Cue weighting in Mandarin tone perception: A comparison between native speakers and learners of Mandarin

Song Jiang

University of Toronto
soong.jiang@mail.utoronto.ca

ABSTRACT

Some studies have shown that advanced learners of Mandarin perceive Mandarin tones more categorically compared to non-native speakers without prior experience. However, these studies have mainly focused on one single cue.

In the current study, the author investigated the interaction between two acoustic-phonetic cues (turning point position (TP) and f0 at the end (End f0)) during perception between Tone 2 and Tone 3 and how English listeners of Mandarin adapt their perceptual strategies to distinguish categories that do not exist in their L1. The results show that native Mandarin listeners tend to use a TP-dominant strategy, whereas non-native listeners rely less on TP. No significant difference was found in cue weighting between the learners and naïve English listeners. We suggest that native Mandarin listeners are able to exploit TP to filter out the uncritical portion, whereas learners of Mandarin focus more on the canonical textbook forms of Mandarin tones.

Keywords: Mandarin tones, contour shape, turning point position, cue weighting, L2 acquisition.

1. INTRODUCTION

In natural speech, phonetic categories are usually defined with multiple acoustic cues [1]. During speech perception, listeners need to integrate acoustic information across multiple dimensions for successful speech categorization. For example, Lisker [2] identified 16 acoustic properties that may play a role in the perception of English voicing. Nevertheless, these acoustic cues do not contribute to speech categorization equivalently. Instead, listeners rely on some cues more than others. This phenomenon has been referred to as cue weighting [3].

There has been controversy over the ability of L2 learners to adapt their perceptual cue weighting in response to different target languages [4, 5, 6]. These studies mainly focused on the acoustic properties that exist in the L1 of the learners (*e.g.*, VOT, formant information, duration). Mandarin tones perception is a good case for us to investigate if learners can use some acoustic cues in Mandarin tone perception that are non-contrastive in non-tonal languages (*e.g.*, f0 curve shape [7]). For example, distinguishing between Tone 2 (a rising tone) and Tone 3 (a low dipping tone) has been considered a more difficult task for both native and non-native speakers due to their similar concave contours [8, 9, 10, 11]. Multiple cues have been found to serve as important cues to the native perception of Tone 2 and Tone 3, such as the temporal position of the f0 minimum (a.k.a. turning point position or TP) [10, 11, 12] and the f0 at the vowel offset (End f0) [13]. A later TP and a lower End f0 serve as a cue for Tone 3. It has also been reported that advanced learners of Mandarin significantly outperformed non-native speakers without prior experience in the identification and discrimination of Tone 2-Tone 3 [8, 9]. What remains unclear is whether advanced learners of Mandarin achieve native-like categorical perception by modulating their cue-weighting strategies.

In the current study, we aim to investigate the relative weights of different cues in Mandarin tone perception. A comparison is made between native Mandarin speakers, English learners of Mandarin, and naïve English speakers. This study focuses specifically on two acoustic-phonetic cues, TP and End f0. Here arise our research questions:

1. What are the cue weights of TP and End f0 for the three groups?
2. How different is the cue weighting between native and non-native speakers?
3. Will L2 learning experience influence the cue weighting in Mandarin tone perception?

TP, which implies the pitch contour shape, is not an acoustic feature that listeners of non-tonal languages can use productively. Hence, I hypothesize that native Mandarin speakers may rely more on TP than non-native speakers. Learners of Mandarin may have acquired the cue weighting that native Mandarin speakers use in tone perception, or learners may use other cue weightings to achieve native-like categorical perception.

2. METHODS

2.1. Participants

One male native Mandarin speaker from North China (Liaoning Province) was recruited for stimuli

recording. Seventy-two participants were recruited for the perception experiment through the Linguistics Participant Pool at the University of Toronto. It consisted of three groups:

- MAN: native Mandarin speakers (n = 26)
- L2M: English learners of Mandarin (n = 26)
- ENG: native English speakers with no previous exposure to Chinese (n = 20)

All participants were first asked to complete a questionnaire on language background before the task. The L2M participants were asked to report their years of learning. Their years of learning averaged 7.5 years (range: 0.125 years – 20 years). None of the ENG participants had learned any tonal language as their second language. No speech or hearing impairments were self-reported.

2.2. Stimuli

Ten repetitions of the syllables /pa, pu, pi, fa, fu, a, u, i/ with Tone 2 and Tone 3 were recorded in the University of Toronto Phonetics Lab. The recording was done in a sound-attenuated booth using an AudioTechnica AT831b microphone and a SoundDevices 722 digital audio recorder, with a 16-bit depth and a sampling rate of 48 kHz.

Figure 1 illustrates the contours along the TP and End f0 continua and the distribution of the stimuli. The stimuli were designed to vary in TP along a continuum from the second to the eighth deciles. The duration was constant at 375 msec, the average of all recording tokens. The End f0 also varies in 7 Hz steps in the range of 144 Hz ~ 200 Hz, which is similar to the range of the End f0 of the recording.

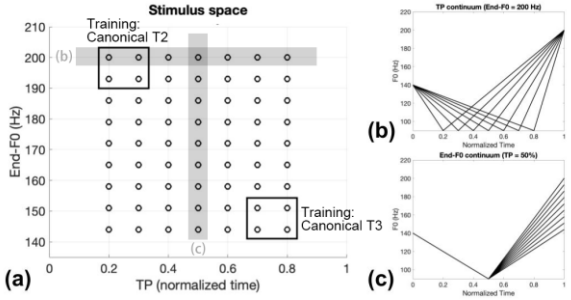


Figure 1: The diagrams of (a) the TP × End f0 stimulus space; (b) a TP continuum (End f0 = 200 Hz), and (c) an End f0 continuum (TP = 50%).

Acoustic measurements for the recordings were performed to determine the f0 range and vowel duration. Tone 2 and Tone 3 have close f0 values at the onset, around 140 Hz. The Onset f0 was set at 140 Hz. According to [12], 50 Hz of the difference between the onset and the minimum (Δf0) results in the most ambiguous perception along the TP continuum. The Δf0 was therefore set at 50 Hz – that is, the minimum is 90 Hz. No onset effect on f0 was found in our data. The onsetless /i/ was chosen as the target syllable because it was also used in previous studies [8, 11]. The target syllable /i/ was synthesized with different f0 contours covarying in TP and End f0, resulting in a set of stimuli spanning a two-dimensional perception space. All stimuli were created by using f0 synchronous overlap and add (PSOLA) method in PRAAT [14]. In total, 63 stimuli were created.

2.3. Procedure

A 2AFC identification task was implemented online using Gorilla [15]. The identification task included two phases: a training phase and a testing phase. In the training phase, two repetitions of "canonical" Tone 2 and Tone 3 stimuli were presented randomly in a single block. As illustrated in Figure 1(a), the four stimuli in the upper left box were the canonical forms of Tone 2, and the four stimuli in the lower right box were the canonical forms of Tone 3. The stimuli in isolation were presented to participants. Then, they were asked to identify the tone and click on the corresponding button on the screen. Depending on their native language, the question was provided in either English ("*Please identify the tone of the syllable you just heard. Did it have Tone 2 or Tone 3?*") or Mandarin ("*请问您刚刚听到的字是二声? 还是三声?*"). Apart from the names of the tones, the diacritics of tones in *Pinyin* were also provided as visual aids to facilitate responses. Feedback was provided during this phase.

Following training, the testing phase began. Ten repetitions of the entire stimulus space in Figure 1(a) were presented to participants. There were ten blocks in this phase. Each repetition (63 tokens) was presented and randomized in a separate block. A short break was provided between blocks. Participants were asked to do the same identification task as in the training phase, but no feedback was provided.

In total, there were 16 trials (2 repetitions × 8 stimuli) in the training phase and 630 trials (10 repetitions × 63 stimuli) in the testing phase. Depending on their first language, participants were provided written instructions in English or Mandarin.

2.4. Statistical analysis

To predict the model of identification results, we fitted generalized logistic models (GLM) to each individual [16, 17] and extended the one-dimensional model to a multi-dimensional stimulus space.[1] This allows us to estimate the relationship between the identification score (p, the probability of Tone 2 responses) and the two predictors, that is, TP and End f0. The MATLAB function *fitglm*() was used to predict the logistic model, using the formula in (1).

(1) $$\ln\left(\frac{1-p}{p}\right) = a_{TP}x_{TP} + a_{F0}x_{F0} + b$$

The categorical boundary can be derived from (1) corresponding to the 50% identification score, as (2). It can be simplified as (3), where the slope *m* represents the cue-weight ratio between TP and End f0. The intercept on the TP axis (1-*c*)/*m* represents the TP step before which all End f0 steps have Tone 2 identification.

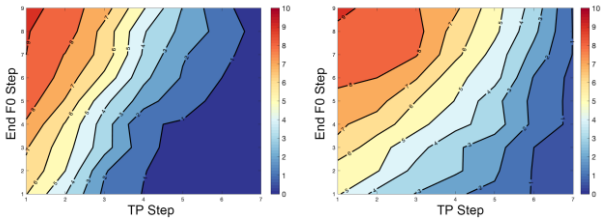
(2) $$a_{TP}x_{TP} + a_{F0}x_{F0} + b = \ln\left(\frac{1-0.5}{0.5}\right) = 0$$
$$\Rightarrow \quad x_{F0} = -\frac{a_{TP}}{a_{F0}}x_{TP} - \frac{b}{a_{F0}}$$

(3) $$x_{F0} = mx_{TP} + c$$

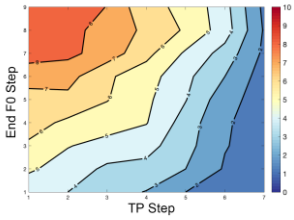
The boundary width of each participant's responses was also attained by calculating the distance between the 20% and the 80% categorical boundaries. It helps us quantitatively evaluate how categorical the perception of each group is.

3. RESULTS

The contour maps in Figure 2 show the average responses for each group. The MAN group exhibits a more vertical boundary than the other two groups. It implicates a higher cue-weight ratio between TP and End f0, which means that MAN listeners relied more on TP. The L2M and ENG groups show boundaries with similar slopes, which implies that the L2M and ENG groups have similar cue-weight ratios.



(a) MAN (*n* = 22) (b) L2M (*n* = 26)



(c) ENG (*n* = 20)

Figure 2: Contour maps of the average responses of each group. (Dark red area: Tone 2-dominant perception; dark blue area: Tone 3-dominant perception)

If we visually inspect the area between two and eight responses, which represents the area of ambiguous identification, the MAN group exhibits a narrower boundary than the other two groups. The ENG group has a wider ambiguous area than the L2M group. That means, in the current space, L2M listeners perceived tones in a more categorical way than ENG listeners since the identification pattern of L2M is more compact. The outperformance of L2M in categorical perception is consistent with [8].

Figure 3 shows the box plot of the boundary width of the ambiguous area. The wider the boundary width, the larger the stimulus space that is perceived ambiguously by the listener, which indicates a worse performance of categorical perception. The ANOVA results showed a significant effect of *Group* on *Boundary width* (F (2, 57) = 9.112, p < 0.001***). The results of Tukey's HSD *post hoc* analyses revealed that the ENG group had a significantly wider boundary width than the other two groups (ENG > MAN, p < 0.001***; ENG > L2M, p = 0.036*). No significant difference between MAN and L2M was found. The results confirmed that L2M identified tones in a native-like categorical way.

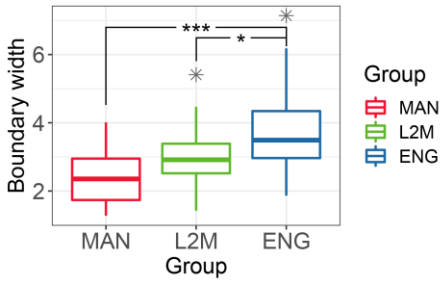


Figure 3: Box plot of the boundary width by *Group*.

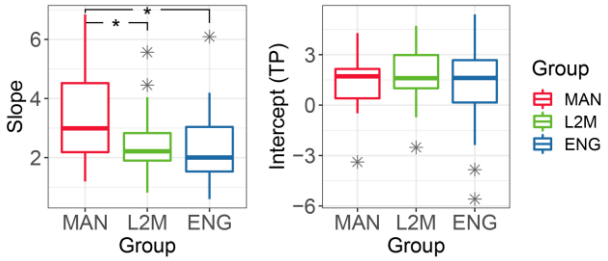


Figure 4: Box plots of the slope *m* (left) and the intercept (1-*c*)/*m* (right) by *Group*.

The slope *m* and the intercept on TP-axis (1-*c*)/*m* were predicted from the generalized logistic model fitted to each individual. If *m* is greater than 1, the speaker relied more on TP than End f0. The greater the slope, the higher the reliance on TP. According to the box plots in Figure 4, most participants show *m* values greater than 1. Only one L2M participant and two ENG participants have a slope of less than 1. The distributions of *Intercept* (*TP*) for the three groups overlapped. The ANOVA results revealed a significant effect of *Group* on *Slope* (F (2, 57) = 4.278, p = 0.018*). The results of Tukey's HSD *post hoc* analyses revealed that the MAN group had a significantly steeper boundary slope than the other two groups (MAN > L2M, p = 0.041*; MAN > ENG, p = 0.038*). No significant difference between L2M

and ENG was found. As for *Intercept (TP)*, no effect of *Group* was found (F (2, 57) = 0.736, p = 0.483).

Figure 5 shows the classification of listeners' reliance types based on the predicted slopes (χ^2 (6) = 10.78, p = 0.096·). The thresholds used here are the same as those used in [5]. The MAN group has 36.4% of speakers who relied exclusively on TP, and the percentage is larger than the other two groups. Moreover, only the L2M and ENG groups have speakers of the "more on f0" type; the ENG group has 14.3% more speakers of this type than ENG. The L2M and ENG groups have similar percentages of "Exclusively on TP" and "more on TP" speakers; however, the L2M group has 12.5% more speakers whose reliance type is "Mainly on TP" than ENG.

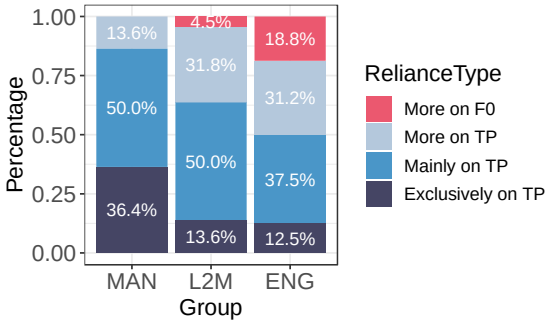


Figure 5: Percentage of each reliance type by *Group*. (More on F0: 0.5<*m*<1; more on TP: 1<*m*<2; mainly on TP: 2<*m*<4; exclusively on TP: *m*>4)

For the L2M group, Pearson's r was used to test the correlation between *Slope* and self-reported learning time of Mandarin. As Figure 6, no correlation between *Slope* and learning years was found (r = 0.0358, p = 0.8774).

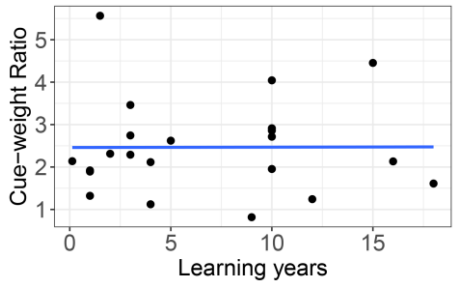


Figure 6: Scatterplot showing the distribution of Cue-weight Ratio (slope) by Learning year(s). Each L2M participant is represented by a dot.

4. DISCUSSION

The present findings demonstrated that the perceptual difference in cue weighting of TP and End f0 is significant between native (MAN) and non-native listeners (learners, L2M, and naïve English speakers, ENG). The results revealed that Mandarin listeners relied more on TP than the other two groups. For native Mandarin speakers, TP is associated with the proportion and duration of the critical portion of Tone 2 and Tone 3. An earlier TP renders a longer duration of the rising part, and a later TP renders a longer duration of the falling part. It has been proposed that the rising part and the falling part are the critical components of Tone 2 and Tone 3, respectively [18]. Moreover, the full dipping Tone 3 (MLM) usually becomes the half falling Tone 3 (ML) when it is not in the phrase-final position. In some Mandarin dialects (*e.g.*, Taiwanese Mandarin), the half Tone 3 is the only form of Tone 3, regardless of the context. Therefore, if the duration of the low falling part is long enough to be perceived, the tone will be identified as Tone 3 by Mandarin speakers. However, non-native listeners, who do not have the same experience with this tone patterning, may focus on the entire pitch contour, since for them, the full Tone 3 is the only prototypical form of Tone 3.

Recall that two predictions were on hold for the question of whether learners have acquired the native-like weighting to achieve native-like categorical perception or learners use other cue weightings, with which learners can achieve the same goal. Our present data did not show that the perception of learners is closer in terms of cue-weighting to that of native Mandarin listeners. Despite similar ability to categorize Mandarin tones, learners still have difficulty modulating their perceptual mechanisms to match native perception during learning. A possible explanation could be that the modulation failure is due to insufficient input; however, we did not find any correlation between the slope and proficiency. Learning experience appears not to take apparent effect on the cue-weighting of perceptual cues to Mandarin tone perception. Similar modulation failure has also been found in Spanish [5] and Japanese [6].

From the "top-down" perspective, adjusting cue weightings is never a specific learning or teaching target in L2 Mandarin courses, and therefore, it is typically not included in Mandarin course materials. Once learners can categorically perceive Mandarin tones, even though their cue weighting is not native-like, they can keep using the cue weight ratio that helps them steadily perform the identification. Moreover, the rising part is overtly emphasized in teaching and learning. Most learners of Mandarin are taught that the pitch contours are consistent with the diacritics in *Pinyin* (Tone 2: *á*, Tone 3: *ǎ*). The "v" shape of the Tone 3 diacritic may emphasize the impression that Tone 3 has a low rising at the end. The survey of Mandarin teaching materials in [19] showed that 14 out of 15 Mandarin textbooks for learners from Mainland China and Taiwan only represented Tone 3 as a low dipping tone (falling-then-rising). This knowledge-driven modulation could also result in attenuating the cue weight of TP, which implies the proportion of the falling part.

5. REFERENCES

[1] Keyser, S. J., & Stevens, K. N. 2006. Enhancement and Overlap in the Speech Chain. *Language* 82(1), 33–63.

[2] Lisker, L. 1986. "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech* 29(1), 3-11.

[3] Schertz, J., & Clare, E. J. 2020. Phonetic cue weighting in perception and production. *WIREs Cognitive Science* 11(2), e1521.

[4] Schertz, J., Cho, T., Lotto, A., & Warner, N. 2016. Individual differences in perceptual adaptability of foreign sound categories. Attention, Perception, and Psychophysics, 78(1), 355-367.

[5] Escudero, P., & Boersma, P. 2004. Bridging the gap between L2 speech perception research and phonological theory. Studies in Second Language Acquisition, 26(4), 551-585.

[6] Yamada, R. A., & Tohkura, Y. I. 1992. The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception and Psychophysics* 52(4), 376-392.

[7] Tupper, P., Leung, K., Wang, Y., Jongman, A., & Sereno, J. A. 2020. Characterizing the distinctive acoustic cues of Mandarin tones. *The Journal of the Acoustical Society of America* 147(4), 2570-2580.

[8] Shen, G., & Froud, K. 2016. Categorical perception of lexical tones by English learners of Mandarin Chinese. *The Journal of the Acoustical Society of America* 140(6), 4396-4403.

[9] Han, J. I., & Tsukada, K. 2020. Lexical representation of Mandarin tones by non-tonal second-language learners. *The Journal of the Acoustical Society of America* 148(1), EL46-EL50.

[10] Shen, X. S., & Lin, M. 1991. A perceptual study of Mandarin tones 2 and 3. *Language and Speech* 34(2), 145-156.

[11] Chow, R., Liu, Y., & Ning, J. 2019. The Categorical Perception of Mandarin Tone 2 and Tone 3 by Tonal and Non-tonal Listeners. In Sasha Calhoun, Paola Escudero, Marija Tabain and Paul Warren (eds.) *Proceedings of the 19th ICPhS*, Melbourne, Australia 2019, 3877-3881.

[12] Moore, C. B., & Jongman, A. 1997. Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America* 102(3), 1864-1877.

[13] Shen, J., Deutsch, D., & Rayner, K. 2013. On-line perception of Mandarin Tones 2 and 3: Evidence from eye movements. *The Journal of the Acoustical Society of America* 133(5), 3016-3029.

[14] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

[15] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 52(1), 388–407.

[16] Xu, Y., Gandour, J. T., & Francis, A. L. 2006. Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America* 120(2), 1063-1074.

[17] Morrison, G. S., & Kondaurova, M. V. 2009. Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis. *The Journal of the Acoustical Society of America* 126(5), 2159-2162.

[18] Liu, S., & Samuel, A. G. 2004. Perception of Mandarin lexical tones when f0 information is neutralized. *Language and Speech* 47(2), 109-138.

[19] Linge, O., 2011. Teaching the third tone in Standard Chinese: tone representation in textbooks and its consequences for students. *LUP Student Papers.* Online reference: http://lup.lub.lu.se/student-papers/record/1979823.

[20] Xu, Y. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics* 25(1), 61-83.

[21] Hallé, P. A., Chang, Y. C., & Best, C. T. 2004. Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of phonetics* 32(3), 395-421.

[1] Participants who did not identify Tone 2 and Tone 3 in a correct way were excluded from the logistic regression because no meaningful boundary could be predicted from their responses.