for /k/ (80 ms vs. 80 ms). The results were evaluated statistically in the following way. The average difference between the two sets of measurements was found to be 3.5 ms. The 450 values (225 obtained using each measurement technique) were submitted to a separate, paired t-test for each place of articulation and following vowel. If the mean values obtained using the two methods did not differ significantly at the 0.05 level, the VOT analyzer would be judged to be a valid measurement technique for VOT. The results of the t-test (Table 1) show that the most significant differences were for the stop consonant and following vowel /ka/ (p = 0.005) and /ku/ (p = 0.002). Another significant difference was for syllable /ta/, at the level p = 0.02.

As shown in Figure 5, a linear relationship was found between the VOT measurements obtained from the VOT analyzer and from waveforms (oscillogram). The slope of the regression line was 0.977 and the intercept was 2.426 ms. The correlation coefficient (r) was 0.9174.

Table 1. The Results of the t-tests Obtained for Each Place of Articulation and Following Vowel

<table>
<thead>
<tr>
<th>Mean VOT* Value (ms)</th>
<th>Oscillograph</th>
<th>Analyzer</th>
<th>t (1,24)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p/</td>
<td>52.4</td>
<td>51.6</td>
<td>0.82</td>
<td>0.419</td>
</tr>
<tr>
<td>/p/</td>
<td>66.3</td>
<td>66.7</td>
<td>-0.12</td>
<td>0.91</td>
</tr>
<tr>
<td>/pu/</td>
<td>65.3</td>
<td>63.2</td>
<td>1.14</td>
<td>0.27</td>
</tr>
<tr>
<td>/ti/</td>
<td>67.4</td>
<td>68.5</td>
<td>-1.27</td>
<td>0.22</td>
</tr>
<tr>
<td>/ta/</td>
<td>71.7</td>
<td>75.8</td>
<td>-2.45</td>
<td>0.02</td>
</tr>
<tr>
<td>/tu/</td>
<td>82.5</td>
<td>78.4</td>
<td>1.97</td>
<td>0.06</td>
</tr>
<tr>
<td>/ki/</td>
<td>85.2</td>
<td>87</td>
<td>-0.45</td>
<td>0.66</td>
</tr>
<tr>
<td>/ka/</td>
<td>82.4</td>
<td>54.2</td>
<td>3.14</td>
<td>0.005</td>
</tr>
<tr>
<td>/ku/</td>
<td>83.4</td>
<td>79</td>
<td>3.5</td>
<td>0.002</td>
</tr>
</tbody>
</table>

*Voice onset time.
DISCUSSION

The VOT analyzer described here has several unique features, the most important being its ability to save substantial amounts of time in making VOT measurements. Its portability makes it easy to conduct field experiments. Because it is a single-function instrument, it is simpler than other measurement techniques for this purpose. The relatively low cost is also attractive.

The result of the regression analysis (Fig. 5) indicates that the VOT analyzer provides a valid measurement technique. This figure illustrates some offset in timing, which is due to the different methods of measurement (intercept). In the laboratory waveform technique, the onset of the vowel is determined by placing the cursor at the first upward-going zero crossing in the periodic portion of the word. However, the VOT analyzer detects vowel onset at the first vowel peak. Some part of the discrepancy may be due to unavoidable interrupts of the software timer during timer counts. Another source of error may be the difference between the hardware processor timers in the two techniques. This difference may affect the result in both the offset and the slope of the regression line. The difference in measurement timings between the analyzer and laboratory waveform analysis can be used as a correction factor for the VOT measurements. Further improvements in the device could be made to eliminate the occasional spurious VOT values (about 5%) by using a laryngeal transducer with less crosstalk.

It would be difficult to accurately test the reproducibility of the VOT measurements with spoken stop consonants. Humans seldom, if ever, produce a particular speech sound identically across trials. Thus, if VOT were to be measured at two separate times, differences could arise owing to differences in how the stop consonants were spoken. The most reasonable method for estimating measurement validity was deemed to be the comparison of measurements made using the VOT analyzer with a second set of measurements made from a tape recording of the same speech material using a conventional technique.

In summary, a new device for measuring VOT in syllable-initial stop consonants has been described. Satisfactory measurement validity was shown by comparing the VOT values obtained using the new device with those from measurements made by hand from digital waveforms. The new device yielded VOT values that were generally quite close to those obtained using the more traditional measurement technique. Further work to optimize the prototype device described here is needed. One goal is to eliminate the need to calibrate the device for each type of syllable to be examined (e.g., “tea” versus “pie”). Another goal is to reduce crosstalk between the signals generated at the mouth and the larynx, which sometimes results in measurement artifacts. With such improvements, the device promises to provide a useful technique for the remediation of the speech of deaf individuals and stroke victims, and for training in the production of stop consonants in a foreign language.

REFERENCES