A preliminary study on naïve listeners’ perception of L3 tone and phonation for English learners of Mandarin

Yufei Niu, Peggy Pik Ki Mok

Department of Linguistics and Modern Languages, The Chinese University of Hong Kong

niuyufei@link.cuhk.edu.hk, peggymok@cuhk.edu.hk

Abstract

Studies have shown that learning a tone language facilitates naïve tone perception for native speakers of non-tone languages. However, tone is a multidimensional concept and phonation also plays an important role in tone perception. In this study, 20 native English speakers of different Mandarin levels, 20 native English and native Mandarin speakers (all naïve to the Wenzhou Wu dialect) participated in an AX discrimination task of the Wenzhou dialect because the eight tones in two registers contrast in both pitch and phonation: high register (modal voice) vs. low register (breathy voice). Mandarin T3 often co-occurs with creaky voice while in English, creaky voice is a sociophonetic marker. We asked whether the exposure to one phonation type (creaky voice) at a different linguistic level (indexical in English, allophonic in Mandarin) could lead to an increased sensitivity to another phonation type (breathy voice). Our results demonstrated that Mandarin experience could increase the sensitivity to both pitch and phonation for native English speakers and their performance was positively correlated with L2 proficiency. Moreover, the high proficiency L2 learners significantly outperformed native Mandarin speakers, which implied a cumulative facilitating effect that L2 learners could benefit from both L1 and L2 experience in terms of pitch and phonation.

Index Terms: L3 tone perception, Chinese tones, phonation, L2 proficiency

1. Introduction

Previous studies on L3 tone perception have demonstrated that for native speakers of non-tone languages, learning a tone language facilitates L3 tone perception for naïve listeners [1, 2]. However, tone is a multidimensional concept and in the current study, we would like to explore whether tonal L2 experience improves the sensitivity to phonation in addition to its effect on tone pitch.

1.1. Linguistic status of phonation

Mandarin is a tone language with four lexical tones (T1 [55], T2 [35], T3 [214], T4 [51]) and studies have shown evidence that Mandarin T3 (the low-dipping tone) often co-occurs with creaky voice [3, 4, 5]. In addition, creaky voice could serve as an enhancement cue in Mandarin T3 perception [6, 7]. [8] conducted a Mandarin tone identification experiment in which the creaky feature was imposed on T2 and the results indicated that phonation cue (creaky voice) can even override pitch cue, which misled native Mandarin speakers to perceive T2 as T3. Wenzhou dialect is a subcategory of the Wu dialect. There are eight lexical tones in Wenzhou dialect with breathy voice in the low register and modal voice in the high register (as shown in Table 1) [9]. Different from Mandarin and Wenzhou dialect, phonation plays more of a sociolinguistic role in English, which is closely related to mood, attitude and gender [10]. Researchers have found that there is an increasing trend of creaky voice among young American and British women which reflects that voice quality is a sociophonetic marker [11, 12, 13].

Several studies have shown that language background affects voice quality perception [14, 15] and the relative amplitude of the first and second harmonics (H1-H2) serves as an important cue in voice quality perception [16]. Based on a discrimination task, [17] found that native Mandarin speakers were significantly more sensitive to H1-H2 than native English speakers were, from which we can infer that they are more sensitive to voice quality.

Table 1: Wenzhou Wu tone system.

<table>
<thead>
<tr>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch</td>
<td>33</td>
<td>31</td>
<td>45</td>
<td>34</td>
<td>42</td>
<td>11</td>
<td>323</td>
</tr>
<tr>
<td>Phonation</td>
<td>M</td>
<td>B</td>
<td>M</td>
<td>B</td>
<td>M</td>
<td>B</td>
<td>M</td>
</tr>
</tbody>
</table>

1.2. The role of L2 proficiency

Most studies on L3 transfer paid attention to the origin of transfer (L1 or L2) and the role of L2 proficiency was not systematically examined. In the morphosyntactic domain, some researchers found evidence that only high proficiency L2 could transfer [18, 19] while others believe that as long as there is typological similarity, L2 could transfer regardless of its proficiency [20]. At the phonological level, there is also mixed evidence: one study found that bilingual proficiency has no influence on the competence of L3 speech perception [21]. However, positive evidence has been found by [22] who tested the perception ability of Japanese phonemes by L1 Korean L2 English speakers and concluded that L3 speech perception is positively correlated with participants’ L2 perceptual ability. Thus, the present study will also focus on the relationship between L2 proficiency and L3 tone perception.

To sum up, phonation (creaky voice) functions at different linguistic levels in Mandarin (allophonic) and English (indexical). In addition, there is a controversy about the relationship between L2 proficiency and L3 performance at both morphosyntactic and phonological levels, and there is a gap in the L3 suprasegmental domain. Therefore, we investigated mainly three research questions: 1) Can the exposure to one phonation type (creaky voice) at a different linguistic level increase the sensitivity to another phonation type (breathy voice)? 2) Is L2 proficiency positively correlated with L3 tone perception performance? 3) Is there a cumulative facilitating effect, i.e., do non-tone language speakers with Mandarin learning experience perform better or worse than native tone-language speakers?
2. Method

2.1. Participants

40 participants were divided into four groups: the English monolingual group (ENG) consisted of 10 native British English speakers (average age: 25.8 years) and the Mandarin group (MAN) consisted of 10 native Mandarin speakers (average age: 22.6 years). L1 English L2 Mandarin speakers were put into two groups according to their Mandarin learning duration and there were 10 participants in each group: the EMH group (high proficiency with an average learning duration of 5.25 years) vs. the EML group (low proficiency with an average learning duration of 7 months). The language background of L2 learners is shown in Table 2. All participants were recruited from universities in the UK and mainland China. The participants were naïve to the Wenzhou dialect and had not lived in Wenzhou before. Participants with musical training (professional music study experience or attendance at a music school) were excluded from this experiment.

<table>
<thead>
<tr>
<th>Language group</th>
<th>Learning duration</th>
<th>HSK level</th>
<th>Living experience in China</th>
</tr>
</thead>
<tbody>
<tr>
<td>English native</td>
<td>5 years</td>
<td>HSK6</td>
<td>1 year in Beijing</td>
</tr>
<tr>
<td>speakers with</td>
<td>6 years</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>high Mandarin</td>
<td>5 years</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>proficiency (EMH)</td>
<td>8 years</td>
<td>HSK6</td>
<td>2 years in Lanzhou and Beijing</td>
</tr>
<tr>
<td></td>
<td>7 years</td>
<td>HSK5</td>
<td>1 year in Chengdu</td>
</tr>
<tr>
<td></td>
<td>4 years</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>5 years</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>3.5 years</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>5 years</td>
<td>HSK5</td>
<td>1 year in Beijing</td>
</tr>
<tr>
<td></td>
<td>4 years</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>6 months</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7 months</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10 months</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English native</td>
<td>6 months</td>
<td></td>
<td></td>
</tr>
<tr>
<td>speakers with</td>
<td>8 months</td>
<td></td>
<td></td>
</tr>
<tr>
<td>low Mandarin</td>
<td>8 months</td>
<td></td>
<td></td>
</tr>
<tr>
<td>proficiency (EML)</td>
<td>8 months</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9 months</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 months</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.2. Stimuli

To manipulate participants' sensitivity to pitch and phonation separately, we manipulated 2 sets of stimuli based on natural productions of the Wenzhou dialect.

Stimuli set 1: 3 syllables of the Wenzhou dialect with a CV structure were used ([tø], [tei], and [ji]). There are 8 lexical tones in Wenzhou dialect (Table 1) and each syllable was recorded with all 8 tones by a native speaker, which resulted in 28 AB tone pairs and 168 trials (1 speaker×8 tone pairs×3 syllables×6 repetitions), so 312 trials in total. All the trials were divided into four groups and there were 78 trials each. In the Wenzhou dialect, T1, T3, T5, T7 are of modal voice and T2, T4, T6, T8 are of breathy voice. To explore the effect of pitch, all 8 tones must have the same phonation type, therefore, pitch synchronous overlap-add (PSOLA) in Praat was applied to adjust F0 while ensuring that the other acoustic dimensions remain unchanged. Specifically, we manipulated the pitch value of modal voiced tones (T1, T3, T5, T7) to be the same as the corresponding pitch value of breathy voiced tones (T2, T4, T6, T8) using the Praat function.

For example, T1 [33] (modal) is manipulated to T6 [11] (modal) and in that way, we can get 8 lexical tones all in modal voice. We also checked that all the manipulated files sounded natural. The inter-stimulus interval (ISI) was 500ms because short ISI helps inexperienced participants to discriminate novel tones [23].

Stimuli set 2: To explore the effect of phonation type, we used the manipulated voices in stimuli set 1 (T2, T4, T6, and T8 in modal voice) together with the natural stimuli (T2, T4, T6, and T8 in breathy voice) and in this case, these four pairs of sounds differ only in phonation type. There were 72 AB trials (4 tone pairs×2 orders×3 syllables×3 repetitions) and 72 AA trials (4 tone pairs×3 syllables×6 repetitions), so 144 trials in total. All the trials were divided into two groups and there were 72 trials each. The inter-stimulus interval (ISI) was also 500ms.

2.3. Procedure

The experiment was conducted through the Gorilla [24] platform due to the pandemic. Firstly, participants were asked to get familiar with the procedure and adjust the volume to a suitable level before the real experiment (the instructions were given in English for native English speakers and Mandarin for native Mandarin speakers). They were required to wear headphones in a quiet environment for the experiment. They were asked to do 5 trial tests in an AX discrimination task. To improve the accuracy of the experiment, we divided 456 trials into 6 blocks and ensured that the numbers of AB and AA pairs in each block were as equal as possible. Stimuli set 1 was used in the first four testing blocks with 78 trials each (36 AA pairs and 42 AB pairs). Stimuli set 2 was used in the last two blocks with 72 tone pairs each (36 AA pairs and 36 AB pairs). All 456 trials were presented to each participant in random order and the participants were allowed to take a 1-minute break between blocks. Participants were asked to do each trial as quickly as possible because both accuracy rate and reaction time were recorded during the process.

3. Results

The result of one participant from the native Mandarin group was excluded from the subsequent data analysis because it deviated from the mean by more than three SDs.

Based on the Signal Detection Theory, we first explored participants' sensitivity to stimuli differences in the AX discrimination task using d' scores (as shown in Figure 1).

Figure 1: d' scores for stimuli set 1 (left) and stimuli set 2 (right)

For stimuli set 1 (tone pairs with different pitch values), the EMH group (M = 3.925, SD = 0.311) was most sensitive to the difference between target tones among all four groups, followed by the EML group (M = 3.616, SD = 0.317), MAN group (M = 3.274, SD = 0.370) and ENG group (M = 2.899, SD = 0.325). The d' scores of the EML, EMH, and MAN groups
were all over 3, which suggested that most participants were highly sensitive to the differences between stimuli. Significant difference was found between groups, F (3,35) = 17.83, p < 0.001 and the post-hoc Bonferroni test revealed that the EMH and EML groups were significantly more sensitive to stimuli set 1 than the ENG group were. Also, there was significant difference between the EMH and MAN group.

For stimuli set 2 (tone pairs with different phonation types), the $d'$ scores showed that participants were not as sensitive as they were to the first stimuli set. The pattern of stimuli set 2 slightly changed: the EMH group (M = 3.213, SD = 0.314) was still most sensitive to the sound difference and the ENG (M = 1.760, SD = 0.366) was still the least sensitive group. The results of the EML and MAN group were reversed compared to the results for stimuli set 1, with the $d'$ score of the MAN group (M = 2.468, SD = 0.345) slightly higher than that of the EML group (M = 2.222, SD = 0.269). In addition, there was also significant difference between language groups, F (3,35) = 34.87, p < 0.001. The post-hoc test indicated that the EMH group was significantly more sensitive to phonation than the other three groups and significant difference was also found between the MAN and ENG group.

3.1. Accuracy rate

We conducted 5 one-way ANOVAs to test whether there is significant difference between language groups in terms of the overall accuracy rate, the accuracy rates on pitch, phonation, level-level tone pairs, and contour-level tone pairs (as shown in Figure 2):

![Figure 2: Accuracy rates across language groups](image)

Significant difference was found for the overall accuracy rate among four language groups, F (3,35) = 80.85, p < 0.001. We then performed a post-hoc Bonferroni test, which indicated that except for the EML and MAN group pair, there was significant difference between all other language groups.

From Figure 2, we can tell that the accuracy of pitch patterns (left panel of the second row) was similar to the overall accuracy (first row) while the accuracy of phonation was much lower (right panel of the second row). There was significant difference between groups in terms of both pitch (F (3,35) = 22.94, p < 0.001) and phonation (F (3,35) = 35, p < 0.001). The post-hoc results demonstrated that there was no significant difference between EMH and EML groups for pitch accuracy while the EMH group performed significantly better than the EML group in phonation. Comparing the mean accuracy rates of EML and MAN groups, despite the slightly reverse patterns in pitch and phonation, no significant difference was found in both aspects.

We then analyzed the pitch accuracy rate in two types of pitch contour (as shown in the third row of Figure 2). For the level-level tone pairs (T1[33]T6[11], T6[11]T1[33], T1[33]T1[33], T6[11]T6[11], left panel), all four groups have achieved high accuracy rate (higher than 95%) with no significant difference between groups, F (3,35) = 1.403, p = 0.258. The pattern of contour level tone pairs (right panel) is consistent with the overall accuracy pattern, with the EMH (M = 0.964, SD = 0.016) group being the highest, followed by the EML (M = 0.958, SD = 0.011) group and MAN (M = 0.934, SD = 0.012) group, and the ENG (M = 0.913, SD = 0.021) group had the lowest accuracy rate. Significant difference was found between groups, F (3,35) = 22.76, p < 0.001 and according to the results of the post-hoc test, the L2 groups performed significantly better than the ENG group, p < 0.001. The EMH group also performed significantly better than the MAN group, p < 0.001.

Among all tone pairs which differ in pitch value, we also selected acoustically similar ones for separate analysis. There are mainly 6 tone pairs that are acoustically similar: T3[45] T4[34], T7[323] T8[212], T2[31] T5[42], T7[323] T1[33], T1[33] T4[34] and T8[212] T6[11]. They accounted for about 50% of the errors (Figure 3). No significant difference was found among group F (3,35) = 0.495, p = 0.688. Acoustically similar tone pair errors had the highest percentage in the MAN group (M = 0.557, SD = 0.193), followed by the EMH group (M = 0.500, SD = 0.228), ENG group (M = 0.481, SD = 0.186) and EML group (M = 0.433, SD = 0.273).

![Figure 3: Percentage of acoustically similar tone pair errors](image)

3.2. Reaction time

As no significant difference was found between groups in terms of the accuracy rate of level-level tone pairs, we tried to see if there were significant differences in reaction time. The result of one-way ANOVA demonstrated that the reaction time of the four language groups was relatively close, so no significant difference was found between groups, F (3,35) = 0.311, p = 0.818. The means and SDs of the four groups are: the EML group (M = 602.74, SD = 169.71) > the EMH group (M = 529.37, SD = 198.06), the ENG group (M = 539.59, SD = 143.68) > the MAN group (M = 529.37, SD = 198.06).

In addition, the standard deviations of reaction time were relatively large for all four language groups, which indicated a wide dispersion in the dataset. This could also be seen from the specific data: for example, the reaction time of level tone pairs for the EMH group ranges from 290.81ms to 1012.05ms. Therefore, in this experiment, the reaction time was more related to individual difference than to participants’ language background.
4. Discussion

Tone is a multidimensional concept, and we focused not only on tone pitch in the process of tone perception but also on phonation type. We tested English learners of Mandarin and used Wenzhou dialect as the L3 because of its breathy feature in the low-register tones. We investigated whether the exposure to creaky voice at different levels in L1 (indexical) and L2 (allophonic) increases L2 learners’ sensitivity to phonation, which is reflected by an increased sensitivity to breathy voice in this experiment. We also asked whether the enhancement in both pitch and phonation sensitivity is positively correlated with L2 (Mandarin) proficiency. In addition, we explored whether there is a cumulative facilitating effect in the process of naïve tone perception that leads to better performance of L2 learners than native Mandarin speakers.

From a within-group perspective, all participants were more sensitive to pitch difference than to phonation difference. Compared with the accuracy rate of stimuli set 1 (tone pairs with different pitch values), the mean accuracy rate of stimuli set 2 (tone pairs with different phonation types) of the ENG group even dropped by nearly 15% (from 91.9% to 77.9%). The other three language groups also had different degrees of decline in accuracy rate: 7.3% for the MAN group (from 93.7% to 86.4%), 10.3% for the EML group (from 95.8% to 85.5%) and 3.0% for the EMH group (from 96.6% to 93.6%).

From a between-group perspective, the results have shown that learning Mandarin does increase English learners’ sensitivity to pitch, and the EMH group performed significantly better than the EML group, which we believe is related to the participants’ language backgrounds. From Table 2, we can see that the mean Mandarin learning duration of the EMH group was over five years, which was much longer than that of the EML group (7 months). Also, several participants in the EMH group had living experience in China for at least one year, and some of them have reached a high level in the HSK test (HSK 5 or HSK 6). All these show that the participants in the EMH group were of very high Mandarin proficiency, which explains their superior performance compared to EML group. However, we did not find significant difference between groups in the level tone pairs (as shown in Figure 2). There are two possible reasons: First, there is only one level tone (T1 [55]) in Mandarin and F0 contour rather than F0 height is the most crucial cue in perceiving Mandarin tones. Previous research has shown that native English speakers are more sensitive to F0 height than native Mandarin speakers [1]. Therefore, for native English speakers, it is reasonable that learning Mandarin does not necessarily facilitate the perception of F0 height regardless of the L2 proficiency level, which leads to the result that no significant difference was found between the ENG, EMH and EML groups. Second, the difference between level tone pairs is quite evident in Wenzhou dialect (T6 [11] and T1 [33]), which resulted in high accuracy rates of all four groups (> 95%) and no significant difference between groups. In addition, acoustic similarity plays an essential role in naïve tone perception. If the target tone pair is very similar acoustically, it would be difficult for all speakers to distinguish regardless of their language backgrounds. As shown in Figure 3, about half of the errors for all four participant groups were related to acoustic similarity, and there was no significant difference between groups.

More importantly, when it comes to tone pairs that differ only in phonation type, L2 learners’ performance is significantly better than native English speakers, which means that learning Mandarin significantly improves native English speakers’ sensitivity to phonation. Phonation plays an allophonic role in Mandarin (co-occurring with the Mandarin low-dipping tone T3) and a phonemic role in Wenzhou dialect, while in English, phonation is at a post-lexical level. Then, why does the increased exposure to one type of phonation (creaky) lead to improved sensitivity to another type of phonation (breathy)? [25] proposed that phonation is a continuum from glottal mostly open to mostly closed, and there are several phonetic properties closely related to phonation type, the major one of which is spectral tilt. One of the methods to measure it is the relative amplitude of the first versus second harmonic (H1-H2), and similarly, the H1-H2 values of different phonation types are also on a continuum, with the creaky voice at the lower end and breathy voice at the higher end. In terms of phonation perception, studies have shown that native language background has a significant effect on H1-H2 sensitivity. By measuring the just-noticeable difference in H1-H2, [17] found that native Mandarin speakers are significantly more sensitive to changes in harmonic amplitude than native English speakers due to the presence of allophonic creaky in Mandarin T3. It is reasonable, then, that for native English speakers, the exposure to allophonic creakiness in Mandarin T3 would make them more sensitive to H1-H2, which leads to the result that they were more sensitive to breathy voice in this experiment.

Interestingly, we found that the EMH group even outperformed the MAN group, and similar patterns had been revealed in previous research on both perception and production [26]. Here are our assumptions: First, there may be a cumulative facilitating effect. Although previous research has shown that native English speakers were not as sensitive to pitch contour as Mandarin speakers were [1], L2 learners could benefit from both L1 and L2 experience in the process of naïve tone perception. Second, L2 learners may pay more attention to lexical tones than native Mandarin speakers because they were fully aware of the difficulty for learning tones and they may have increased metalinguistic awareness about tone. Since F0 plays a vital role in Mandarin tone perception, English learners would naturally pay more attention to F0, that is, to the frequency of H1, and then incidentally to the amplitude of H1. Therefore, they would be more sensitive to both pitch and phonation than native Mandarin speakers are.

In conclusion, for native English speakers, the sensitivity to pitch and phonation could be increased through learning Mandarin, and L2 proficiency has been demonstrated to be an essential factor. When L2 proficiency has reached a high level, the performance of naïve tone perception of native English speakers improved considerably, even surpassing that of native Mandarin speakers. This study provides evidence for the influence of L2 experience in L3 suprasegmental domain based on the AX discrimination task, and future research can focus on why learning Mandarin has such a substantial impact for native speakers of non-tone languages in terms of tone perception, and whether there will be a similar pattern in the identification task. In addition, participants’ language backgrounds can be more diverse in the future study to investigate whether the enhancement effect of Mandarin learning experience in naïve tone perception is general, and how it relates to different L1 backgrounds.

Acknowledgments: This study was partially supported by the RGC General Research Fund 2019/20 Project Reference 14607619 awarded to the last author.
5. References


