A Perceptual Study on Circumflex Tone in Mandarin*

GENG Puyang, FAN Ningxue
Nanjing Normal University, Nanjing, China

The present study investigated the perceptual characteristics of the circumflex Tone 3 (T3) in Mandarin, by conducting a perceptual experiment on two sets of stimuli synthesized from the time-normalized base syllables /i1/ and /i3/. The rate of identification for T3 was significantly higher on the set of stimuli derived from /i3/ than on that derived from /i1/. Moreover, acoustic analysis of the stimuli revealed that not only $F_0$ but also intensity and voice quality differed between the two sets, suggesting that intensity and voice quality could be secondary cues for T3. In sum, we propose that T3 of Mandarin be preferably defined on the basis of both voice source and vocal tract characteristics, instead of only on $F_0$ variation.

Keywords: perception, Tone 3, $F_0$ manipulation, intensity, voice quality

Introduction

As one of the most well-known tonal languages among the world, Mandarin has four lexical tones: Tone 1 (level-tone), Tone 2 (rising-tone), Tone 3 (circumflex-tone), and Tone 4 (falling-tone). These tones carry not only much linguistic information, but also intonation implication as a salient factor. Among all these tones, Tone 3 (henceforth T3) is mostly disputable in terms of the perspectives of its tonal contour. The early studies proposed that T3 should be best described as a raising contour (Liu, 1951; Karlgren, 1930), while some other scholars suggested that T3 should be best described as a low-falling/low-level tone, which is also known as /21/ or /11/ in five-tone-letter system (Chao, 1922; Lin, 1965; Shi & Ran, 2011; Chao, 1930). Another viewpoint held that T3 had a low-concave curve (Lin & Wang, 1922; Cao & Wei, 2016) and it was the ideal manifestation for T3. Voice quality of T3 had also been emphasized because that fundamental frequency ($F_0$) was not the only variation mattered to the realization of T3. It had been noted that some more acoustic correlates (e.g., creaky voice, amplitude) contributed to the production of T3 (Chao, 1980; Kong, 2001) as well. However, apart from $F_0$, none of the above researches took these parameters as indispensable factors but appendant ones in T3 production.

Experimentally, numerous perceptual researches had been performed in the field and intensified the debate. Through the observation of digital manipulated $F_0$ contours of T3, some had reported that an ideal shape of T3 contour should be a low-concave curve (Kong, 2001). Results from tonal perception for foreign learners of Mandarin implied that judgments of T3 showed different patterns between Thailand students and native Mandarin listeners (Wu, 2014). Another perceptual study carried out by Chen and Liu (2010) on cochlea-implanted children found that duration and intensity largely accounted for the tonal recognition when

* The study is funded, in part, by “China Scholarship Council” scholarship to the first author.

GENG Puyang, Ph.D. candidate, School of Chinese Language and Literature, Nanjing Normal University, Nanjing, China; visiting research student, UC Berkeley, Berkeley, U.S.

FAN Ningxue, Ph.D. candidate, School of Education Science, Nanjing Normal University, Nanjing, China.
This paper aims to uncover a range of variations that contribute to T3 perception instead of merely focusing on relations between tonal recognition and $F_0$ as did in previous studies. This would be a reference for tonal study of Mandarin, and also help teachers and foreign students with Mandarin tone in the process of teaching and learning Chinese as a foreign language.

**Methods**

**Materials**

The original speech materials were downloaded from the speech corpora of “Center Studies of Chinese as Second Language” (CSCSL, 2019-07-20) of Beijing Language and Culture University. Materials pronounced by speaker CM 3 with /i/ in T1 (/i1/) and T3 (/i3/) were selected for digital manipulations. It was reported that T1 could be considered as a stable tone with no marks, which also represented its steady pitch and intensity (Cao & Wei, 2016). By comparing synthesized /i1/ and /i3/, some more acoustic correlates except $F_0$ may be further investigated. Furthermore, the duration of /i3/ was 413 ms, which was in line with the standard timing normalization. Another note is that, /i/’s in T1 to T4 were used to calculate the range of the speaker’s pitch.

$F_0$ values of the four syllables were extracted at 10 ms intervals using autocorrelation analysis in Praat (Boersma & Praat, 2018) and the time-normalized $F_0$ contours of the four were obtained by extracting $F_0$ values at 10 equally-spaced points in these syllables. As shown in Figure 1, four tonal contours were plotted across -2 st to 11 st (measured with reference to 100 Hz). Notably, no particular voice qualities (e.g., creaky voice, breathy voice) were found in the original two basic syllables before manipulation.

![Figure 1. $F_0$ contours of /i/.](image)

**$F_0$ Manipulation**

Two sets of stimuli for perceptual experiment were synthesized on the two basic syllables “/i1/ and /i3/” (normalized to 413 ms) using PSOLA in Praat (Boersma & Praat, 2018). As shown in Figure 2, taking the time-normalized $F_0$ contours of the T1 and T3 syllables respectively, as the -1st and 9th steps for the offset $F_0$, a 10-step rising $F_0$ continuum was constructed by successive linear interpolation on the three feature points (onset, turning, offset), with a same 3 st $F_0$ at the onset point, a 20% duration stepwise percentage difference and same -1 st $F_0$ at the turning point, lastly a 1 st stepwise $F_0$ difference at the offset point. Only one acoustic parameter (here, $F_0$) is typically varied along the two sets of continuum, while all other acoustic parameters are kept intact. The basic syllables (/i1/, /i3/) were manipulated into 40 stimuli (4 * 10) each, thus 80 stimuli in total.
Participants

The perceptual experiment was conducted in a quiet room. All 80 (2 * 4 * 10) word stimuli and 10 original materials of /i/ in T2 and T4 were mixed into the stimuli in order to avoid participants’ habitual selecting. All stimuli were divided into two blocks (45 in each) and presented in random order. The stimuli were presented through headphone and responses were sent through key-pressing. The participants were required to decide the sounds presented belonged to which tone then press the corresponding keys, which is known as identification task. All participants (20) were postgraduates and were recruited in Beijing Language and Culture University.

Results

Identification

The identification experiment involved selecting one out of four candidate tones. Because the sum of the identification rates for the four candidate tones equals one, only the identification rate for one tone (say, T3) needs to be calculated. The rate of a stimulus being identified as T3 is defined.

Rates of tone identification were shown in Figure 3. The identification rates of synthesized stimuli based on /i1/ (mean 11%) were far less than the other set of stimuli (mean 58%), which implied that the manifestation of T3 might consist of some other acoustic correlates as a whole rather than F0 alone. And probably, the results also reveal some certain distinguishable phonation features behind /i1/ and /i3/ since the F0 had been manipulated to the same semitone. In this section, we plan to focus on the stimuli that originated from /i3/, and the acoustic differences excluding F0 between /i1/ and /i3/ will be discussed in next section.
Figure 3. Identification rates of stimuli.

For the stimuli that synthesized on base of /i3/, different identification patterns were found depending on the position of turning point of the tonal contour. A significant rising of the T3 recognition occurred when the offset $F_0$ was set 4 st or below, as the dipping part accounts for 20% or 40% length of the duration of one syllable. The identification rates dramatically dropped under conditions that the offset $F_0$ was set 2 st or below and the dipping part accounts for a proportion of 60% or 80% of syllabic duration. It is supposed to be an inference that there were correlations between the values of onset $F_0$ and offset $F_0$ which would affect the T3 identification since the $F_0$ of onset point was 3 st laying between 2 st and 4 st as we mentioned above.

To further demonstrate the reasoning that the relationship between values of onset $F_0$ and offset $F_0$, all tonal contours over 50% of identification were plotted in Figure 4 with a horizontal dotted line pointed to 3 st. As we can see in the upper layer of the Figure 4, the offset $F_0$ of the tonal contours should be lower than 3 st if to be perceived as T3, provided that the dipping part accounted for less than 40% length of the syllabic duration. On the contrary, when the turning point was located at the three fifth of the entire syllabic duration, all the contours would be perceived as T3. The contours were perceived as T3 with the offset $F_0$ at 4 st or higher, when turning point was set at 80% of the syllabic duration.
As shown in Figure 5, the general intensity contours of the two basic syllables /i1/ and /i3/ were plotted, and the mean intensities of these two were 52.64 dB and 56.28 dB respectively. Besides, no evidence of previous studies had suggested that intensity would display differently as the change of $F_0$. The intensity curve of /i3/ was shaped as double hump which resulted from its unique $F_0$ contour of T3, while the other curve of /i1/ presented as a unimodal. The difference in steepness between the two and the less fluctuated intensity of /i1/ verified its featured pitch and intensity stability as mentioned in Section 2.1. Especially, value distinctions in the maximum and the range explained intensity between the two as well. Accordingly, the remarkable differences revealed by the two intensity contours in this study indicated that intensity could be a potential salient factor contributing to the T3 identification.

Voice Quality

Two parameters (jitter and shimmer) were measured for the stimuli that synthesized on base of /i1/ and /i3/ as the index of voice quality, which were listed in Table 1. In case that these parameters would vary along with the $F_0$ manipulation, especially given that jitter was known as the cycle-to-cycle variation of fundamental frequency (Farrús, Hernando, & Ejarque, 2007), independent-sample test was taken for the two sets of stimuli.
derived from the syllable bases /i1/ and /i3/. Significant differences were found in both two sets of stimuli ($p < 0.05$).

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Or. /i1/</th>
<th>Me. /i1/</th>
<th>Or. /i3/</th>
<th>Me. /i3/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jitter</td>
<td>0.0205</td>
<td>0.0212*</td>
<td>0.0143</td>
<td>0.0151*</td>
</tr>
<tr>
<td>Shimmer</td>
<td>0.1018</td>
<td>0.0944</td>
<td>0.1017</td>
<td>0.0936</td>
</tr>
</tbody>
</table>

Notes. “Or.” stands for original material; “Me.” stands for mean value of corresponding parameters of synthesized materials; Where * indicates significant difference between the two sets of stimuli ($p < 0.05$).

Since no evident creaky voice quality or other atypical phonation types (e.g., breathy voice) was produced in /i3/, which means the possibility of influence from accompany features of T3 could be reasonably excluded, the inter-group difference we found in jitter was the only factor responsible for the fundamental frequency variation of the stimuli.

**Discussion and Conclusions**

A perceptual experiment was conducted for a series of stimuli that generated from digital $F_0$ manipulation based on two base syllables /i1/ and /i3/, while other acoustic features remained intact for the stimuli with time normalized to 413 ms.

According to the perceptual results of the experiment, apparent distinctions in identification rates of T3 were respectively found for the two sets of stimuli synthesized on base of /i1/ and /i3/. As to materials originated from /i1/, the rates of T3 recognition were 30% or below. On the contrary, the majority of the tonal contours synthesized from /i3/ were identified as T3 with 50% rates or above.

Furthermore, the acoustic analysis on the two sets of stimuli based on /i1/ and /i3/ was employed to verify the conjecture of T3 identification which was possibly influenced by other phonation features in addition to $F_0$. As a result, the significant differences in intensity and jitter between the two sets of stimuli demonstrated that the production of T3 was comprised of multiple factors rather than $F_0$ alone. To put it another way, such argument about the ideal $F_0$ contour of T3 (Cao & Wei, 2016) or tonal paradigm depending on $F_0$ (Shi, 1991) should be barely proposed. However, no one could deny the significance that fundamental frequency had T3 production since the tonal contours were supposed to accord with basic T3 curves (e.g., circumflex tone, or low-concave tone) to be well-perceived (over 50% rate of identification). Moreover, duration was considered as another salient factor contributing to T3 production and perception (Lin & Wang, 1922; Cao & Wei, 2016), which was not examined in this study since it had been normalized for all the stimuli.

Referring to the basic curves of T3 shown in Figure 4 as well, two situations are brought into view. On the one hand, when the dipping part accounts for less than 40% of syllabic duration, the value of offset $F_0$ is supposed to be equal or lower than that of the onset $F_0$. On the other hand, when more than 60% part of the syllable presents a declining curve, the value of $F_0$ of offset shall be consistent with that of onset point or higher. Therefore, the proportion of the dipping part of certain tonal contour is not rigid, and only through a combination of position of turning point and offset $F_0$ can we explain the variations of T3 curve and the well-perceived tonal identification. Hence, a T3 contour is supposed to comprised by inferior offset $F_0$ and a turning point locates at preceding part within one syllable, since higher $F_0$ in the endpoint possibly results in a confusion with T2 (a rising-tone), while the other feasible manifestation of T3 contour is composed with longer
duration of decline and superior $F_0$ at the offset point, otherwise lower offset $F_0$ may be hard to differentiate from T4 (a falling-tone).

Apart from fundamental frequency, intensity is deemed as another significant factor that affects T3 identification as well. The intensity differences we find in this study reflect traits in tonal curves of the two basic syllables which shall be considered as one accompanying feature within the T3 production. According to the study of tonal recognition conducted by Chen and Liu (2010) on cochlea-implanted children, duration and intensity account for a major part of the tonal recognition when $F_0$ is no longer perceived by these hearing-impaired children. Hence, the intensity, which appears to be a simultaneous feature during tonal production, has a great impact in identification of T3.

Finally, the study has examined the highly controversial issue and even its opposite views on voice quality differences between the two basic syllables. The interaction between jitter variation and the corresponding $F_0$ manipulation brings about the differences mentioned for each stimulus. Consequently, different tonal productions lead to different jitter outputs along with $F_0$ variations, which reflect different phonation characteristics of vocal tract as well. Therefore, the jitter is supposed to be integrated into the entire tonal production process along with $F_0$, duration and intensity, instead of being isolated.

Further studies are suggested to focus on some other acoustic information of phonation/articulatory tract during the tonal production. All types of tones in Mandarin shall be taken into consideration as one whole unit. In addition, some other techniques, like Electroglottography (EGG) and Electro-Magnetic Articulograph (EMA), shall see their roles in future exploration on more features of tonal phonation.

References