INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

The acoustic realization of Mandarin tones in fast speech

Ping Tang, Shanpeng Li

School of Foreign Studies, Nanjing University of Science and Technology, China
ping.tang@njust.edu.cn, lspqdu@foxmail.com

Abstract

Many studies have demonstrated that acoustic contrasts between speech segments (vowels and consonants) were reduced when speaking rate increases, while it was unclear whether tones in tonal languages also undergo similar modifications. Mandarin Chinese is a tonal language, while results regarding the rate effect on Mandarin tones in previous studies were mixed, probably driven by the material difference, i.e., the position of target tones within a sentence. Therefore, the present study examined the effect of speaking rate on Mandarin tones, comparing the pitch contour and tonal contrast of Mandarin tones between normal and fast speech across utterance initial, medial and final positions. The results showed that, relative to normal speech, lexical tones in Mandarin Chinese exhibited overall higher and flatter pitch contours, with smaller tonal space. Moreover, the rate effect on tones did not vary with position. The current results and previous studies on segments thus revealed a universal pattern of speech reduction in fast speech at both segmental and suprasegmental levels.

Index Terms: Mandarin tones, fast speech, acoustic realization, tone space

1. Introduction

Speech units such as vowels or consonants are contrasted in various acoustic parameters such as the first and second formants (F1 and F2) or voicing onset time (VOT). It has been suggested that, in fast speech, acoustic distinctions between vowels or consonants typically reduced, resulting in the target-undershoot phenomena [1]. However, most languages (more than 60%) on the world are tonal, using lexical tones in additional of segments to contrast word meanings [2], while only a few studies have explored the effect of speaking rate on tonal realization, and the results are mixed [3-5]. It was unclear whether tones also undergo tonal target-undershot in fast speech. Therefore, the current study focused on a tonal language Mandarin Chinese and explored the acoustic realization of Mandarin tones in fast speech.

Mandarin has four lexical tones with distinct pitch contours, i.e., level (Tone 1, hereafter "T1"), rising (T2), dipping (T3) and falling (T4; see figure 1). The four tones are primarily contrasted in contour features, i.e., pitch height (T1: high), pitch slope (T2: rising; T4: falling) and curvature (T3: dipping). Lexical tones are important to delivery meanings, i.e., "ma1" *mother*, "ma2" *hemp*, "ma3" *horse* and "ma4" *scold*.

There have been several studies investigating the effect of speaking rate on the acoustic realization of Mandarin tones, with mixed results [3-5]. For instance, in a pioneering work, [3] analysed the pitch realization of Mandarin tones produced at slow, normal and fast speaking rates. Based on observation, the author argued that there were no systematically tonal changes in terms of pitch height or slope as a function of speaking rate. However, the author did not provide any statistical evidence to support this finding. Moreover, since there were only four speakers in that study, a large individual variability was also observed in the results. Despite this, a later work from [5] partly supported his findings. The authors analysed the pitch realization of Mandarin T2 and T3 produced at various speaking rates, finding that speaking rate did not change the pitch realization of the T2 and T3. On the contrary, [4] reported that speaking rate does modify the realization of Mandarin tones: the onset and offset of T1 and T4 were higher in fast speech, resulting in increased pitch heights for the two tones; the turning points of T2 and T3 moved forward in fast speech, leading to modifications on the shape of pitch contours.

A potential reason leading to the discrepancy might be the material difference, especially the position of target tones within utterance. In the studies of [3, 5], the target tone was embedded in the utterance-medial or utterance-final position, while it was in isolation in the study of [4]. As argued by [7], the modification of speaking rate on tones might be reduced as the position of target tones move towards the end of a sentence due to the declination effect. Therefore, tones may undergo smaller modifications in studies of [3, 5] as compared to those in the study of [4], thus leading to different results.

Despite the mixed results, these studies only focused on the rate effect on the realization of individual tones rather than the acoustic contrasts between tones. It was unclear whether tonal contrasts are also reduced in fast speech like those of segments. This issue could be addressed by comparing the size of tone space between fast and normal speech. The tone space maps three simple-contour tones (T1, T2 and T4) into a two-dimensional quadrilateral using two contrastive features: pitch onset and offset [8]. Similar to the vowel space, a smaller tone space suggests reduced tonal distinctions. The tone space has been proved to be a valid and effective tool to quantify the degree of tonal distinctions across different tonal languages, including Cantonese [8-10], Thai [9] and Mandarin Chinese [9, 11-12].

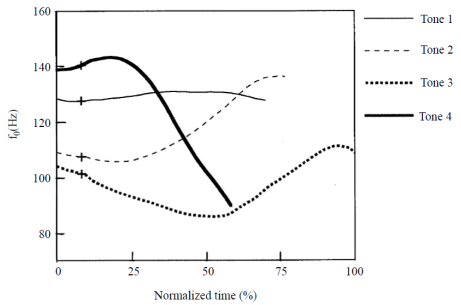

Figure 1: *Pitch contour of Mandarin tones, from [6].*

Therefore, the current study compared the tonal realization and the degree of tonal distinction (as measured by the size of

http://dx.doi.org/10.21437/Interspeech.2020-1274

tonal space) in fast speech across various positions (utterance initial, medial and final). We asked whether and how Mandarin tones change when speaking rate increases, and the consequence of this change on tonal contrasts.

2. Method

2.1. Participants

Fourteen Mandarin-speaking adults participated in the current study (seven males and seven females; age: Mean = 26.96 years, SD = 2.35 years). All participants provided written informed consents to participate and were reasonably remunerated for their participation.

2.2. Procedure

Stimuli included three target syllables /ta/, /ti/ and /tu/ with four lexical tones (T1-4), embedded in the initial, medial and final positions of utterances. Three carrier sentences were used, i.e., "_ zhe4 ge0 zi4 wo3 hui4 fa1" _ *word I know how to say* (utterance-initial; the tone "0" represents the neutral tone in Mandarin Chinese), "wo3 fa1 _ gei3 ni3 ting1" *I say _ for you* (utterance-medial) and "zhe4 ge0 zi4 ying1 gai1 fa1 _" *This word should be pronounced as* _ (utterance-final). In the utterance-medial and utterance-final positions, the target syllables were always preceded by a syllable carrying T1 to minimize the coarticulation effect. Each sentence was produced in normal (as control) and fast speech conditions in random order across participants (see below section for detailed procedure to elicit different speaking rates). Therefore, there were 72 tokens for each speaker (3 syllables × 4 tones × 3 positions × 2 rates).

2.3. Procedure

Each participant took part in the experiment individually in a sound-proof room. Target sentences were printed on A4 papers with randomized order across participants. Prior to experiments, participants were allowed to familiarize themselves with the target sentences. In the testing phase, two blocks were adopted to elicit normal and fast speech in random order. In the block for normal speech, for instance, participants were asked to produce the target sentences at a normal, comfortable speaking rate. In the block for fast speech, participants were instructed to produce the target sentences as fast as possible without forsaking accuracy or naturalness. Participants were able to rest between blocks as long as they want. They were also instructed to break for 1 or 2 seconds for breathing between sentences. Recordings were made via a Neumann U87Ai cardioid condenser microphone and a Fireface 800 audio interface with a sample rate of 44100 Hz.

2.4. Coding and measurement

Data were acoustically coded using PRAAT [13]. A total of 1292 tokens were coded and included in further analysis, with four tokens being excluded due to the poor acoustic quality. Vowel onset and offset were marked for each target syllable based on clear F2 energy in the spectrogram. For each target syllable, ten equidistant pitch points were extracted from 5% to 95% of the vowel portion. Pitch points were manually checked and corrected to avoid the "doubling" and "halving" errors in pitch tracking. The observed pitch values (in Hz) were transformed into semitones (St = 12 × $\log_2$ (Observed Hz / 50Hz)) and z-score normalized across participants.

The tone space was calculated based on the pitch onset and offset values of three simple-contour tones, i.e., T1, T2 and T4. The area of tone space was calculated using the same method used in [11].

The growth curve analysis (GCA) was adopted to compare the pitch contours of Mandarin tones across different conditions using three pitch parameters, i.e., height, slope and curvature [14], implemented by the linear mixed regression model using the R package "lme4" [15]. Linear mixed-effects models were adopted to compare the tone space areas across conditions. In both types of models, the significance of the fixed effects was obtained using the anova function in the R package "lmerTest" [16], which provides omnibus effects for multilevel factors or interactions using F-tests rather than comparisons with the baseline level using t- or z-tests. When a significant main effect of a multilevel factor or a significant interaction effect was observed, Tukey HSD post-hoc comparisons were performed on the multilevel factor, as well as interactions, using the R package "lsmeans" [17].

3. Results

3.1. Speaking rate

Averaged vowel duration of target syllables was 0.23s (SD = 0.08s) and 0.15s (seconds; SD = 0.05s) for normal and fast conditions across participants. A linear mixed-effects model was performed on the duration of target syllable. Two fixed factors "Rate" (normal and fast) and "Position" (initial, medial and final) and two random factors "Participant" (14 participants) and "Syllable" (/ta/, /ti/ and /tu/) were included in the model.

The results showed that the effects of "Rate" and "Position" were both significant on the vowel duration (Rate: $F(1, 986) = 851$, $p < 0.001$); Position: $F(2, 986) = 128$, $p < 0.001$), while the interaction of "Rate × Position" was not significant: $F(2, 986) = 0.174$, $p = 0.840$. The Tukey-HSD post-hoc test on the main effect of "Rate" showed that the vowel duration in fast speech was significantly shorter than that in normal speech ($\beta = -0.093$, $t(984) = -29.177$, $p < 0.001$).

3.2. Pitch contour

Figure 2 shows the pitch contour of lexical tones in normal and fast speech. A linear mixed regression model was conducted on pitch parameters (pitch height, slope and curvature), with three fixed factors "Tone" (T1-4), "Rate" (normal and fast) and "Position" (initial, medial and final) and two random factors "Participant" and "Syllable". The results showed that the interaction of "Rate × Tone" was significant on pitch height ($F(3, 9663) = 4.170$, $p = 0.006$) and pitch slope ($F(3, 9963) = 12.572$, $p < 0.001$), and the interaction of "Rate × Tone × Position" was significant on the pitch curvature ($F(6, 9963) = 2.396$, $p = 0.026$).

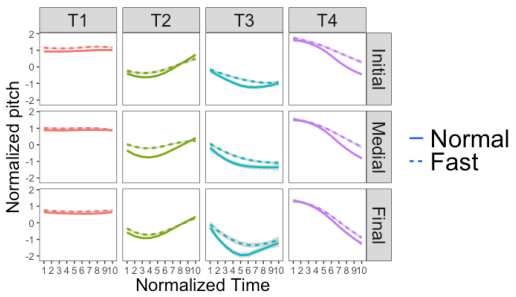

Figure 2: *Pitch contour of lexical tones in normal and fast speech across positions.*

To further explore the effect of speaking rate on tonal realizations, Tukey HSD post-hoc tests were then performed on the two-way interaction (on pitch height and pitch slope) and the three-way interaction (on pitch curvature) to compare the pitch difference between fast and normal conditions across tones and positions. The results showed that, in the fast condition, all tones exhibited higher (larger pitch height values) and flatter (smaller absolute pitch slope and curvature values) contours across all positions, except for T1, i.e., the level tone, which maintains the same level contour in both conditions.

3.3. Tone space

Figures 3 and 4 illustrates the acoustic tone spaces and the tone space areas of normal and fast speaking rates across positions. A linear mixed-effects model was performed on the tone space area with two fixed factors "Rate" (normal and fast) and "Position" (initial, medial and final) and a random factor "Participant".

The results showed that only the main effect of "Rate" was significant on the space area (F (1, 230) = 36.547, $p < 0.001$), while the main effect of "Position" (F (2, 230) = 1.196, p = 0.304) or the interaction of "Rate × Position" (F (2, 230) = 2.561, p = 0.080) was not significant. The Tukey-HSD post-hoc test on the main effect of "Rate" showed that, relative to normal speech, the tone space area was significantly smaller in the fast speech ($\beta = 0.327$, t (230) = 6.045, $p < 0.001$).

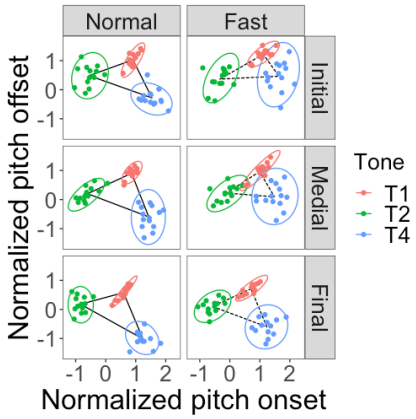

Figure 3: *Acoustic tone spaces in normal and fast speech across conditions.*

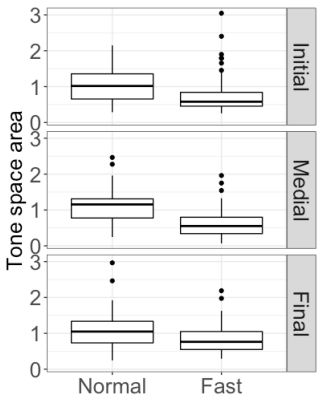

Figure 4: *Tone space areas in normal and fast speech across conditions.*

4. Discussion

The current study examined the acoustic realization of Mandarin tones in fast speech across utterance initial, medial and final positions. The results showed that, relative to normal speech, tones in fast speech exhibited higher and flatter pitch contours (reflected by higher pitch height and smaller pitch slope and curvature). The tone space area was also smaller in fast speech, suggesting reduced tonal contrasts. Moreover, the effect of speaking rate on tones did not vary with position.

The current results therefore supported previous findings of [4] that speaking rate does modify the realization of Mandarin tones, with higher pitch for T1 and T4 in fast speech. Moreover, our results also extended her findings and showed that this is also the case for T2 and T3. The increased pitch height in fast speech were also consistent with findings from previous physiological studies showing that speakers tend to increase fundamental frequency (f0) values as speaking rate increases due to the increased vocal fold tension [18].

However, our results are not consistent with findings from [3, 5], which suggested that Mandarin tones were unlikely to be modified by speaking rate. This inconsistence was unlikely to be driven by the positional effect as we initially proposed, because our results demonstrated that the rate effect on tones did not change with position. We thus propose that the inconsistence might be related to the methodological difference. For instance, conclusions regarding the rate effect in [3] were primarily made by observation rather than acoustic or statistical evidence, and a close inspection on figures in [3] revealed that the tonal realization does differ across speaking rates, while the graphs could not provide any numeric evidence to support this. Furthermore, although [5] were based on acoustic analysis, their pitch analysis focused on T2 and T3 and only measured ΔF0 values (the pitch change from onset to the lowest value of the pitch contour), which primarily reflect tonal changes in the initial part of a pitch contour. In contrast, the present study focused on all four tones and measured global features of a pitch contour, i.e., pitch height, slope and curvature, which can better capture key pitch changes across the entire pitch contour when speaking rate alters.

The reduced tonal contrasts in fast speech, i.e., flatter pitch contours and reduced tone spaces, suggested tonal target-undershoot in fast speech. This is also in line with previous findings in other tonal languages, e.g., Thai [18], which found that pitch contours of Thai tones were flatter in fast speech as compared to those in normal speech, leading to reduced tonal contrasts in fast speech. Our results are also consistent with previous studies finding that segments such as vowels and consonants are target undershoot and acoustically less contrastive in fast speech, as reflected by the reduced vowel spaces and voicing contrasts [18-19]. Taken together, the present and previous studies thus reveal a big picture showing that, when speaking rate increases, speech units exhibit a universal pattern of speech reduction at both segmental (vowels and consonants) and suprasegmental (tones) levels.

There are some limitations of the present study that could be addressed in future research. First, the current study focused on fast speech only, and future work could explore the tonal realization in slow speech, which has not been investigated systematically so far. Secondly, the current study mainly focused on the pitch information of tones, and future studies could further explore the impact of speaking rate on the voice

quality of Mandarin tones, which is an important feature for the perception of T3.

5. Conclusions

The current study showed that Mandarin tones underwent modifications when speaking rate increases, exhibiting overall higher and flatter pitch contours in fast speech, with reduced acoustic pitch contrasts. This is in line with previous findings on segments, exhibiting an universal pattern of speech reduction in fast speech at both segmental and suprasegmental levels.

6. Acknowledgements

This research was funded by the The National Social Science Fund of China (20CYY012).

7. References

[1] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory", Speech Production and Speech Modelling, Springer, Dordrecht, pp. 403-439, 1990.
[2] M. Yip, Tone, Cambridge University Press, 2002.
[3] Y. Xu, "Consistency of tone-syllable alignment across different syllable structures and speaking rates", Phonetica, vol. 55, no. 4, pp. 179-203, 1998.
[4] Y. H. Stockton, "The effects of speaking rate on Mandarin tones", Doctoral dissertation, University of Kansas, 2008.
[5] J. A. Sereno, H. Lee and A. Jongman, "Effects of speaking rate and context on the production of Mandarin tone", In Proceedings of the 18th International Congress of Phonetic Sciences, 2015.
[6] Y. Xu "Contextual tonal variations in Mandarin", Journal of Phonetics, vol. 25, no. 1, pp. 61-83, 1997.
[7] R. Nitisaroj, "Thai tonal contrast under changes in speech rate and stress", In Speech Prosody, pp. 2-5, 2006.
[8] J. G. Barry and P. J. Blamey, "The acoustic analysis of tone differentiation as a means for assessing tone production in speakers of Cantonese", The Journal of the Acoustical Society of America, vol. 116, no. 3, pp. 1739-1748, 2004.
[9] N. Xu Rattanasone, V. Attina, B. Kasisopa and D. Burnham, "How to compare tones", South and Southeast Asian psycholinguistics, pp. 233-246, 2013.
[10] N. Xu Rattanasone, D. Burnham and R. G. Reilly, "Tone and vowel enhancement in Cantonese infant-directed speech at 3, 6, 9, and 12 months of age", Journal of Phonetics, vol. 41, no. 5, pp. 332-343, 2013.
[11] P. Tang, N. Xu Rattanasone, I. Yuen and K. Demuth, "Phonetic enhancement of Mandarin vowels and tones: Infant-directed speech and Lombard speech", The Journal of the Acoustical Society of America, vo. 142, no. 2, pp. 493-503, 2017.
[12] N. Zhou and L. Xu, "Development and evaluation of methods for assessing tone production skills in Mandarin-speaking children with cochlear implants", The Journal of the Acoustical Society of America, vo.l. 123, no. 3, pp. 1653-1664, 2008.
[13] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer" (Version 6.0. 23) [Computer software], Amsterdam, The Netherlands: Authors, 2016.
[14] D. Mirman, Growth Curve Analysis and Visualization Using R, CRC press, 2016.
[15] D. Bates, M. Maechler, B. Bolker and S. Walker, lme4: Linear mixed-effects models using Eigen and S4, R package version 1.1–7, 2015.
[16] T. A. Kuznetsova, P. B. Brockhoff and R. H. B. Christensen, lmerTest: Tests for Random and Fixed Effects for Linear Mixed Effect Models (lmer objects of lme4 package), version 2.0-6, 2014.
[17] R. V. Lenth, "Least-squares means: the R package lsmeans", Journal of Statistical Software, vol 69, no. 1, pp. 1-33, 2016.
[18] W. E. Cooper and J. M. Sorensen, Fundamental frequency in sentence production. Springer Science & Business Media, 2012.
[19] J. L. Miller, "Effects of speaking rate on segmental distinctions", Perspectives on the Study of Speech, pp. 39-74, 1981.
[20] M. J. Solé, "Controlled and mechanical properties in speech", Experimental Approaches to Phonology, pp, 302-321, 2007.