Vocal Attack Time of Different Pitch Levels and Vowels in Mandarin

*Ruifeng Zhang, †,‡R. J. Baken, and *Jiangping Kong, *Beijing and ‡Hong Kong SAR, China, and †Valhalla, New York

Summary: 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 and electroglottographic signals were recorded simultaneously from 53 male and 53 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 1 to 5, 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 F_0 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.

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

INTRODUCTION

Vocal attack time

Vocal attack time (VAT) is the time lag between the rise of the simultaneously recorded sound pressure (SP) and electroglottographic (EGG) signals, measured at the onset of phonation.[1] When airflow goes through the glottis during the initiation of a vowel or a voiced consonant, the vocal folds oscillate with very small amplitudes before their first contact is achieved and stabilized. Therefore, the SP signal, which is a record of SP emitted from the mouth, begins its growth of amplitude well before the vocal folds touch each other, but the EGG signal, as a record of vocal-fold contact area, has nearly no amplitude until the vocal fold contact occurs, and only after that does its magnitude show up and grow. This is the occasion when VAT values are positive. However, in other cases when the initiation of EGG signals leads that of SP signals, namely, when vocal fold contact precedes the appearance of SP signals such as in a hard glottal attack, VAT is negative. If the two sorts of signals rise at the same point of time, VAT equals zero. Consequently, VAT can be taken to be the duration from the start of vocal cord oscillation to the instant for the first vocal cord contact and provides a useful index to indicate the prephonation laryngeal adjustment.

The effectiveness of VAT measurement was experimentally verified by Orlikoff et al.[1] 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. The DKG attack duration data obtained by hand were then compared with the VAT measures extracted by using computer programs. The strong and direct relationship between the VAT- and DKG-measured data proved VAT to be a valid and convenient measure of vocal attack. In 2012, a figure of merit (FOM), which assesses a critical assumption of vocal startup on which the VAT measure is based and therefore represents integrity of the derived measure, was proposed by Roark et al[2] for the VAT measurement of sustained /a/. SP and EGG signals from 102 tokens were visually inspected to empirically derive a criterion level of FOM less than 0.75 to indicate when the assumption underlying the measurement had failed and the VAT value obtained should be disregarded. The example of using VAT for nonlinguistic research was the measurement by Roark et al[3] acquiring normative data of VAT in healthy young adults. They collected SP and EGG signals from 55 males and 57 females performing multiple tokens of three tasks (sustained /a/, "always," and "hallways") at comfortable pitch and loudness. The average VATs were significantly shorter for females than for males and mean VAT was 1.98 milliseconds in the screened sample of normal young speakers. The use of VAT in linguistic research was exemplified by the measurements done by Ma et al[4] examining the association between VAT and tone in Cantonese speakers.

Pitch levels and vowels

It is well known that pitch increases along with the acceleration of vocal fold oscillation, resulting from step-by-step augmentation of the vocal folds' tension. The VAT study of three phonation types by Orlikoff et al[1] seems to suggest that tenser vocal folds tend to be associated with smaller VAT values. Therefore, how VAT varies with increasing pitch in Mandarin Chinese appears to be an attractive research subject that has never been touched upon. We required the subjects to produce vowels at five different linguistically unconstrained pitch levels out of two considerations. For the purpose of devising tone letters, Chao[5] divided the pitch range of a person into four equal parts with five points numbered 1, 2, 3, 4, and 5, corresponding to low, half-low, medium, half-high, and high, respectively. It has also been found by subsequent linguistic researchers that no language uses over five pitch levels to distinguish its tones.[6] On the other hand, many subjects felt it natural and easy to

Accepted for publication July 15, 2014.
This research was funded by the National Natural Sciences Foundation of China and the grant number was 61073085.

From the *Department of Chinese Language and Literature, Research Center for Chinese Linguistics and Joint Center for Language and Human Complexity, Peking University, Beijing, China; †Department of Speech-Language Pathology, New York Medical College, Valhalla, New York; and the ‡Hong Kong University, Hong Kong SAR, China.
Address correspondence and reprint requests to Jiangping Kong, Department of Chinese Language and Literature, Peking University, Beijing 100871, China. E-mail: kongjp@gmail.com

Journal of Voice, Vol. 29, No. 5, pp. 542-547
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2014.07.010

TABLE 1.
Maximum, Minimum, Mean, and Standard Deviation of ST and VAT Across Pitch Levels in Groups A: Males, B: Females, and C: Total

Groups	A: Males (n = 1458)				B: Females (n = 1369)				C: Total (n = 2827)			
Pitch Levels	Max	Min	Mean	SD	Max	Min	Mean	SD	Max	Min	Mean	SD
1												
Pitch (ST)	18.87	3.09	11.46	3.09	26.40	16.45	20.75	1.93	26.40	3.09	15.96	5.32
VAT (ms)	39.27	−21.11	2.69	8.64	53.13	−46.24	3.33	12.26	53.13	−46.24	3.00	10.54
2												
Pitch (ST)	21.37	5.19	14.25	3.20	29.37	17.97	23.46	1.94	29.37	5.19	18.74	5.32
VAT (ms)	22.83	−22.70	0.59	7.35	27.73	−36.74	−0.27	8.82	27.73	−36.74	0.17	8.10
3												
Pitch (ST)	23.03	7.91	16.15	3.32	31.44	19.55	25.22	2.10	31.44	7.91	20.53	5.33
VAT (ms)	20.45	−21.72	0.39	6.56	31.93	−29.37	0.07	8.18	31.93	−29.37	0.24	7.38
4												
Pitch (ST)	23.40	8.73	17.70	3.24	34.07	21.02	26.94	2.49	34.07	8.73	22.21	5.45
VAT (ms)	20.11	−21.22	0.59	7.00	26.44	−23.61	0.95	7.69	26.44	−23.61	0.77	7.34
5												
Pitch (ST)	26.06	11.01	19.11	3.26	35.31	23.26	28.66	2.36	35.31	11.01	23.68	5.57
VAT (ms)	15.85	−25.67	0.27	7.36	27.57	−56.26	−1.20	10.98	27.57	−56.26	−0.43	9.29

space five pitch levels equally within their voice range, but difficult to manage six or more levels that way. So the present study, by focusing on five sustained pitch heights that are not linguistically distinctive, is intended to be conducive to future work on language tones. The three vertex vowels /A/. (Since /A/ in Mandarin is the lowest central vowel and different from the front /a/ and back /a/ on the IPA (International Phonetic Association) chart, Chinese linguists prefer to represent it with a capital letter.) /i/ and /u/ in Mandarin Chinese were chosen to be produced at five pitch levels because they occupy the utmost points on the vowel chart and represent the entire scope of tongue movement during articulation. All in all, the purpose of this study was 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 (aged 18–22 years) and 53 males (aged 18–22 years), all of whom were college students, participated in the research. They spoke standard Mandarin Chinese for daily communication, had no voice or hearing problems, and were all in good health at the time of recording. The recording was accomplished in the sound-treated booth at the Language Laboratory of the Chinese Department, Beijing University, where the background noise was below 25dBA. The *Adobe Audition 2.0* on the computer (Lenovo, x220i, Lenovo, Beijing, China) was set at the stereo interface with a sampling rate of 44 100 Hz and a resolution of 16 bits for each channel. The electroglottograph (Model 6103) used for collecting EGG signals, and the microphone and sound card (Creative Labs Model No.sb1095) used to gain SP signals were synchronously connected to the computer through a sound console (Behringer XENYX502, MUSIC Group Macao Commercial Offshore Limited, Zhongshan, Guangdong, China). With their lips about 10 cm away from the microphone, the subjects were asked to say sustained /A/, /i/, and /u/ at five pitch heights, each of them being higher than the preceding one. All pitch levels were repeated twice and 30 tokens (3 × 5 × 2) were obtained from each speaker.

Parameter extraction

F0, VAT, and FOM measures were extracted largely automatically from the speech samples using the software developed by Roark et al,[3] which processed signals in four stages: signal verification, signal segmentation, F_0-based frequency filtering and signal modeling, and extraction of measures. 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 the poor quality of their EGG signals.

Data preprocessing

Because our research required the subjects to pronounce each of the three vowels with increasing pitch heights, the F_0 values they produced for each vowel should theoretically increase with the level shift from 1 to 5. Consequently, a correlation analysis was first done that discarded 210 ineffective speech samples (140 for females and 70 for males) whose pitch values had a negative correlation with the pitch level numbers. The 2955 measures left were then divided into 10 groups: male-level 1, male-level 2, male-level 3, male-level 4, male-level 5, female-level 1, female-level 2, female-level 3, female-level 4, and female-level 5. Each group was processed separately in the same way: because vocal fold vibrations normally do not go beyond 500 Hz, measures that were beyond ±3 standard deviation (SD) from the mean F_0 were first removed from each group; and second, according to the observation that there