Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Tones do not disappear in singing: the duration of Mandarin tones in the music context

Qianyutong Zhang[1], *Lei Zhu*[1], *Xiaoming Jiang*[1]

[1]Institute of Linguistics, Shanghai International Studies University
qianyutong_zhang@shisu.edu.cn, zhulei@shisu.edu.cn, xiaoming.jiang@shisu.edu.cn

Abstract

The primary acoustic cue of tones is F0, but secondary cues such as duration and intensity can also distinguish tones from one another. When pitch information is unavailable, do speakers utilize any of the secondary acoustic features to realize tones? This study situates Mandarin tones in a music context and explores tonal production patterns by speakers under different musical notes. The vowel duration of the four Mandarin tones in the context of normal speech and singing from sixteen native Mandarin speakers were analyzed. It was found that tones in the music context shows partially similar patterns with those in a normal speech context: syllables with T4 had significantly shorter vowel durations than those with other tones. However, the vowel duration of T1 increased, hence T3 was no longer the longest one. The results suggest that when pitch information is not available for the realization of tones, speakers may partially rely on vowel duration cues to express tonal contrasts, though the duration patterns of tones are affected by different communicative contexts.

Index Terms: Mandarin tones, music context, vowel duration

1. Introduction

Mandarin Chinese is a tonal language with four lexical tones, characterized by different patterns of fundamental frequency (F0) changes [1]: Tone 1 (T1) is high-level, Tone 2 (T2) is mid-rising, Tone 3 (T3) is low-dipping, and Tone 4 (T4) is high-falling [2-3]. These tones are highly functional in the distinction between morphemes, which are mostly monosyllabic. For instance, the syllable /ma/ with Tones 1 to 4 (denoted as /ma1/, /ma2/, /ma3/ and /ma4/) can mean “mother”, “hemp”, “horse” and “name-calling” respectively.

Besides primary cues in pitch height and contour, however, Mandarin tones also show significant differences in temporal cues, such as duration. In isolated monosyllables, T3 has the longest duration and T4 the shortest, with T1 and T2 falling in between [4-5]. As for why duration varies among tones, it is presumed that a more complex contour causes a longer articulation time [6].

In normal communicative contexts, duration cues are secondary and redundant for the realization and recognition of tones. However, when the primary cues are insufficient or unavailable, the secondary cues may play an indispensable role [7-10]. Tones in whispered speech are typical cases that several previous studies have focused on, since whisper offers a context in which F0 is eliminated, so that speakers can only use secondary cues to convey tone contrasts. In the whisper context, speakers tend to reserve or even exaggerate the secondary cues of tones [11-13]. For example, it has been found that the difference in duration between the longest T3 and the shortest T4 in Mandarin is more distinct in whispered than normal speech [12-13].

However, in addition to the whisper context, there are other contexts in which pitch may not be available for the realization of tonal contrasts, such as singing. Unlike in whisper, in the singing context, the F0 exists but carries musical notes whose demands on the vocal folds often conflict with those of lexical tones [6]. It is still unknown whether the pattern of secondary cues of tones is maintained as in normal or whispered speech. Specifically, for example, if the longer duration for T3 is caused by its complex dipping contour [6], such a cause will no longer exist in the singing context, where tone contours, whatever they are, are replaced by single musical notes. It remains to be answered whether in this situation T3 will still have the longest duration or is expected to be sung for as long as the other tones?

Hence, this study aims to explore whether speakers maintain intrinsic duration patterns of different tones in a singing context where pitch is regulated. As previous studies on whisper found expressions of tones in the absence of F0 in line with normal speech, we expect to find a similar duration pattern in the singing context.

2. Method

2.1. Participants

Sixteen native Mandarin speakers were recruited to record the materials (8 males and 8 females; mean age = 24.1 yr). All speakers reported Mandarin Chinese as their first and daily-used language, and English as a foreign language from school education. Among all the speakers, 7 reported music experiences (such as playing the violin or singing Peking Opera), 9 reported other foreign language skills (such as French, Korean, or Russian). None of the speakers have reported language or speech disorders, or any neurological or psychiatric disorders.

2.2. Materials

Three syllables (/ta//ti//tu/) with four Mandarin tones (T1, T2, T3 and T4) and 3 musical notes (G3, A3 and B3, only in music context task) were selected as materials. The 12 combinations of syllables and tones (3 syllables × 4 tones) were presented in the form of Chinese characters (see Table 1), which are all commonly used in daily communication. Another 8 words with various syllables and tones were set as fillers to prevent fatigue in target tone production.

10.21437/SpeechProsody.2024-70

Table 1: *Word list for recording*

	Tone 1	Tone 2	Tone 3	Tone 4
/ta/	搭	答	打	大
/ti/	低	笛	底	地
/tu/	督	读	堵	度

2.3. Procedure

Materials were recorded in a quiet room using a TASCAM DR-07X microphone for 10 speakers, and the other 6 speakers recorded remotely using their built-in mobile phone recorders. All recordings were made at a sampling rate of 44.1 kHz and a 16-bit resolution.

The whole recording task was composed of two sessions: normal speech task and singing task. All recording materials were presented randomly one by one via PsychoPy [14]. For both sessions, participants were asked to repeat each word on the screen orally five times, and then press "space" to switch to the next word. Before the singing task, speakers were taught to familiarize themselves with certain music notes, namely G3 ($\approx$ 197 Hz), A3 ($\approx$ 221 Hz) and B3 ($\approx$ 247 Hz), played by the experimenter on an app named "Perfect Piano". For each word in the singing session, participants first saw a Chinese character on the screen, and then heard a music note lasting for 2 seconds after 1 second. When the sound ended, participants sang the word five times with the given note.

In the normal speech session, each participant produced 60 target tokens (3 syllables $\times$ 4 tones $\times$ 5 repetitions). In the singing session, the target token number for each speaker was 180 (3 syllables $\times$ 4 tones $\times$ 3 notes $\times$ 5 repetitions). The recordings were saved as WAV files. With certain tokens excluded due to low quality or errors of recording, a total of 3444 pieces of data were included in the following analysis.

2.4. Acoustic analysis and statistic modeling

The tone of each token was annotated manually using TextGrids in Praat version 6.1.24 [15], and the average duration of each annotated tone was extracted using ProsodyPro [16]. Following previous researches [2, 17-18] which suggested that it is only the syllable nucleus that carries the tone, the present study excluded the initial consonants when analyzing tone duration. For each participant, the duration data was normalized via the formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Semitones were calculated as a measurement of participants' singing accuracy based on the pitch of each token, using the formula below:

$$Semitone = 12 \times \log_2 \frac{Singing\ Pitch}{Note\ Pitch}$$

For those male speakers who sang in an octave lower (N = 5), semitones were calculated by doubling their singing pitch.

Linear mixed-effect models were built for duration data between sessions and in each session, using the *lme4* package [19] in R. In the model for the singing session, the fixed factors were Tone (T1, T2, T3 and T4) and Note (G3, A3 and B3). In the model for all the data across sessions, the fixed factors were Session (normal speech and singing) and Tone (T1, T2, T3 and T4). Subject and Vowel were set as random factors in both models. For example, the model for all duration data is listed below:

$$NomDuration \sim Session * Tone + (1|Subject) + (1|Vowel)$$

In order to prevent the influence of factors which caused unbalanced data distribution or irrelevant to the question of interest (as shown in Table 2), speaker gender, music experience and accuracy of singing were separately modeled as fixed factors in three linear mixed-effect models prior to our formal test, and the results presented no main effects of these factors.

Table 2: *Subjects' information and singing accuracy calculated by semitone*

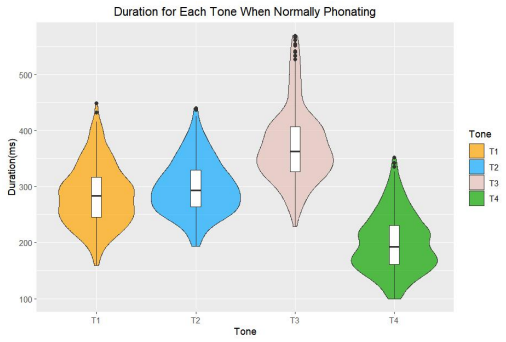
Subject	Gender	Music Experience	Mean accuracy of singing under each musical note (semitone)			
			G3	A3	B3	Mean
GMY	F	No	0.62	0.62	-0.18	0.35
GYD	F	Yes	1.49	1.56	1.15	1.40
GLS	M	Yes	-4.39	-4.64	-4.57	-4.54
LHY	M	Yes	-4.51	-5.64	-6.90	-5.69
LKX	F	No	-1.28	-0.20	-0.19	-0.56
LMG	F	No	1.58	1.46	1.16	1.40
LYC	F	No	0.38	0.81	-1.30	-0.04
MYS	M	Yes	-0.97	0.08	-0.04	-0.31
NJY	F	Yes	5.68	4.50	7.14	5.81
RTQ	M	Yes	1.57	1.08	0.29	0.99
ST	F	Yes	-1.28	-0.02	-0.06	-0.46
TR	M	No	-3.12	-3.93	-4.45	-3.83
XRX	F	No	3.51	2.08	1.33	2.31
YJT	M	No	1.16	0.88	0.99	1.01
ZTY	M	No	-1.65	0.18	-0.65	-0.68
ZZX	M	No	2.43	1.79	1.20	1.81

3. Results

3.1. Durations of four tones across two sessions

As shown in Table 3 and Figure 1, in the normal speech session, T3 had the longest duration and T4 had the shortest. In the singing session, T4 was still the shortest but T1 was the longest. In addition, the mean durations of the four tones in the singing session were significantly longer than in the normal speech session (for all the tones, $p < .0001$) and the differences between the four tones decreased.

(a)



(b)

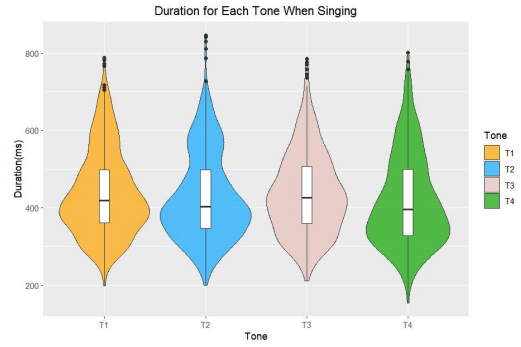


Figure 1: *Violin plots of tone durations in normal speech (a) and singing speech (b)*

Table 3: *Mean durations of four tones in two sessions*

Scene	Mean Vowel Duration（ms）			
	T1	T2	T3	T4
Normal phonation	284.18	298.77	372.40	199.45
Singing	437.12	429.42	436.91	418.58

Table 4: *Results of linear mixed-effects on tone durations in two sessions*

	Sum Sq	Mean Sq	Num DF	Den DF	F value	*p*
Session	33.662	33.662	1	3420.2	2227.06	<.0001
Tone	12.797	4.266	3	3419.0	282.21	<.0001
Scene: Tone	6.810	2.270	3	3419.0	150.19	<.0001

The linear mixed-effects model (as shown in Table 4) revealed a significant main effect of Session (F (1, 3420.2) = 2227.06, $p < .0001$) and a significant main effect of Tone (F (3, 3419.0) = 282.21, $p < .0001$), as well as a significant two-way interaction between Session and Tone (F (3, 3419.0) = 150.19, $p < .0001$).

The Tukey-HSD post-hoc comparison for the two-way interaction effect showed that the pattern of tone-difference significance varied between the normal speech and the singing session. In the normal speech session, T1 and T2 had no significant differences in duration, while T4 was significantly shorter and T3 significantly longer than all the other tones. In the singing session, T4 was still shorter than the other tones, and T3 significantly longer than T2 and T4, but T3 and T1 were not significantly different in duration (see Table 5).

Table 5: *Post-hoc comparison of duration between tones in two sessions*

Scene	contrast	estimate	SE	z.ratio	*p*.value
Normal	T1 - T2	-0.02504	0.01072	-2.337	0.090
	T1 - T3	-0.16433	0.01065	-15.424	<.0001
	T1 - T4	0.15605	0.01069	14.603	<.0001
	T2 - T3	-0.13929	0.01062	-13.113	<.0001
	T2 - T4	0.18109	0.01065	16.996	<.0001
	T3 - T4	0.32038	0.01059	30.245	<.0001
Sing	T1 - T2	0.01916	0.00706	2.715	0.034
	T1 - T3	-0.00429	0.00706	-0.607	0.930
	T1 - T4	0.04734	0.00710	6.665	<.0001
	T2 - T3	-0.02345	0.00701	-3.345	0.005
	T2 - T4	0.02817	0.00705	3.996	<0.001
	T3 - T4	0.05162	0.00705	7.317	<.0001

3.2. Durations of four tones in singing context

As shown in Table 6, the linear mixed-effects model revealed a significant main effect of Tone (F (3, 2365.1) = 21.811, $p < .0001$) and a significant two-way interaction between Tone and Note (F (6, 2365.0) = 4.590, $p < 0.001$).

Table 6: *Results of linear mixed-effects on tone duration in singing context*

	Sum Sq	Mean Sq	Num DF	Den DF	F value	*p*
Tone	0.978	0.326	3	2365.1	21.811	< .0001
Note	0.021	0.010	2	2365.0	0.696	0.499
Tone: Note	0.412	0.069	6	2365.0	4.590	< .001

Post-hoc comparisons revealed that among all the musical notes, the consistent result was that T4 was significantly shorter than T1 and T3 (see Table 7). T4 was significantly shorter than T2 only under note G3, where T3 was significantly longer than all the other tones.

Table 7: *Post-hoc comparison of tone durations under three musical notes (only significant contrasts are listed)*

Note	contrast	estimate	SE	t.ratio	*p*.value
G3	T1 - T3	-0.040	0.0123	-3.239	0.007
	T1 - T4	0.032	0.0124	2.589	0.047
	T2 - T3	-0.043	0.0122	-3.531	0.002
	T3 - T4	0.072	0.0123	5.873	<.0001
A3	T1 - T3	0.038	0.0122	3.096	0.011
	T1 - T4	0.071	0.0123	5.764	<.0001
	T2 - T4	0.050	0.0123	4.054	<0.001
	T3 - T4	0.033	0.0123	2.689	0.036
B3	T1 - T4	0.036	0.0123	2.919	0.019
	T2 - T3	-0.042	0.0121	-3.439	0.003
	T3 - T4	0.052	0.0122	4.249	<0.001

4. Discussion

In the current study, we explored whether speakers maintain the secondary duration cues for Mandarin tones when F0 is restrained to carry musical notes in the singing context. We conducted two comparisons of tone duration, one across the normal speech and the singing session, and the other across musical notes within the singing session.

In the analysis across sessions, the post-hoc comparison for Tone × Session interactions revealed that all the four tones are significantly longer in singing context than in normal speech. This result is consistent with [20-21] who directly compared the acoustic parameters between speech and singing and found mean syllable durations to be higher in singing than in speaking. The longer syllable duration in the singing context is reasonable, since lungs and the larynx need more time to arrive at a given pitch and resonance [22]. However, inconsistent with previous studies focused on tones in whisper [12-13] which suggested the enhancement of duration contrasts, the present study revealed a reduced duration variation between tones in the singing context.

In addition, the post-hoc comparison for Session × Tone interaction also indicate different duration patterns of the four Mandarin tones between singing and speaking. In the normal speech session, syllables with T4 were significantly shorter than syllables with the other three tones, and T3 is significantly longer than the others, consistent with previous studies [4-5]. In the singing session, as in whisper contexts [11-13], T4 is still the shortest tone, and T3 is significantly longer than T2 and T1, suggesting a solid tone duration pattern in different communication contexts. However, no significant difference in duration has been found between T1 and T3, the reason for which is still unknown. But it is speculated that T1, the only level tone in Mandarin Chinese, may have some superposition effect on duration with a single note which sounds like another level tone.

In the analysis within the music context, the linear mixed-effects model revealed a main effect of Tone and an interaction between Tone and Note, but no main effect of Note itself. Further studies are required to unmask the reason for the different duration patterns of tones under different notes, including more notes with greater pitch differences as predictors.

5. Conclusions

The present study demonstrated similar but unidentical patterns in the durations of the four Mandarin tones between the normal speech and the singing context, the latter placing conflicting demands on pitch realization. Specifically, 1) the mean tone duration in the singing context is longer than in normal speech; 2) the variance of tones decreases in the singing context; 3) T1 has a relatively increased duration in the singing context. However, in both contexts, T4 is consistently the shortest, and T3 is the longest if the exceptive singing T1 is not taken into consideration. Moreover, although musical notes have no significant main effect, they do to some extent affect the pattern of tonal duration, and further studies are required to unmask the inner reason. In sum, the similarity in tone duration pattern between normal speech and singing indicates that the pattern has been internalized in speakers, no matter what the prescribed pitch contour and the communication context is.

6. References

[1] J. Gandour, "Tone perception in Far Eastern languages," *Journal of Phonetics*, vol. 11, no. 2, pp. 149-175, 1983.

[2] J. M. Howie, "Acoustical studies of Mandarin vowels and tones," *Cambridge University Press*, 1976.

[3] R. H. Wang, "Chinese phonetics," *Speech Signal Processing*, pp. 37-64, 1989.

[4] A. T. Ho, "The acoustic variation of Mandarin tones," *Phonetica,* vol. 33, no. 5, pp. 353-367, 1976.

[5] D. H. Whalen and Y. Xu, "Information for Mandarin tones in the amplitude contour and in brief segments," *Phonetica,* vol. 49, no. 1, pp. 25-47, 1992.

[6] L. H. Wee, "Phonological tone," *Cambridge University Press,* 2019.

[7] M. Lin and J. Yan, "A perceptual study on the domain of tones in Standard Chinese," *Chinese Journal of Acoustics,* vol. 14, no. 4, pp. 350-357, 1995.

[8] J. Yang, Y. Zhang, A. Li, and L. Xu, "On the Duration of Mandarin Tones," *in INTERSPEECH*, pp. 1407-1411, 2017.

[9] W. F. Heeren and C. Lorenzi, "Perception of prosody in normal and whispered French," *The Journal of the Acoustical Society of America*, vol. 135, no. 4, pp. 2026-2040, 2014.

[10] M. Higashikawa and F. D. Minifie, "Acoustical-perceptual correlates of 'whisper pitch' in synthetically generated vowels," *Journal of Speech, Language, and Hearing Research,* vol. 42, no. 3, pp. 583-591, 1999.

[11] L. L. Holt and A. J. Lotto, "Cue weighting in auditory categorization: Implications for first and second language acquisition," *The Journal of the Acoustical Society of America,* vol. 119, no. 5, pp. 3059-3071, 2006.

[12] F. Llanos, O. Dmitrieva, A. Shultz, and A. L. Francis, "Auditory enhancement and second language experience in Spanish and English weighting of secondary voicing cues," *The Journal of the Acoustical Society of America,* vol. 134, no. 3, pp. 2213-2224, 2013.

[13] M. Gao, "Tones in whispered Chinese: articulatory features and perceptual cues," *University of Victoria,* 2002.

[14] H. Zhang, S. Wiener, and L. L. Holt, "Adjustment of cue weighting in speech by speakers and listeners: Evidence from amplitude and duration modifications of Mandarin Chinese tone," *The Journal of the Acoustical Society of America*, vol. 151, no. 2, pp. 992-1005, 2022.

[15] S. Liu and A. G. Samuel, "Perception of Mandarin lexical tones when F0 information is neutralized," *Language and Speech*, vol. 47, no. 2, pp. 109-138, 2004.

[16] J. W. Peirce, "PsychoPy—psychophysics software in Python," *Journal of Neuroscience Methods,* vol. 162, no. 1-2, pp. 8-13, 2007.
[17] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [Computer program](Version 6.1. 24)," Retrieved from www. praat. org, 2021.
[18] Y. Xu, "ProsodyPro—A tool for large-scale systematic prosody analysis," *Laboratoire Parole et Langage,* France, 2013.
[19] D. Bates, M. Mächler, B. Bolker, S. Walker, R. H. Christensen, H. Singmann, and B. Dai, "lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7," 2014.
[20] S. R. Livingstone, K. Peck, and F. A. Russo, "Acoustic differences in the speaking and singing voice," *in Proceedings of Meetings on Acoustics*, vol. 19, no. 1, June 2013.
[21] B. R. de Medeiros and J. P. Cabral, "Acoustic distinctions between speech and singing: Is singing acoustically more stable than speech?," *in International Conference on Speech Prosody*, pp. 542-546, 2018.
[22] D. Burrows, "Singing and Saying," *The Journal of Musicology,* vol. 7, no. 3, pp. 390–402, 1989.