The Effect of Tonal Changes on Voice Onset Time in Mandarin Esophageal Speech

*Hanjun Liu, †Manwa L. Ng, *Mingxi Wan, *Supin Wang, and *Yi Zhang

*Xi’an, P.R. China and †Hong Kong SAR, China

Summary: The present study investigated the effect of tonal changes on voice onset time (VOT) between normal laryngeal (NL) and superior esophageal (SE) speakers of Mandarin Chinese. VOT values were measured from the syllables /pʰa/, /tʰa/, and /kʰa/ produced at four tone levels by eight NL and seven SE speakers who were native speakers of Mandarin. Results indicated that Mandarin tones were associated with significantly different VOT values for NL speakers, in which high-falling tone was associated with significantly shorter VOT values than mid-rising tone and falling-rising tone. Regarding speaker group, SE speakers showed significantly shorter VOT values than NL speakers across all tone levels. This may be related to their use of pharyngoesophageal (PE) segment as another sound source. SE speakers appear to take a shorter time to start PE segment vibration compared to NL speakers using the vocal folds for vibration.

Key Words: Voice onset time—Esophageal speech—Tone—Mandarin.

INTRODUCTION

Voice onset time (VOT) is defined as the time interval between the release of a stop and the onset of the following vowel. It serves as an important perceptual cue to distinguish voicing and aspiration. VOT has been reported to be the primary determinant between voiceless and voiced stops. Voiceless stops are associated with longer VOT values than voiced stops; aspirated stops have longer VOT values than their unaspirated counterparts.

In normal laryngeal (NL) phonation, the absolute and relative VOT differences between cognate pairs vary depending on a number of factors. Lisker and Abramson indicated that there are smaller VOT values and less VOT difference between cognates in sentences compared to single syllables. Stress pattern also affects VOT; longer VOT values were seen in stressed voiceless stops than unstressed counterparts, although the effects of stress on VOT are limited. Similar results were reported by Löfqvist. In addition, VOT varies as a function of speaking rate, in which longer VOT values are associated with slower speaking rate resulting in longer word duration.

After the surgical procedure of removing a pathological larynx, laryngectomized patients using superior esophageal (SE) phonation have shown
difficulty producing correct voicing contrasts between homorganic stop consonants. It is generally believed that, with the use of the pharyngoesophageal (PE) segment as a new sound source, SE speakers are experiencing difficulty with the precise control of VOT. A large number of studies have explored the VOT characteristics of SE speech in comparison with NL speech. It is generally agreed that SE speakers tend to produce stops with smaller VOT values than NL phonation. This is especially true for voiceless stops. Errors in voicing distinction were reported in SE phonation. Sacco et al reported that approximately half of all listener errors involved misidentification of voicing features based on the VOT perceived. Christensen et al investigated the VOT produced by SE and NL speakers in English. Results revealed that VOT values associated with voiceless stops produced by SE speakers were significantly shorter than those produced by NL speakers, while no significant differences were found in VOT values associated with the voiced stops between the two speaker groups. Christensen et al suggested that the shorter VOT values for voiceless stops by SE speakers are due largely to the normally adducted PE segment. An additional explanation is offered that air volume in the cavity between the vibratory source and lips in SE speakers is relatively smaller after total laryngectomy, so that it takes less time than NL speakers for sufficient air pressure drop following stop release to allow for initiation of PE segment vibration.

Connor et al compared high- and low-intelligibility productions of /t/ and /d/ in SE speakers, and acoustic comparison indicated a longer VOT for /t/ than /d/ for high-intelligibility productions, but not for low-intelligibility productions. This revealed that SE speakers used VOT to differentiate /t/ and /d/ in high-intelligibility productions. Shorter VOT values for voiceless phonemes produced by SE speakers compared to NL speakers were also found by Robbins et al. They also reported that, for voiceless stops, VOT differences between SE and NL speakers were vowel dependent. VOT values associated with /p/ and /k/ in a C/i/C syllable did not differ between SE and NL speakers but SE speakers had significantly shorter VOT values associated with /p/ and /k/ in a stimulus with the vowel /a/ or /u/. To analyze the articulatory ability of skilled SE speakers in terms of the voicing distinction in consonants, Hirose conducted the perceptual and acoustical studies to examine the pattern of confusion made for normal listeners in comprehending consonants spoken by SE speakers. The results indicated that VOT values in word-initial position and the presence or absence of voicing in word-medial position were the most important cues for voicing distinction of consonants in SE speech.

Although there are many studies reporting the VOT characteristics of SE speech, most have reported data from English SE speakers. A review of the literature indicates that few investigated SE speech of a tone language. In studying Thai SE speech, Gandour et al measured the VOT values of word-initial stops from Thai-speaking SE and electrolarynx speakers. Their data showed that SE speakers were capable of signaling voicing distinctions for /b, d/ and /p, t, k/, but they were unsuccessful in signaling /pʰ, tʰ, kʰ/. They then concluded that the realization of stop voicing contrasts in alaryngeal speech depends on the number of voicing categories in a language, the relative positions of the voicing categories on the VOT continuum, and the form of alaryngeal speech. As a tone language, Mandarin has four contrastive lexical tones: high-level (HL), mid-rising (MR), falling-rising (FR), and high-falling (HF). A syllable such as /t ha/ produced at different tone levels will result in different meanings. Different tones are generally associated with different fundamental frequencies (F₀) and pitch levels. Physiologically, these are determined largely by the tension of the vibrating structure, and thus the rate of vibration. To prepare different tensions in the vibratory structure, different amounts of time may be needed. As a result, vibration may start at different times, and thus VOT may vary. Although many studies have reported the VOT characteristics of different languages, the present study germinated from two existing issues. First of all, few studies reported VOT information of a tone language, especially that of Mandarin. Iwata and Hirose reported VOT values of both stops and affricates of Mandarin. They found that VOT of the aspirated consonants was longer than that of the unaspirated
counterparts, and VOT of affricates was longer than that of stops. Yet, only two subjects who were non-native speakers of Mandarin participated in that study. The lack of statistically generalizable results due to the insufficient subject sample reduces the representativeness of the study. Liu et al reported differences in VOT values between aspirated and unaspirated stops of Mandarin produced by normal subjects and individuals with cerebral palsy. However, the interrelationship between VOT values and other factors such as tone levels is still unknown. The second issue is the lack of sufficient information on how VOT is affected by using an alternative voicing source, as in the case of esophageal speakers using the PE segment as the new voicing source. There is paucity of information on the VOT characteristics exhibited by the SE speakers of a tone language. Using the PE segment as an alternative voicing source, the onset of phonation may be different from using a normal larynx. Consequently, VOT may be affected.

Therefore, the present study aimed at investigating two aspects of VOT characteristics: (1) determine the influence of different tone levels of Mandarin Chinese on VOT and (2) determine how modes of phonation (NL and SE) influence VOT, and determine if Mandarin SE speakers are able to linguistically distinguish and maintain similar VOT patterns as in NL speakers.

METHOD

Participants

Two groups of adult males (eight NL and seven SE) who were native speakers of Mandarin participated in the present study (Table 1). All speakers were physically healthy with no known history of speech and hearing problems, except that related to laryngectomy for alaryngeal participants. The participants were all natives of Beijing and reportedly spoke the Beijing dialectal variation of Mandarin. All alaryngeal speakers received total laryngectomy as a result of late stage laryngeal cancer. At the time of experiment, they had used esophageal speech as their primary means of verbal communication for an average of 8.93 years. Only those judged as superior speakers by the ENT doctors in charge of their speech rehabilitation were recruited in the present study. Participants were all age-matched and average age (SD) was 52.50 (8.00) years for NL speakers and 57.86 (7.38) for SE speakers. All speakers were literate and were able to read the speech materials used in the study.

Speech materials

To study the VOT differences in Mandarin NL and SE speech, the voiceless stops /pʰ, tʰ, kʰ/ followed by a vowel /a/ were used. The CV syllable (stop + /a/) was produced at four tone levels of Mandarin: HL, MR, FR, and HF. Table 2 depicts the stimulus words used in the study at four contrastive tones, in which these words were combined with real words and nonsense words. To ensure a more natural production, all stimulus words were embedded in a carrier phrase (meaning “I read _____ to you”) during the recording.

<table>
<thead>
<tr>
<th>TABLE 1. Brief Information of the Laryngectomized Subjects (S1 to S7) and the Normal Subjects (S8 to S15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>S1</td>
</tr>
<tr>
<td>S2</td>
</tr>
<tr>
<td>S3</td>
</tr>
<tr>
<td>S4</td>
</tr>
<tr>
<td>S5</td>
</tr>
<tr>
<td>S6</td>
</tr>
<tr>
<td>S7</td>
</tr>
<tr>
<td>S8</td>
</tr>
<tr>
<td>S9</td>
</tr>
<tr>
<td>S10</td>
</tr>
<tr>
<td>S11</td>
</tr>
<tr>
<td>S12</td>
</tr>
<tr>
<td>S13</td>
</tr>
<tr>
<td>S14</td>
</tr>
<tr>
<td>S15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE 2. Stimulus Words Used as Speech Materials</th>
</tr>
</thead>
<tbody>
<tr>
<td>/pʰ/a/</td>
</tr>
<tr>
<td>Prostrate</td>
</tr>
<tr>
<td>Crawl</td>
</tr>
<tr>
<td>Nonsense</td>
</tr>
<tr>
<td>Scare</td>
</tr>
</tbody>
</table>

All of the esophageal speakers were using injection method for air intake.
Recording procedures
Before the actual recording, each speaker was given a brief practice period to become familiar with the speech materials, recording format, and recording instrumentation. This helped to ensure that the speech samples represented the participant’s best production. Each speaker was instructed to read the speech materials as though they were conversing with a person at a distance of approximately one meter. All recordings took place in a soundproof room. Speech signals were collected by using a cardioid microphone (Model RE10, Electrovoice, Lincoln, NE) that was mounted at a distance of 15 cm from the mouth. This placement was chosen to minimize the recording of potential stoma noise in SE speakers. The recorded speech samples were digitized at a sampling frequency of 22 kHz at 16 bits/sample and amplified by using a multichannel conditioning amplifier (Brüel and Kjær, Denmark).

During the recording, the subjects were provided with cards on which Chinese characters representing the stimulus words were printed. The order at which the cards were presented was randomized to avoid the ordering effect. Instructions were given to the speakers before the recordings. They were instructed to read the speech materials two times at a normal loudness and speaking rate.

Measurements
Temporal measurements were obtained from the digitized signals of the CV syllables by using Praat, a signal analysis software package. VOT was determined as the time interval between the release of the stop occlusion and the onset of the following vowel. Such measurements were obtained based on a waveform and a wideband spectrogram (window size = 5 milliseconds). The use of a waveform with the help of a wideband spectrogram in the measurement of VOT values was found to be reliable than measuring VOT based on the time-domain waveform only. When measuring VOT, two vertical markers were placed on the waveform. The first marker was used to indicate the release of burst, and the second marker was placed at the first identifiable period at the level of the second formant. VOT values (in milliseconds) were then measured as the time interval between the two markers.

To ensure the consistent and reliable measurement of VOT, interjudge and intrajudge reliability were assessed. To determine the intrajudge reliability, 10% of the entire data corpus (36 out of 360 speech samples) were randomly selected and reanalyzed by the same investigator. To determine the interjudge reliability, 10% of the entire data corpus (36 out of 360 speech samples) was randomly selected and analyzed the second time by the second investigator who was also experienced in VOT measurement. Absolute percent errors were calculated to indicate the intrajudge and interjudge reliability of measurement.

RESULTS
Reliability measurements
Absolute errors were used to determine the intra-judge and interjudge reliability. Results indicated an absolute error of 4.76% between the first and second measurements by the primary investigator. An absolute error of 13.94% was found between the measurements made by the primary and secondary investigators. Both intrajudge and interjudge reliability measures indicated that the primary investigator was consistent and reliable in the data analysis procedure.

VOT measurements
VOT values from speech samples produced by NL and SE speakers of Mandarin were measured and analyzed (Table 3). Repeated-measures (RM) ANOVA analyses were used to determine if VOT values were significantly different between NL and SE speakers and among the four tone levels. RM ANOVA revealed that significant main effects were found in speaker group \( F(1,172) = 104.905, \ P < 0.01 \), which indicated that VOT values related to NL speakers (99.62 ± 16.45 milliseconds) were significantly longer than those related to SE speakers (71.88 ± 21.44 milliseconds). There was no significant interaction between two independent variables: speaker group and tone level \( F(3,172) = 3.457, \ P < 0.02 \). There was no significant interaction between two independent variables: speaker group and tone level \( F(3,172) = 0.189, \ P = 0.904 \).
mean VOT values associated with different tones produced by SE and NL speakers. It was found that, for both SE and NL groups, VOT values associated with HL and HF tones were shorter than those associated with MR and FR tones. With respect to NL speakers, there was a significant difference in VOT among the four tone levels \( F(3,92) = 3.566, P < 0.02 \). Tukey’s post hoc tests indicated that VOT values associated with HF tone (90.94 ± 17.12 milliseconds) were significantly shorter than those associated with MR tone (104.35 ± 13.23 milliseconds) \( P < 0.01 \) and FR tone (103.19 ± 15.34 milliseconds) \( P < 0.05 \). With respect to SE speakers, no significant difference was found in VOT among the four tone levels \( F(3,80) = 0.894, P = 0.488 \).

Figure 2 shows the mean VOT values of three consonant stops in different speaker groups. It was found that the mean VOT values associated with SE speaker group were shorter than those associated with NL speaker group. RM ANOVA tests indicated that significant differences were found...
between SE and NL groups in VOT associated with /pʰ/ \([F(1,58) = 38.159, P < 0.01]\), /tʰ/ \([F(1,58) = 23.648, P < 0.01]\), and /kʰ/ \([F(1,58) = 50.321, P < 0.01]\). SE speakers consistently showed significantly shorter VOT values than NL speakers.

**DISCUSSION**

VOT is the relative timing measurement of two events: the release of the oral occlusion and the onset of voicing. Both the voicing source and the supralaryngeal system contribute to VOT. Apparently, using an alternative voicing apparatus, SE speakers are likely to exhibit VOT characteristics that are different from NL speakers. A number of other factors that may affect VOT have been suggested in the literature.\(^5,24-27\) This includes the tension of the vibratory structure and the pressure requirement below the voicing apparatus. The present study attempted to investigate the influence on VOT with the use of the PE segment as a new voicing source in SE speech. In addition, as tone levels are associated with different tension in the vibratory structure and pressure below the vibratory structure, the effect of tone levels on VOT was also investigated.

**Tone levels**

Despite the generally shorter VOT in SE speech than in NL speech (see discussion below), our data indicated that both NL and SE speakers showed similar VOT patterns across the four tone levels. Both exhibited longer VOT values in the MR and FR tones than the HF tone (Figure 2). This finding on NL speakers is consistent with that reported by Chen and Ng, in which VOT values were measured from /tʰV/ syllables produced at four contrastive tones.\(^28\) The longer VOT values in the MR and FR tones may be related to the fact that both tones have a rising component in their pitch contours.\(^29\) According to Wang et al, \(F_0\) and pitch level increase at certain points during the production of the MR and FR tones.\(^29\) The anticipated increase of tension in the voicing source and pitch may have delayed the onset of vibration of the vibratory structure. In studying excised larynges, Finkelhor et al suggested that VOT values increase with vocal fold longitudinal tension.\(^30\) With a greater vocal fold tension, the phonation threshold pressure (PTP) required start vocal fold vibration will concomitantly increase. According to Verdolini-Marston et al, threshold subglottal pressure positively affects VOT.\(^31\) It is noted that, in MR and FR tones, the anticipated increase in the tension of the vibratory structure, and the consequently greater PTP requirement may have delayed the onset of phonation. This suggestion may perhaps be further supported by the data reported by Xu and Sun.\(^32\) They found that NL speakers reduce their \(F_0\) more rapidly than when they raise \(F_0\) for inflections. This indicates that, physiologically, human vocal mechanism generally requires more time for preparation for higher pitch production as it needs to increase the tension in the voicing structure to prepare for a higher \(F_0\) during speech production. Yet, the detailed relationship between tension of the voicing source (vocal folds for NL speakers and PE segment for SE speakers) and VOT is still unknown. More information including aerodynamic and electromyographic data is needed to better understand the reason(s) for longer VOT in MR and FR tones.

In the present study, the four Mandarin tone levels were categorically represented by tone labels: HL, MR, FR, and HF tones. No attempt was made to examine how the actual \(F_0\) values affect VOT for the following reasons: (1) the \(F_0\) contours for different tones are complicated, (2) the \(F_0\) contours produced by the speakers were not consistent, and (3) the speakers were instructed to produce the different Mandarin tones, not the different \(F_0\) values.

**Speaker group**

The present study also attempted to determine if VOT varies with phonation type. Our data indicated that SE speakers exhibited significantly shorter VOT values than NL speakers. This indicates that, with the use of the PE segment as the voicing source, SE speakers tend to start phonation sooner than NL speakers. It takes a shorter time for SE speakers to start PE segment vibration than NL speakers to start vocal fold vibration during the production of stop consonants. This is consistent with the data previously reported on English-speaking,\(^13,16\) Thai-speaking,\(^19\) and Hebrew-speaking\(^33\) SE speakers.
The shorter VOT in SE phonation is believed to be related to the use of PE segment as the vibratory source in SE speakers.

In the production of a stop, the two events that determine VOT are the release of oral occlusion and the onset of sound source vibration. Since the release of the stop occlusion can be viewed as the reference point in time, the onset of vibration of sound source is more deterministic of the VOT value. A number of factors that may contribute to VOT have been discussed. The greater intraoral pressure in the cavity before and behind the occlusion, the slower movement of articulators, or the wider contact area forming the occlusion implies that the supraglottic pressure will drop more slowly, and the resulting VOT will be longer.24–27

In the case of NL phonation, the onset of vocal fold vibration starts when two conditions are fulfilled: (1) the two vocal folds are sufficiently close to each other, and (2) the transglottal pressure differential is greater than the PTP. Intrinsic laryngeal muscles allow the voluntary abduction and adduction of the vocal folds.34 The onset of phonation in the syllable /hə/, for example, is initiated by the adduction of vocal folds. Vocal folds are being adducted at the same time when intraoral pressure behind the occlusion (or the supraglottic pressure) drops. If it takes a longer time for the vocal folds to be sufficiently adducted, or for the transglottal pressure differential to achieve the required PTP level, VOT will become longer.

However, PE segment vibration in SE phonation is a totally different situation. Composed of mainly the cricopharyngeus, the inferior pharyngeal constrictor, and the upper esophagus,35 the PE segment is normally constricted when not in phonation. In fact, spastic constriction of the PE segment after laryngectomy is not rare and patients often require subsequent myotomy to reduce muscle tonicity.36 This is in part attributed to the fact that SE speakers are using the upper part of the esophagus as the air reservoir. To retain the air in the high-pressure esophageal air reservoir, PE segment has to be constricted and tensed. Moreover, the exact location of the vibratory structure in SE phonation is not clear. Various locations of the vibratory structure in SE speech have been suggested. The most widely accepted one believes that the cricopharyngeus muscle, which is located between the fifth and the sixth cervical vertebrae, contributes the major part of the vibratory structure in SE speech.37 The nearby mucous membrane, cricotracheal band, and muscular remnant may serve as the basis of PE segment vibration.36 Fine muscular control of adduction and abduction of the PE segment is not naturally possible. SE phonation appears to be initiated by relaxing the PE segment. There are certain muscles around the PE region that can counteract the tight contraction of the PE segment, so that the mucosal edge of the PE segment can be set into vibration.37 Hirose15 reported that voicing distinction is contributed by the mucosal margin of the esophageal orifice and that vibration appears to develop at the highest level of the reconstructed esophageal passage. In fact, it has been reported that the control of PE segment relaxation is closely related to successful esophageal voice.36 The more relaxation achieved in the PE segment, the better the voice production.

According to Isshiki and Tanabe,39 and Hirose,15 intraoral pressure found in SE speakers during the production of a voiceless stop is much higher than that found in NL speakers, sometimes reaching as high as 60 cm H2O. Assuming an adducted neoglottis, such high intraoral pressure is created by quick and tense movements of the articulators in the oral cavity. This may be created to offset the high subpseudoglottic pressure found in the esophagus. Therefore, the suggestion that the low intraoral pressure in SE speakers helps achieve the required PTP more quickly appears to be incorrect.

The significantly shorter VOT in SE phonation should be accounted for by the fact that, unlike the vocal folds in NL speakers which have to be adducted to start vibration, the PE segment just needs to be relaxed to start phonation. It might take a shorter time for the PE segment to achieve a certain relaxed state for PE segment vibration to occur, as compared with vocal fold vibration. In addition, Christensen and Dwyer suggested that the intraoral pressure drops more quickly in SE speakers than NL speakers due to the smaller cavity size behind the occlusion.18 Scar tissue is not uncommon in laryngectomized individuals after radiation therapy which unavoidably reduces the cavity size in SE speakers. The smaller supraglottic cavity in SE
speakers and shorter time to relax the PE segment to achieve appropriate tension for vibration may justify the shorter VOT values in SE speakers than NL speakers. Apparently, such speculations should further be confirmed by additional data such as visual images of a vibrating PE segment. Future research should focus on the muscular activity in the PE segment during the production of stops.

**Clinical implications**

VOT is an important perceptual cue for aspiration and voicing. In Mandarin Chinese where all stop consonants are voiceless, VOT is used by listeners to distinguish between aspirated and unaspirated stops. Aspirated stops are generally associated with longer VOT values than the unaspirated counterparts. However, VOT is significantly shorter in SE speakers than NL speakers. Should there be no other perceptual cues available, distinguishing between aspirated and unaspirated Mandarin stops in SE speech relying only on VOT may be problematic. Data obtained from the present study provide insights into the characteristics of VOT production in individuals using the PE segment as a new voicing source and should be of assistance to speech-language pathologists in charge of developing speech therapy regimes for the SE speakers of Mandarin. To improve SE speech, postlaryngectomy speech therapy should target at such VOT discrepancy and should focus on how to attain proper VOT values, or more importantly focus on achieving categorical VOT difference between aspirated and unaspirated stops in SE speech.

**Acknowledgments:** This work was supported by National Natural Science Foundation of China (Grant No. 30070212 and 69925101).

**REFERENCES**


