The Role of Creaky Voice in Mandarin Tone 2 and Tone 3 Perception

Rui Cao\textsuperscript{1}, Ratree Wayland\textsuperscript{1}, Edith Kaan\textsuperscript{1}

\textsuperscript{1}Department of Linguistics, University of Florida, Gainesville, FL, U.S.A.  
ruicao@ufl.edu, ratree@ufl.edu, kaan@ufl.edu

Abstract

This study investigated the role of creaky voice in the perception of Mandarin Tone 2 and Tone 3 among native English and native Mandarin listeners. Results showed that the presence of creaky voice affected the perception of Mandarin Tone 2 and Tone 3. Both native and non-native listeners reported hearing these two tones at a later F\textsubscript{0} turning point with the presence of creaky voice.

Index Terms: speech perception, Mandarin tones, creaky voice

1. Introduction

Among the four tones, Tone 2 and Tone 3 are very similar in their pitch contour shapes and thus can be difficult to distinguish in native and non-native perception [1]. A phonation type, creaky voice, has been reported to be associated with the production of Tone 3 [2], but its role in the perception of either tone has not been established. The goal of this study was to investigate the effect of creaky voice in the perception of Mandarin Tones 2 and 3 among native Mandarin listeners and native English listeners who are learning Chinese. Two tasks were administered: the baseline task and the tone identification task.

There have been some studies on the perception of Tone 2 and Tone 3, e.g. [3, 4]. [4] manipulated Tone 3 in two conditions by varying the duration of the dip (the so-called “turning point” [5] in Tone 3 is often seen as a low dip in natural speech) and the timing of the turning point. However, it did not include a condition that the timing of the dip varies, which is closer to the true tone contour in natural speech. Therefore, in the tone identification task of the current study, the perception of Tone 2-Tone 3 on a tone continuum with varied timing of the dip was examined. In addition, the role of creaky voice was explored by adding creakiness to a duplicated set of stimuli to see how it affects perception of Tone 2 and 3.

2. Baseline Task

There are two purposes of the baseline task-- to make sure all the participants have the satisfactory level of accurate Mandarin tone identification, and to divide non-native speakers of Chinese into different proficiency groups.

2.1. Method

2.1.1. Participants

Group 1: Native Chinese Speakers (NC)

There were 30 native speakers of Chinese (16F, 14M, 28.1±5.1 years). Half were born and grew up in Beijing and the other half were from ten other provinces in Mainland China. All of them speak standard Mandarin Chinese and were studying or working at the University of Florida. They had lived in the United States for 2.9±2.4 years. They were all paid for participation.

Group 2: Native English Speakers (L2 learners of Chinese, NE)

There were 40 native speakers of American English (14F, 26M, 20.9±1.4 years) participated in the study. They were all undergraduate students at the University of Florida, late learners of Mandarin Chinese (AoA>16) with the length of study of 1-4 years. Seventeen of them were paid for participation and the rest received extra credit for their Chinese courses.

2.1.2. Stimuli

The stimuli used in the tone baseline task consisted of 40 monosyllabic words of Mandarin Chinese, with 10 Consonant+Vowel combinations produced with all four tones by two native Chinese speakers from Beijing, one female and one male, resulting in 80 target words total. The 40 words were chosen from the first-year Chinese textbook (Integrated Chinese I) used by the Department of Languages, Literatures, and Cultures of the University of Florida, so that all L2 learners of Chinese should have encountered all the words prior to participating in the study. The recording was conducted in the soundproof booth in the Linguistics lab, using a digital recorder (Marantz PMD660) at a sampling rate of 44.1 kHz. All tokens were normalized at 98% peak intensity with the UAB software developed by Steve Smith at the University of Alabama in Birmingham.

2.1.3. Procedure

All experiments were conducted at the linguistics laboratory of the University of Florida. The tone baseline task was carried out in a quiet room located in the Linguistics lab, through a computer. The presentation of the stimuli in the tone baseline task was controlled by the UAB software. The 80 stimuli were presented in two blocks, one for the 40 words read by the female, and the other for the 40 words read by the male. The 40 target words in each block were randomized.

Presentation of the two blocks of words was counter-balanced for all participants. There was a break after the first 40 words. Participants were tested individually. After a participant entered the room and sat in front of the computer, s/he was instructed (in her/his native language) to identify the tone by clicking one of the four buttons marked with “Tone 1, Tone 2, Tone 3 and Tone 4” on the computer screen after hearing each word through a pair of headphones connected to the computer.

There was no time constraint in this task; that is to say, participants were allowed to take as much time as they needed to make a choice. There was no feedback provided throughout the task, except that a red dot would appear on top of the tone square a participant clicked, as an acknowledgement of a response to a stimulus.
2.2. Results
The baseline score of each participant was the percentage of the total number of correct answers divided by 80. The mean score for the Chinese speakers was 98.9±1.9%. Since there is noticeable variance among the scores of native English speakers (range 53.5%-100%), the English group was further divided into two subgroups: NE-High and NE-Low, with a cut-off point of 90. The NE-High group included 18 participants and their mean score was 94.9±2.9%. The NE-Low group had 22 participants and their mean score was 80.8±10.1%. The post-hoc test from one-way ANOVA showed that the two NE groups (High and Low) are significantly different from each other (p=0.00), whereas the NE-High and NC were not (p=.90). However, for the purpose of the current study, it is necessary to regard the NE-High as a different group from the native Chinese speakers. Therefore, in the following section of perception experiment, participants are considered in one of three proficiency groups: NC, NE-High, and NE-Low.

3. Tone Identification Experiment
There were three goals in the tone identification experiment: 1) to examine the turning point at the categorical boundary between Chinese Tone2-Tone3 by native and non-native (L2) listeners, 2) to investigate whether the presence of creaky voice affects the perceptual boundary, and 3) to explore the role of proficiency of L2 in Tone2-Tone3 perception. Two identical sets of tone stimuli were created for this experiment, except that creaky voice was present in one set, but was absent in the other.

3.1. Method
3.1.1. Participants
They were the same 30 native Chinese speakers and 40 native English speakers from the baseline task. NE-High consisted of 18 and NE-Low 22 participants.

3.1.2. Materials
The first set of tone stimuli consisting of 34 different tokens with no creaky voice was generated from a single token of a [ma] (Tone 3, ‘horse’, /ɔ/) syllable produced by a female native Chinese speaker from Beijing. An examination of the spectrogram generated by PRAAT indicated that this word was produced with no creaky voice. In addition, the total duration of this syllable was 460 milliseconds (ms) long with F0 at syllable onset of 203.8 Hz and offset of 253.5 Hz. The length of the low dip was 68ms and the minimal F0 during the low dip was at 170.5 Hz. All 34 tokens of the non-creaky stimuli were created from this naturally produced model [ma] syllable by manipulating pitch contours using PRAAT.

To generate the 34 tokens of creaky stimuli, another [ma] syllable produced with Tone 3 by the same female speaker was used as the model for examining the natural parameters of creakiness. Examination with PRAAT indicated that this utterance was produced with creakiness at the dip of the Tone 3 pitch contour with duration of 68 ms. Therefore, the dip of the non-creaky tokens in the first group was also set to 68 ms, with the minimal observed F0 of 170.5 Hz from the non-creaky [ma] utterance.

Stimulus Set 1: Non-Creaky Tokens

For the first set of non-creaky tokens, an empty pitch tier (length=460 ms) was created in PRAAT for each token, then 5 pitch points were added to each empty pitch tier (Figure 1), 1) syllable onset (0 ms, F0=203.8 Hz), 2) starting of vowel (57 ms, F0=203.8 Hz), 3) starting point of the dip (starting from 58 ms to 388 ms, in an increment of 10 ms, F0=170.5 Hz), 4) ending point of the dip (starting from 126 ms up to 456 ms, in an increment of 10ms, F0=170.5 Hz), and 5) syllable offset (460 ms, F0=253.5 Hz). Therefore, without manipulating the total length of the syllable, the dip was kept constant at 68ms, only the timing of the turning point (dip) was altered; in other words, the time that the turning point (dip) occurred in each token differed by 10 ms, resulting in 34 different pitch tiers. The modal [ma] token was then synthesized with these 34 different pitch tiers, generating 34 non-creaky tone stimuli.

Stimulus Set 2: Creaky Tokens
The second set of perception stimuli consisted of 34 tokens, which were identical to the 34 tokens in the first set, except that tokens in this group were manipulated again with PRAAT so that each of them had creaky voice at the dip.

Just as in the first set, an empty pitch tier (length=460 ms) was created in PRAAT for each token, and the same five pitch points were added. In addition, the 68 ms dip was manipulated again to create a creaky voice effect. The prominent characteristics of creaky voice include irregular periodicity and sudden decrease in fundamental frequency and intensity. Therefore, creakiness was generated by randomly adding extremely low and irregular pitch points at a level far below the non-creaky dip to create “jitter” (Figure 2). The creakiness sounded quite close to a naturally produced creakiness according to two native Chinese speakers. Thus the values of each of the pitch points in the “jitter” dip were kept the same for all creaky tokens. Each of the pitch tiers was then synthesized with the naturally produced [ma], generating 34 creaky tone tokens for the perception experiment.

Altogether, 68 different tokens were created for the tone categorization experiment. The presentation of these stimuli was again operated with UAB software. They were first normalized at 98% peak intensity with the UAB program. Then the two groups of creaky and non-creaky tokens were mixed together and randomized. The stimuli were presented to each participant in three blocks, with two breaks to reduce the fatigue effect. In each block, the mixed 68 tokens were repeated two times (68*2=136

Figure 1: An example of an empty pitch tier with five pitch points added to create a non-creaky token.

Figure 2: An example of a pitch tier of a creaky token.
the order of the mixed 68 tokens was different in each block. The presentation of the three blocks was counter-balanced across the participants. In total, there were 6 repetitions for each stimulus.

3.1.3. Procedure

The experiment was also conducted in the same lab. Participants were individually tested. Each participant sat in front of a computer, wearing headphones, and was instructed (in their native language) to identify the tone after hearing each stimulus by clicking one of the four squares on the computer screen with texts of “Tone 1, Tone 2, Tone 3, and Tone 4” on each of them. Responses had to be made within 3 seconds; otherwise, the following stimulus would be presented. There was no feedback provided throughout the experiment, except that a red dot would appear on top of the button to acknowledge the response. The participants were also instructed that they should respond according to their first intuition as soon as possible and that they should take two breaks by taking off the earphones and resting for a few minutes.

3.2. Results

3.2.1. Averaged percentages of the four tone responses

The averaged percentages of four tone responses for each stimulus under two conditions were the averaged value of the percentages of each tone response for each stimulus from all participants in a group (example shown in Table 1 and the average percentages are shown in Table 2).

A repeated measures ANOVA with group (3 levels: NC, NE-H, NE-L) as a factor, “Tone response” as the dependent variable, showed no significant main effects of group [F(2, 99)=.463, p=.631] and group x tone response [F(6, 297)=57.4, p<.001] interactions were significant. The stimulus type x tone response x group [F(6, 297)=7.1, p<.001] interactions were significant.

Table 1. Percentages of four tone responses received per stimulus (examples from the first two stimuli among three groups).

<table>
<thead>
<tr>
<th>Group</th>
<th>NC</th>
<th>NE-H</th>
<th>NE-L</th>
<th>Creaky</th>
<th>T1 (%)</th>
<th>T2 (%)</th>
<th>T3 (%)</th>
<th>T4 (%)</th>
<th>T1 (%)</th>
<th>T2 (%)</th>
<th>T3 (%)</th>
<th>T4 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td>47.89%</td>
<td>52.33%</td>
<td>61.31%</td>
<td>60.02%</td>
<td>52.33%</td>
<td>61.31%</td>
<td>60.02%</td>
<td>52.33%</td>
<td>47.89%</td>
<td>52.33%</td>
<td>61.31%</td>
<td>60.02%</td>
</tr>
<tr>
<td>NE-H Low</td>
<td>46.57%</td>
<td>47.89%</td>
<td>52.33%</td>
<td>60.02%</td>
<td>47.89%</td>
<td>52.33%</td>
<td>60.02%</td>
<td>46.57%</td>
<td>47.89%</td>
<td>52.33%</td>
<td>60.02%</td>
<td>46.57%</td>
</tr>
<tr>
<td>NE-H High</td>
<td>47.89%</td>
<td>52.33%</td>
<td>61.31%</td>
<td>60.02%</td>
<td>52.33%</td>
<td>61.31%</td>
<td>60.02%</td>
<td>47.89%</td>
<td>47.89%</td>
<td>52.33%</td>
<td>61.31%</td>
<td>60.02%</td>
</tr>
<tr>
<td>Average</td>
<td>47.89%</td>
<td>52.33%</td>
<td>61.31%</td>
<td>60.02%</td>
<td>52.33%</td>
<td>61.31%</td>
<td>60.02%</td>
<td>47.89%</td>
<td>47.89%</td>
<td>52.33%</td>
<td>61.31%</td>
<td>60.02%</td>
</tr>
</tbody>
</table>

Table 2. Averaged percentages of four tone responses from 3 groups.

Averaged value from all stimuli showed that most of the tone responses are Tone 2 and Tone 3 (Non-creaky 82.65%, Creaky 79.73%). In Non-creaky condition, all three groups heard more Tone 3 (NC 61.31%, NE-H 60.02%, and NE-L 46.57%) than Tone 2 (25.52%, 23.68%, and 30.83%). In Creaky condition, however, only the NC group heard more Tone 3 (43.1%) than Tone 2 (40.92%); the other two NE groups heard fewer Tone 3 (NE-H 29.01%, HE-L 25.91%) than Tone 2 (52.33%, 47.89%).

A repeated measures ANOVA with group (3 levels: NC, NE-High, NE-Low) as the between-subject factor, and stimulus type (2 levels: Non-creaky and Creaky) and tone response (4 levels: Tone 1, Tone 2, Tone 3, and Tone 4) as the within-subject factors yielded a significant main effects of tone response [F(3, 297)=57.4, p<.001], but not of stimulus type [F(1, 99)=.084, p=.773] and group [F(2, 99)=.463, p=.631]. The stimulus type x tone response [F(3, 297)=205.5, p<.001] and stimulus type x tone response x group [F(6, 297)=7.1, p<.001] interactions were significant.

Paired-samples t-tests were run to examine how stimulus type interacts with tone responses. In all three groups, the presence of creakiness resulted in more Tone 1, Tone 2, and Tone 4, but fewer Tone 3 responses. In the NE-Low group, Tone 2 responses were higher in the Creaky condition (47.89%) compared to the Non-creaky condition (30.83%). Tone 3 responses were lower in the Creaky (25.91%) condition compared to the Non-creaky (46.57%) condition. In the NE-High group, the Tone 2 responses were higher in the Creaky condition (52.33%) compared to the Non-creaky condition (23.68%); Tone 3 responses were lower in the Creaky condition (29.01%) compared to the Non-creaky condition (60.02%). Similarly, in the NC group, Tone 2 responses were higher in the Creaky condition (40.92%) compared to the Non-creaky condition (25.52%); Tone 3 responses were lower in the Creaky condition (43.11%) compared to the Non-creaky condition (61.31%). It is very clear to see that the presence of creakiness leads to an increase in Tone 2 responses, and a decrease in Tone 3 responses, regardless of language proficiency.

One-way ANOVA was run for the three groups (with “group” as a factor, “Tone response” as the dependent variable). Under the Creaky condition only, the Tone 3 responses were different among the three groups: NC (43.1%) had more Tone 3 responses than NE-Low (25.9%) and NE-High (29.0%), p=0.04, while between the two NE groups, there was no difference. When looking at the change in Tone 3 responses across conditions, it was found that NC showed the least amount of decrease from Non-creaky (61.3%) to Creaky (43.1%), when compared to both NE-High and NE-Low (60% to 29%, 46.6% to 25.9%). This suggests, perhaps, that the English speakers are more affected by the presence of creaky voice than native Mandarin listeners when it comes to Tone 3 perception.

3.2.2. Average location of turning points for Tone 2 and Tone 3

The average location of turning points (turning point refers to the start of the low dip in the current study) was calculated from stimuli that received 100% Tone 2 or Tone 3 responses. This means that what was included were those stimuli that either a) received six Tone 2 or Tone 3 responses out of 6 repetitions; or b) received five Tone 2 or Tone 3 responses out of five repetitions when there was a missing response. The turning point of across these stimuli was then averaged for each participant.

Table 3 summarizes the results of average location of turning points for Tone 2 and Tone 3 from NC and combined NE-H and NE-L groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>NC</th>
<th>NE-H</th>
<th>NE-L</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATC of Tone 2</td>
<td>73.9</td>
<td>84.1</td>
<td>10.2</td>
<td>191.8</td>
</tr>
<tr>
<td>ATC of Tone 3</td>
<td>94.2</td>
<td>103.2</td>
<td>9</td>
<td>179.1</td>
</tr>
</tbody>
</table>

Table 3. NC and Combined NE of their Tone 2 and Tone 3 ATPs (ms)
The Average Turning Point of Tone 2 Responses:

The NC’s average turning point was 10.2 ms earlier in a Non-creaky (73.9 ms) condition compared to a Creaky condition (84.1 ms, p=.003). Although descriptively, the data seems to hold true for both the NE-High and NE-Low groups, the data, nonetheless, shows no significant difference within each group from Non-creaky (NE-High 90.3ms; NE-Low 97.4 ms) to Creaky (NE-High 101.3 ms; NE-Low 105.5 ms). When these two groups are combined, however, the difference (9 ms) is significant (Non-Creaky ATP for NE: 94.17 ms; Creaky ATP for NE: 103.2; p=0.00). In other words, the NE’s average turning point was 9 ms earlier when there was no creaky voice. Therefore, both native and non-native speakers of Chinese show a later turning point for Tone 2 responses in Creaky condition.

The fact the Tone 2 perception occurs at a later turning point in creaky stimuli suggests that the presence of creaky voice may have perceptually shortened the initial falling portion of the tonal contour before it rises (i.e., the turning point) among native Mandarin listeners. On the other hand, creaky voice does not seem to show the same effects on native English groups.

The Average Turning Point of Tone 3 Responses:

A Paired-sample Test was run for the ATPs within each group and the results were similar to the results from the ATPs of Tone 2 responses. The NC group showed a significant difference in Tone 3 ATPs across conditions—19.6 ms earlier in Non-creaky (191.8ms) than in Creaky (211.4 ms, p=.012). Neither of the NE groups showed a significant difference in their Tone 3 ATPs across conditions (p=.43 for NE-High; p=.39 for NE-Low); however, when these two groups were combined, the difference (14.9 ms) in their Tone 3 ATPs was significant (Non-Creaky ATP for NE: 179.1 ms; Creaky ATP for NE: 194.0 ms; p=0.015). In other words, disregarding tone proficiencies, both native and non-native speakers of Chinese show a later turning point for Tone 3 response in Creaky condition. These results, similar to those from the Tone 2 ATP, again suggest that the perceptual boundary seems to occur at a later point when creaky voice is added.

4. Discussion

Three important results were found from the current study. First, the presence of creaky voice during the turning point of the pitch contour led to a higher percentage of Tone 2 than Tone 3 responses among native Mandarin as well as native English listeners. Secondly, Tone 3 perception occurred at a later turning point with the presence of creaky voice among both native Mandarin and native English listeners (NE-low). Thirdly, in comparison to Mandarin listeners, the amount of Tone 3 responses among native English listeners was affected by the presence of creaky voice to a larger extent.

These results suggested that a) the presence of creaky voice appeared to have perceptually shortened the actual location of the turning point (i.e., the duration of the falling before the dip) among native Mandarin listeners and among native English listeners leading, therefore, to a higher percentages of Tone 2 responses. Specifically, since the turning point of Tone 2 is shorter than that of Tone 3, with the presence of creaky voice, more stimuli with shorter turning points (thus Tone 2 response) were heard among the stimuli presented, and b) creakiness affected Tone 3 perception among native English listeners more than native Mandarin listeners.

5. Conclusions

In summary, fewer Tone 3 was heard with the presence of creakiness suggesting that creakiness is not a direct perceptual cue to Tone 3 identification. Instead, the presence of creakiness perceptually shortens the duration of the turning point leading to more Tone 2 responses among both native English and native Mandarin listeners. On the other hand, the effect of creakiness on Tone 3 responses appeared to be more pronounced among native English than native Mandarin listeners.

6. References