Vocal Attack Time of Different Pitch Levels and Vowels in Mandarin

Authors: Ruifeng Zhang,, R.J. Baken, Jiangping Kong

Date: 2019-06-20T00:00:00+00:00

Abstract

The purpose of this study was to investigate how vocal attack time (VAT) varies when young adults articulate the three vertex vowels in Mandarin Chinese at five linguistically unconstrained pitch levels. Sound pressure (SP) and electroglottographic (EGG) signals were recorded simultaneously from fifty-three male and fifty-three female subjects saying sustained /A/, /i/ and /u/ at five equally spaced pitch heights, each being higher than the preceding one. Then analyses of means, variance and correlation were performed to explore the relationships of VAT/pitch levels and VAT/vowels. Findings were: As mean STs (semitone) increase linearly from levels one to five, mean VATs decrease nonlinearly in a big group of subjects but increase nonlinearly in a small group of them. Based on the body-cover model of F0 control, data here lead to the guess that different people incline to use different strategies in increasing pitch height. When males, females and males plus females are considered as a whole, average STs and VATs tend to be positively correlated among the three vertex vowels.

Full Text

Preamble

Vocal Attack Time of Different Pitch Levels and Vowels in Mandarin
Ruifeng Zhang, R.J. Baken, and Jiangping Kong
Beijing, China; Woodstock, New York, USA; and Hong Kong University, Hong Kong SAR, China

Summary: This study investigated how vocal attack time (VAT) varies when young adults articulate the three vertex vowels in Mandarin Chinese at five linguistically unconstrained pitch levels. Sound pressure (SP) and electroglottographic (EGG) signals were recorded simultaneously from fifty-three male and fifty-three female subjects producing sustained /A/, /i/, and /u/ at five equally spaced pitch heights, each higher than the preceding one. Analyses of means,

variance, and correlation were performed to explore the relationships between VAT and pitch levels and between VAT and vowels. The findings revealed that as mean semitones (ST) increased linearly from levels one to five, mean VATs decreased nonlinearly in a large group of subjects but increased nonlinearly in a small group.

Based on the body-cover model of F0 control, these data suggest that different individuals tend to use different strategies when increasing pitch height. When males, females, and the combined group are considered as a whole, average STs and VATs tend to be positively correlated across the three vertex vowels.

Key Words: Vocal attack time–Pitch levels–Vertex vowels–Semitone.

INTRODUCTION

Vocal attack time (VAT) is the time lag between the rise of simultaneously recorded sound pressure (SP) and electroglottographic (EGG) signals, measured at the onset of phonation [**?**]. When airflow passes through the glottis during vowel or voiced consonant initiation, the vocal folds oscillate with very small amplitudes before their first contact is achieved and stabilized. Consequently, the SP signal, which records sound pressure emitted from the mouth, begins growing in amplitude well before the vocal folds touch each other. In contrast, the EGG signal, which records vocal-fold contact area, shows nearly no amplitude until vocal fold contact occurs, after which its magnitude appears and grows. This results in positive VAT values. However, in cases where EGG signal initiation precedes SP signal initiation–such as in a hard glottal attack where vocal fold contact occurs before the appearance of SP signals–VAT is negative. When both signals rise simultaneously, VAT equals zero. Therefore, VAT can be understood as the duration from the start of vocal-cord oscillation to the instant of first vocal-cord contact, providing a useful index of pre-phonation laryngeal adjustment.

The effectiveness of VAT measurement was experimentally verified by Orlikoff et al. [**?**] using five vocally normal subjects. EGG and SP signals of different phonation types were recorded synchronously with high-speed video-endoscopy, from which a digital kymogram (DKG) was generated. DKG attack duration data obtained manually were then compared with VAT measures extracted using computer programs. The strong and direct relationship between VAT and DKG-measured data proved VAT to be a valid and convenient measure of vocal attack. In 2012, Roark et al. [**?**] proposed a figure of merit (FOM) that assesses a critical assumption of vocal startup underlying the VAT measure and thus represents the integrity of the derived measure. SP and EGG signals from 102 tokens were visually inspected to empirically derive a criterion FOM level of less than 0.75, indicating when the measurement assumption had failed and the obtained VAT value should be disregarded. An example of VAT use in nonlinguistic research is the normative data measurement by Roark et al. [**?**], who collected SP and EGG signals from fifty-five males and fifty-seven females performing multiple

tokens of three tasks (sustained /a/, "always," and "hallways") at comfortable pitch and loudness. Average VATs were significantly shorter for females than for males, with a mean VAT of 1.98 milliseconds in the screened sample of normal young speakers. The use of VAT in linguistic research was exemplified by measurements conducted by Ma et al. [**?**], who examined the association between VAT and tone in Cantonese speakers.

It is well-established that pitch increases with acceleration of vocal fold oscillation, resulting from progressive augmentation of vocal fold tension. The VAT study of three phonation types by Orlikoff et al. [**?**] suggests that tenser vocal folds tend to be associated with smaller VAT values. Therefore, how VAT varies with increasing pitch in Mandarin Chinese represents an attractive research subject that has not previously been explored. We required subjects to produce vowels at five different linguistically unconstrained pitch levels for two reasons. First, for the purpose of devising tone-letters, Chao [**?**] divided a person' s pitch range into four equal parts with five points numbered 1, 2, 3, 4, 5, corresponding to low, half-low, medium, half-high, and high, respectively. Subsequent linguistic researchers have found that no language uses more than five pitch levels to distinguish its tones [**?**]. Second, many subjects found it natural and easy to space five pitch levels equally within their voice range but difficult to manage six or more levels in this manner. Thus, by focusing on five sustained pitch heights that are not linguistically distinctive, the present study is intended to facilitate future work on language tones. The three vertex vowels /A/, /i/, and /u/ in Mandarin Chinese were chosen for production at five pitch levels because they occupy the extreme points on the vowel chart and represent the entire scope of tongue movement during articulation. In summary, this study aims to explore how VAT varies when young people produce the three vertex vowels in Mandarin Chinese at five linguistically unconstrained pitch heights.

METHOD

Subjects and Instrumentation

Fifty-three females (18 to 22 years old) and fifty-three males (18 to 22 years old), all college students, participated in the research. They spoke standard Mandarin Chinese for daily communication, had no voice or hearing problems, and were in good health at the time of recording. Recording was conducted in a sound-treated booth at the Language Laboratory of the Chinese Department, Beijing University, where background noise was below 25 dBA. Adobe Audition 2.0 on a Lenovo x220i computer was set to stereo interface with a sampling rate of 44100 Hz and a resolution of 16 bits per channel. The electroglottograph (Model 6103) used for collecting EGG signals and the microphone and sound card (Creative Labs Model No. sb1095) used to obtain SP signals were synchronously connected to the computer through a sound console (Behringer XENYX502).

With their lips approximately 10 cm from the microphone, subjects were asked to produce sustained /A/, /i/, and /u/ at five pitch heights, each higher than

the preceding one. All pitch levels were repeated twice, yielding 30 tokens (3 × 5 × 2) per speaker.

Parameter Extraction

F0, VAT, and FOM measures were extracted largely automatically from the speech samples using software developed by Roark et al. [**?**], which processed signals in four stages: signal verification, signal segmentation, F0-based frequency filtering and signal modeling, and measure extraction. From the 3180 samples (1590 for males and 1590 for females), 3165 values (1590 for males and 1575 for females) were obtained for each of the three parameters, with 15 female speech recordings unable to be evaluated by the software, possibly due to poor EGG signal quality.

Data Preprocessing

Since our research required subjects to pronounce each of the three vowels with increasing pitch heights, the F0 values they produced for each vowel should theoretically increase with level shifts from one to five. Consequently, a correlation analysis first discarded 210 ineffective speech samples (140 for females and 70 for males) whose pitch values showed a negative correlation with pitch level numbers. The remaining 2955 measures were then divided into ten groups: male-level1, male-level2, male-level3, male-level4, male-level5, female-level1, female-level2, female-level3, female-level4, and female-level5. Each group was processed separately in the same manner: first, measures beyond $±$3 standard deviations from the mean F0 were removed from each group, since vocal fold vibrations normally do not exceed 500 Hz; second, based on the observation that there were more outliers among VAT values, measures beyond $±$2 standard deviations from the mean VAT were eliminated.

Among the 2827 measures that remained, F0 ranged from 77.3 Hz to 497.06 Hz, with a mean (SD) of 220.35 Hz (75.26 Hz), and VAT ranged from -56.26 ms to 53.13 ms, with a mean (SD) of 0.75 ms (8.69 ms). For Excel and SPSS analyses below, all F0 values were converted to semitones (ST) re 64.66 Hz. This reference level was chosen not only because it was close to the minimum pitch value of 77.3 Hz but also because Liu [**?**] had used it in his groundbreaking research on Mandarin tones.

RESULTS

VAT and Pitch Levels

Among the 1458 male speech samples, pitch ranged from 3.09 ST to 26.06 ST (mean = 15.72 ST; SD = 4.19) and VAT from -25.67 ms to 39.27 ms (mean = 0.91 ms; SD = 7.46). Among the 1369 female speech samples, pitch ranged from 16.45 ST to 35.31 ST (mean = 24.98 ST; SD = 3.49) and VAT from -56.26 ms to 53.13 ms (mean = 0.59 ms; SD = 9.84 ms). One-way analyses of variance and post

hoc tests showed that at significance level p = 0.01 for both males and females, ST values were significantly different between any two of the five pitch levels (p < 0.01 for all), exceeding a preselected $\alpha = 0.01$. For VAT, significant differences were only observed between pitch level one and each of levels two, three, four, and five (p < 0.01 for all). According to correlation analysis, ST showed a significant negative correlation with VAT across all 2827 speech samples (N = 2827, r = -0.077, p < 0.01). When male and female measures were calculated separately, the significant negative correlation remained for both groups but with a higher correlation strength for females than for males (N = 1458, r = -0.084, p < 0.01 for males; N = 1369, r = -0.115, p < 0.01 for females).

Table 1 lists the maximum, minimum, mean, and standard deviation of ST and VAT across pitch levels calculated according to three groupings: A: males (1458 tokens); B: females (1369 tokens); and C: total (2827 tokens). Here, vowel considerations and individual differences are temporarily set aside. As pitch levels shift from one to five, all three groups show a linear increase in pitch but a nonlinear and non-monotonic decrease in VAT: unlike ST means, each mean VAT value from Levels Two to Five is not always larger than the one that follows. However, in all cases, the average VAT at Level One is always the largest among the five pitch levels and is substantially larger than the mean VATs at Levels Two, Three, Four, and Five, creating an overall downward trend. Mean STs and VATs as a function of pitch levels for the three groups are shown in Figure 1 [Figure 1: see original paper], which illustrates VAT variations more intuitively compared with ST changes: from Level One to Levels Two and Three, there is a sharp decline, followed by a clear upturn at Level Four, and then a steep dip at Level Five to the minimum mean VAT value.

Now consider how individuals produced the five pitch levels differently. Since each pitch level was repeated consecutively (e.g., /A/level 1 → /A/level 1 → /A/level 2 → /A/level 2 → /A/level 3 → /A/level 3 → /A/level 4 → /A/level 4 → /A/level 5 → /A/level 5, etc.) and the pitch values were nearly identical, they can and must be averaged to examine inter-subject differences associated with ST-VAT correlation while ignoring slight differences between tokens repeated by the same person. The findings after this averaging are displayed in Figure 2 [Figure 2: see original paper]. Among the 103 subjects (52 males and 51 females) whose data were retained for analysis, 36 (35%) produced all /A/, /i/, and /u/ with negative VAT-ST correlation coefficients, 25 (24%) uttered two of the three vowels with negative VAT-ST correlation coefficients and one with positive coefficients, 20 (19%) pronounced one vowel with negative VAT-ST correlation and two with positive correlation, and 11 (11%) articulated all three vowels with positive VAT-ST correlation coefficients. The remaining 11 subjects (11%) could not be categorized because measures for one or two of their vowels were culled during data preprocessing. In summary, many subjects produced increasing pitch heights with declining VAT, but a small number did so with increasing VAT, and naturally there were also subjects belonging to intermediate types.

Table 2 lists the means and standard deviations of ST and VAT across five pitch levels for the 36 subjects who uttered all three vowels with negative VAT-ST correlation coefficients and the 11 who did so with positive coefficients. Their mean STs and VATs as a function of pitch levels are shown in Figure 3 [Figure 3: see original paper]. The VAT-ST co-variation of the former group is illustrated in Figure 3(a): as pitch levels shift linearly from one to five, mean VATs drop gradually except for a slight upturn at pitch level four. The latter group shows a quite different pattern: as pitch levels increase linearly from one to five, average VATs increase gradually except for small turns at pitch levels two and four (Figure 3b).

VAT and Vowels

To examine VAT variation among different vowels while setting aside pitch level considerations, mean VATs and STs of /A/, /i/, and /u/ were calculated for Groups A: males (1458 tokens), B: females (1369 tokens), and C: total (2827 tokens). The resultant means for the three groups are listed in Table 3 and displayed in Figure 4 [Figure 4: see original paper], which shows that the average STs of /A/, /i/, and /u/ in Groups A, B, and C are all ordered as: /u/ > /i/ > /A/, although they differ only slightly from one another. The mean VATs for these three groups follow the same pattern: /u/ > /i/ > /A/, with average VATs of /u/ being the largest among the three vowels and substantially larger than those of the other two vowels. High vowels tend to have both larger pitch values and longer VATs than low vowels.

DISCUSSION

All analyses concerning VAT and pitch levels point to the same finding. As shown in Figure 1, when all males, females, or males plus females are considered as a group, a nonlinear contra-variant relationship can be observed between average VAT and ST values at five pitch levels, because the best trend line for increasing pitch is linear while that for decreasing VAT is cubic. When individuals are considered, a slightly different picture emerges: many people tend to produce the three vertex vowels (/A/, /i/, /u/) with negative VAT-ST correlation coefficients, but a small number of subjects incline to produce them with positive coefficients (see Figures 2 and 3). However, the nonlinear relationship still holds in both cases, because although Figure 3(a) presents a contra-variant VAT-ST relationship while Figure 3(b) shows an orthokinetic one, the best-fit lines for increasing pitches in both graphs are linear while those for decreasing or increasing VAT remain cubic. The author listened to and compared subjects who pronounced five pitch heights of a vowel with significantly negative VAT-ST correlation and those who did so with significantly positive correlation, and the impression was that the former sounded much more laborious than the latter at pitch level five.

Watson et al. [**?**] reported that adjusted mean VAT for their high frequency condition was smaller than adjusted mean VAT for their mid and low frequency

conditions and that VAT appears to be sensitive to increases in vocal fold tension in normal speakers. The present findings not only accord with their statements but also suggest different strategies people tend to use when increasing pitch height.

The body-cover model of F0 control proposed by Titze [**?**] concerns the activities of cricothyroid (CT) and thyroarytenoid (TA) muscles. CT contraction elongates the vocal folds, while TA contraction tends to shorten them. In combination, these two muscles are responsible for most of the length change that can be achieved. TA muscle activity also regulates the effective depth of vocal fold vibration, which reduces rapidly as pitch increases. At low to intermediate F0 and loud productions—that is, in modal voice conditions—both CT and TA activities are relatively low, and a significant portion of the vocal fold body vibrates while the mucosa and ligament remain somewhat lax. A rise in F0 is generally obtained by increased TA activity, as long as CT activity is not near its maximum. As pitch increases from high to falsetto or when a high F0 is to be achieved (as in high-pitched singing), CT activity gradually becomes dominant while TA action decreases. Only the surface of the vocal fold vibrates, with a stiff ligament and loose mucosa in combination [**?**]. Some hypotheses can be made based on this model of pitch control and the data reported here. Many people are not skilled at using falsetto, and when required to utter vowels at five increasingly higher pitch levels, tend to start at a very low point and progress step by step through the modal voice register with vocal fold tension being increased gradually. However, quite a few speakers, untrained but skillful in voice use, prefer to start at a relatively higher point and gradually approach falsetto, where the vocal folds conversely become somewhat slack. These may be the reasons why the former group displayed a negative VAT-ST correlation while the latter showed a positive one.

Zhang [**?**] and Dong [**?**] reported that the three Chinese vertex vowels, when pronounced comfortably, display their intrinsic F0 in the pattern: /u/ > /i/ > /A/. The results here support their finding. Figure 4 indicates that both mean STs and mean VATs of /A/, /i/, and /u/ in men, women, and all subjects combined are ordered as: /u/ > /i/ > /A/, suggesting that VAT of the three vowels normally tends to be positively correlated with their intrinsic pitch. However, what causes such a relationship needs further exploration.

CONCLUSION

The purpose of this study was to investigate how VAT varies when young adults articulate the three vertex vowels in Mandarin Chinese at five linguistically unconstrained pitch levels. In a large group of young adults, pitch and VAT changed in opposite directions, but in some individuals the two varied in the same direction. In all cases, however, pitch and VAT tended to present a nonlinear relationship. It is possible that people, out of habit or due to laryngeal physiology, tend to manipulate their vocal folds in two different ways when increasing pitch from low to high levels. Those who utter vowels at increas-

ing pitch levels with positive VAT-ST correlation coefficients may have greater potential for using falsetto, which of course needs further investigation.

Following this research, our study on VAT of lexical tones in Mandarin Chinese is now underway.

Acknowledgements: Thanks go to all the voice experts for their kind participation in the investigation. This research was funded by the National Natural Sciences Foundation of China, grant number 61073085.

REFERENCES

1. Orlikoff RF, Deliyski DD, Baken RJ, Watson BC. Validation of a glottographic measure of vocal attack. J Voice. 2009; 23:164-168.

2. Roark RM, Watson BC, Baken RJ. A figure of merit for vocal attack time measurement. J Voice. 2012; 26:8-11.

3. Roark RM, Waston BC, Baken RJ, Brown DJ, Thmas JM. Measures of vocal attack time for healthy young adults. J Voice. 2012; 26:12-17.

4. Ma EP-M, Baken RJ, Roark RM, Li P-M. Effect of tones on vocal attack time in Cantonese speakers. J Voice. 2012; 26: 670.e1-670.e6.

5. Chao Y R. A system of "tone-letters". 方言. 1980; 2: 81-83.

6. Maddieson I. Universals of tone. Universals of human language. 1978; Volume 2: 338.

7. 刘复. 乙二声调推算尺. 史语所集刊. 1934; 4 本 4 分: 355-361.

8. Watson BC, Baken RJ, Roark RM, Reid S, Ribeiro M, Tsai W. Effect of fundamental frequency at voice onset on vocal attack time. J Voice. 2013; 27: 273-277.

9. Titze IR. Principles of voice production. 2nd ed. USA: National Center for Voice and Speech; 2000: 211-242.

10. 张家騄. 元音的内在基频与讲话方式对共振峰的影响. 声学学报. 1989; 14: 401-406.

11. 董倩倩. 汉语普通话元音音高再探. 文教资料. 2010: 27-28.

Note: Figure translations are in progress. See original paper for figures.