CONFUSABILITY OF MANDARIN TONE 3 AND TONE 4: EFFECTS OF FOCUS AND SYLLABLE POSITION

Angeliki Athanasopoulou[+], Irene Vogel*, Chao Han*, Yue Yuan*

*University of Delaware, [+]University of Calgary
hanchao@udel.edu, ivogel@udel.edu, angeliki.athanasopou@ucalgary.ca, yueyuan@udel.edu

ABSTRACT

It is well known that Mandarin tones are not always clearly produced and perceptually distinguished, especially Tones 3 (dipping) and 4 (falling). Prosodic structure affects the production of tones, for example pitch range is expanded under focus. We investigate the effects of prosodic structure (syllable position and focus) on the perceptibility of Mandarin tones. We find that focus and syllable do affect tone perceptibility; tone perceptibility increases when words are focused, but it is reduced when the word is in the second syllable of a trisyllabic word. These prosodic effects are the same across the four tones, but as reported in the literature, Tones 3 and 4 are less clearly distinguished compared to Tones 1 and 2. We discuss the role of pitch contour and phonation properties of the tones in their perception.

Keywords: Mandarin, tones, perception, production, Focus, Syllable Position

1. INTRODUCTION

Prescriptively, Mandarin Chinese is described as having four contrastive tones based on their pitch properties: Tone 1 (T1=high), Tone 2 (T2=rising), Tone 3 (T3=dipping), and Tone 4 (T4=falling). This characterization is primarily based on monosyllabic words in isolation. It has also been noted that duration, intensity, and phonation may be secondary cues for the tones, and it is well known that prescriptive characterizations of sounds, including tones, are not always observed in more natural connected speech. For example, in connected speech, tones show minimal durational differences [34], and words often exhibit less prosodic strength than when they are uttered in isolation [37]. Other factors, however, may provide enhancements, for example, focus may increase a tone's duration, pitch range and intensity [5, 6, 9, 18, 27, 33, 36].

In the present study, we examine the perception of Mandarin tones in real words drawn from connected speech to assess the effects of prosodic structure (i.e., focus and syllable position), on the clarity of their production, and thus their distinguishability for native speakers. We primarily consider Tones 3 and 4, which seem to be prone to confusion, likely due to their shared initial falling contour [1]; Tones 1 and 2 are also considered for comparison.

2. MANDARIN TONES

The main properties of Mandarin tones in both production and perception pertain to pitch: F0 height, contour, change or turning points. (e.g., [10] and references therein). Other properties, (e.g., intensity, duration, phonation) may enhance the perception of the tones [10, 35], but on their own, they are not reliable for successful perception [3, 7, 8, 24].

While most of the previous claims about Mandarin tones rely on monosyllabic words produced in isolation, some studies investigate the properties of tones in connected speech. For example, [22, 31] tested the automatic recognition of the tones in trisyllabic words produced by native speakers, and found that tone identification is better in Syllable 1 (Syll1) than in Syllables 2 and 3 (Syll2, Syll3). In some cases, tone recognition in Syll2 is also weaker than Syll3. The effect of syllable may be due to tonal coarticulation and word prosodic structure (e.g., Feet) in 2- and 3-syllable words, with tone reduction in Syll2 (i.e., the F0 contour is less defined than in Syll1) [15, 16, 19, 20, 23, 28, 29, 32, 37]. In both cases, Syll2 is a less prominent position.

Focus is especially relevant, since it affects the prosodic structure of a sentence, introducing a strong boundary following the focused item [26]. It has been reported that tones in focus positions often show increased duration, intensity, and pitch range [5, 6, 9, 18, 27, 33, 36]; though, the last property is limited for T3, which already extends to the lowest portion of the pitch range [3, 5, 18, 33]. Moreover, since words following a focus tend to be reduced [33, 36], and those in sentence-final position show the lowest perceptual accuracy rate [18, 21, 36], it is crucial to examine words produced in connected speech, not just in isolation, to understand the production and perception of tones.

3. PRESENT STUDY

The present study tests the potential perceptual confusion of Mandarin T3 and T4 in CV words originally produced as parts of trisyllabic (compound) words in connected speech, where they may exhibit

similar falling contours, especially when the final rise of T3 is absent. The corpus was constructed with controls for a number of factors that may be relevant to tone perception, permitting us to examine, in particular, the effects of syllable position and focus.

Since the beginnings and ends of words tend to be the most salient positions, we test Hypothesis 1:

(1) *Hypothesis 1*: T3 and T4 are perceived more successfully when they are at the edge of a word than in the middle of a word.

Moreover, since Focus tends to enhance prosodic properties, we test Hypothesis 2:

(2) *Hypothesis 2*: T3 and T4 are perceived more successfully when they are produced in a focus context than a non-focus context.

Note that the presence of focus may additionally contribute to a positive finding for Hypothesis 1, since both the word edge and focus factors would be present on the last syllable of a focused word.

4. METHODOLOGY

4.1. Auditory Stimuli

The stimuli, monosyllabic CV words, were extracted from a large corpus collected for acoustic analysis of word and phrase prosody [1]. The full corpus consists of 10 native speakers of Mandarin (4F; ages 18-28) producing real three-syllable (compound) words containing 6 instances of /i, u, a/ in each syllable position, with each of the 4 tones. Although the compounds were left-headed (e.g., *bā xiān zhuō* 'square table'), right-headed (e.g., *dī zī tài* 'modesty'), or without a clear head (e.g., *sū mù zhē* 'poem style'), an initial analysis showed that the head position did not affect the production of the tones [1]. All of the words were produced in two dialogues, priming focus either on the target compound word (Focus Condition) or on a word after the compound (Non-Focus Condition), yielding 432 targets per speaker. Only congruous tones appeared in the words adjacent to the targets (in the compound itself or an adjacent word in the carrier sentence) to limit tonal coarticulation. For example, T1 was preceded by tones with a final high target (T1 or T2), and followed by tones with an initial high target (T1 or T4). Sequences of T3 were excluded to avoid tone sandhi.

In the present perception study, a subset of the corpus was used, drawn from 4 female and 4 male speakers. An experimental block was constructed for each voice, consisting of 72 target CV words, with /i, u, a/ bearing T3 or T4, taken equally from all three syllable positions, and both Focus contexts. Another 36 CV syllables, half with T1 and half with T2, served as distractors. Each participant heard two blocks of male and two blocks of female speakers, presented randomly (e-prime); all voices appeared the same number of times.

4.1.1. Acoustic Properties of Auditory Stimuli

To understand how the phonetic properties of the auditory stimuli may play a role in the perception of their tones, we first summarize the relevant acoustic properties of the corpus. For a full acoustic analysis of the corpus, the reader is referred to [1].

T3 and T4 differ crucially in pitch and phonation; while duration and intensity play little or no role. Figure 1 shows that both tones have a falling F0 contour, but T3 is lower than T4. We also observe a general expansion of F0 range under focus, resulting in larger pitch differences between the tones. Figure 2 shows that the phonation property, HNR (Harmonic -to-Noise Ratio), is low across T3, indicating creaky phonation throughout the tone, while there is a drop in HNR in the latter part of T4, where the tone has a low pitch target. Though they are not directly relevant here, the properties of T1 and T2 are also shown for comparison

Figure 1. Normalized F0 by Syllable and Focus

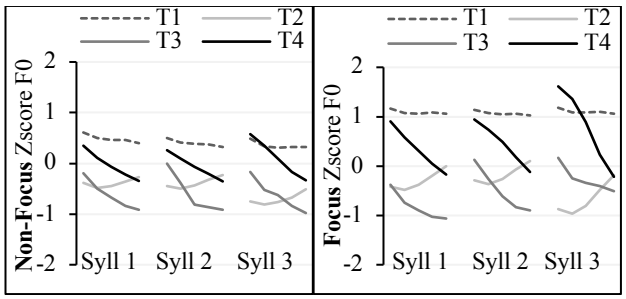
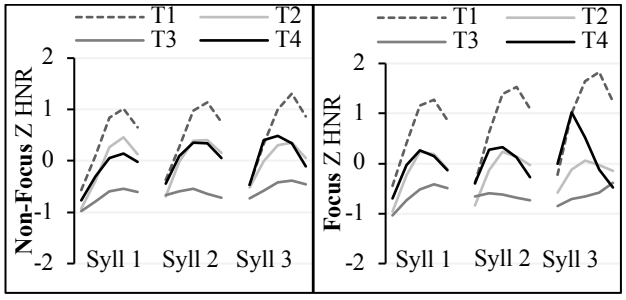


Figure 2. Normalized HNR by Syllable and Focus



4.2. Procedure

Seventeen native speakers of Mandarin (11 Females), were tested by a native speaker at a US University.

Each participant heard two repetitions of each stimulus (monosyllabic CV word), and selected one of four characters corresponding to words differing only in tone (e.g., /ma/: T1 妈 'mother', T2 麻 'hemp', T3 马 'horse', T4 骂 'scold'). Before each block of stimuli, the participants heard two full dialogues

recorded for the production experiment, to familiarize them with the voice of the speaker.

4.3. Analysis

The responses in the perception experiment were first analyzed for accuracy of T3 and T4 identification, (i.e., selection of the character corresponding to the auditory stimulus). Corresponding results for the distractors, T1 and T2, were also examined, for the purposes of comparison. Since we found a bias in the responses, we also ran d' analyses for sensitivity.

5. RESULTS

5.1. Overall Accuracy and Sensitivity

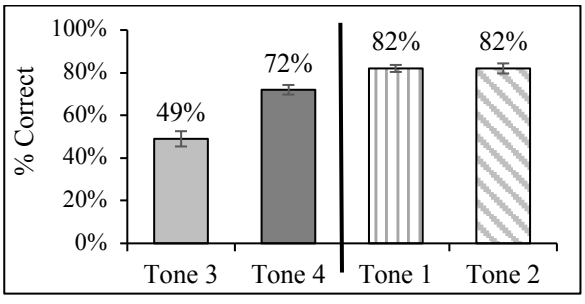
An initial check of the overall accuracy rates for the 8 stimulus voices, (unsurprisingly) revealed some speaker differences; however, there is no pattern related to the gender of the speaker's voice (Table 1). The data from the 8 voices were thus combined.

Table 1. Overall accuracy by speaker's gender

Speaker Gender	F1	M1	F2	M2	F3	M3	F4	M4
Overall Accuracy	56%	64%	65%	65%	70%	73%	75%	77%

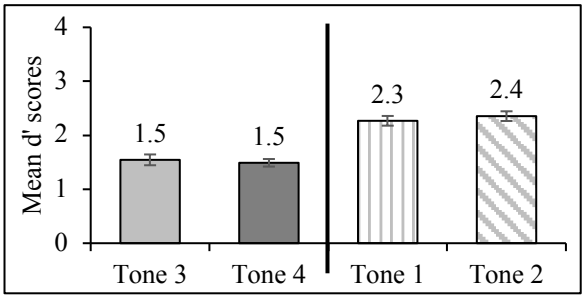
Figure 3 provides the rate of correct selection of each tone. A generalized linear model analysis showed a main effect of Tone on the accuracy rate ($\chi^2(3) = 44.91, p < .001$). Pairwise t-tests (Bonferroni correction) revealed that the response rate for T4 was considerably higher than for T3 ($p < .001$); both were lower than the rates for T1 and T2 ($p < .001$).

Figure 3. Correct perception for each tone.



The difference in accuracy between T3 (49%) and T4 (72%) might be a result of T4 being better perceived than T3, or a result of a bias towards perceiving T4. We thus tested the sensitivity to the tones using a d' analysis that takes bias into consideration. Figure 4 gives the d' values for each tone. Pairwise t-tests between all pairs of tones showed that T3 and T4 d' scores are not statistically different, nor are the T1 and T2 scores. The sensitivity to both T3 and T4, however, is lower than to T1 and T2. So, when the T4 bias is removed, we no longer see a difference between T3 and T4 perception. In the following sections, we only consider the d' values, to avoid the effect of bias in the perception.

Figure 4. Mean d' scores for each tone.



5.2. Effects of focus and syllable position

The d' values for T3 and T4 are shown in Figure 5, for each syllable, and each Focus condition. A repeated-measures ANOVA assessed d' as a function of focus and syllable and revealed a main effect of Syllable ($p < .001$) and Focus ($p < .001$).

With respect to the Syllable effect, post-hoc tests (Bonferroni correction) show that tone sensitivity is lower in Syll2 than the other two syllables ($p < .01$); Syll1 and Syll3 are not significantly different from each other. Note that the low d' in Syll2 arises in both T3 and T4. Focus also affects both tones similarly, significantly increasing their perceptibility ($p < .05$).

Figure 5. T3 & T4 d' scores by Syllable and Focus

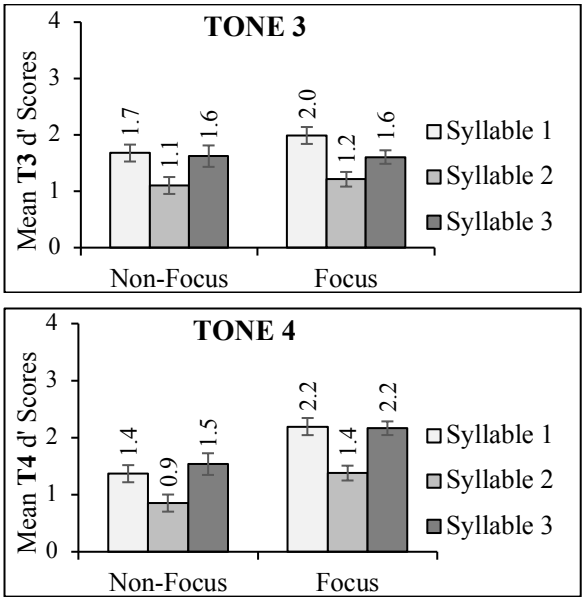
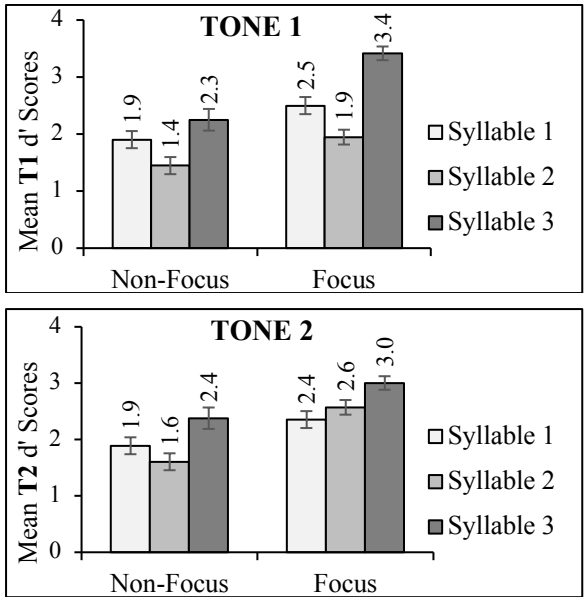


Figure 6 gives the corresponding findings for T1 and T2, which serve as examples of tones expected to show strong perceptibility. Again, we see an effect of focus and syllable: Syll2 has the lowest d' values, while Focus increases d'.

Figure 6. T1 & T2 d' scores by Syllable and Focus



6. DISCUSSION

The present study investigates the perception of Mandarin T3 and T4, and the possible effects of prosodic structure (syllable position and focus) on their perception. Overall, we find that the two tones are highly confused with each other, especially in comparison with the greater accuracy seen in the perception of T1 and T2.

When viewed in light of the acoustic properties of the stimuli, the perceptual confusion of T3 and T4 suggests that that this pattern likely arises as result of the acoustic signal. That is, while these tones exhibit similar pitch contours overall, there is nevertheless a considerable difference in their pitch height, as well as in their phonation, in particular, creaky phonation. The fact that we observed a bias towards perceiving T4 in place of T3, however, suggests that the listeners rely more on the contour of the tones, and somewhat less on the pitch height, or on phonation.

With respect to prosodic structure, we found a significant effect of syllable position and focus. Specifically, T3 and T4 were perceived more successfully in syllables 1 and 3 compared to Syll2, confirming Hypothesis 1. Our results agree in part with the previous automatic recognition studies that found more successful tone identification in Syll1 than in Syll2 and Syll3 [22, 31]. The difference for Syll3 may be a result of the manifestation of the final syllable of a word when produced in isolation (as in previous studies) and when produced within a sentence (as in the current study). Our results also agree with previous acoustic analyses where Syll2 was found to be the weakest syllable in trisyllabic words based on consonant lenition and tone coarticulation patterns (e.g., [15]).

In addition, in support of Hypothesis 2, we found that T3 and T4 are perceived more successfully when they were produced in a focus context than a non-focus context. This is consistent with previous studies showing Mandarin tones to have enhanced pitch range, duration and intensity when they are focused [5, 6, 9, 18, 27, 33, 36]. This acoustic enhancement, in turn, leads to their more successful perception.

7. CONCLUSIONS

We investigated the possible perceptual confusion of Mandarin tones 3 and 4 drawn from three-syllable words in a large corpus with connected speech, where their manifestations may not always conform to their prescriptive descriptions. Tones 1 and 2 were also tested for comparison, showing clear tone perception.

When selection of the correct tone was examined, it appeared that only T3 was not reliably perceived; T4 was quite successfully identified, more like T1 and T2. The distribution of the errors, however, suggested a response bias favoring T4, so additional d' analyses were conducted. When the bias was removed, T3 and T4 were instead found to be similar, and distinct from T1 and T2, the clear cases. That is, T3 and T4 both showed relatively little sensitivity, compared to the greater sensitivity shown for T1 and T2. The different interpretations of the confusability of T3 and T4 show, moreover, that we cannot just rely on correct scores, without considering possible response biases.

The two prosodic factors tested were both shown to affect the perception of T3 vs. T4. That is, perceptibility was reduced in syllable 2, compared with syllables 1 and 3. Focus, however, improved perceptibility of the T3 vs. T4 distinction.

Since perception is directly connected with the production properties, we also considered whether particular acoustic properties could account for the perception patterns. The main source of the problem with T3 and T4 was found to be the fact that their pitch contours are quite similar, since both begin with a fall, and T3 often loses its final rise. Creaky phonation, often associated with T3, however, did not lead to better perception of that tone.

8. REFERENCES

[1] Athanasopoulou, A., Vogel, I. In Prep. The acoustic properties of Mandarin tones in connected speech: effects of focus and syllable position.

[2] Belotel-Grenié, A., Grenié, M. 1994. Phonation types analysis in Standard Chinese. *ICSLP*, 343-6.

[3] Cao, R. 2012. *Perception of Mandarin Chinese Tone 2/ Tone 3 and the role of creaky voice*. University of Florida.

[4] Cao, W., Zhang, J. 2008. Tone-3 accent realization in short Chinese sentences. *Tsinghua Science and Technology* 13, 533-9.

[5] Chen Y., Gussenhoven, C. 2008. Emphasis and tonal implementation in Standard Chinese. *J. Phonetics* 36, 724-46.

[6] Chen, Y., Braun, B. 2006. Prosodic realization in information structure categories in standard Chinese. *Speech Prosody*, 54-7.

[7] Cheng, C.-C., Sherwood, B. 1982. Technical aspects of computer assisted instruction in Chinese. *The Tsing Hua J. Chinese Studies*, 35-49.

[8] Gårding, E., Kratochvil, P., Svantesson, J.-O., Zhang, J. 1986. Tone 4 and Tone 3 discrimination in Modern Standard Chinese. *Language and Speech* 29, 281-93.

[9] Jin, S. 1996. *An acoustic study of sentence stress in Mandarin Chinese*. Ohio State University.

[10] Jongman, A., Wang, Y., Moore, C., Sereno, J. 2006. Perception and production of Mandarin Chinese tones. In: Li, P., Tan, L., Bates, E., Tzeng O. (eds), *The Handbook of East Asian Psycholinguistics.* Cambridge: Cambridge University Press, 209-17.

[11] Kong, J. 2007. *Laryngeal dynamics and physiological models: High speed imaging and acoustical techniques.* Beijing: Peking University Press.

[12] Kuang, J., Liberman, M. 2018. Integrating voice quality cues in the pitch perception of speech and non-speech utterances. *Front. Psychol.* 9, 2147.

[13] Kuang, J. 2013. *Phonation in Tonal Contrasts.* University of California, Los Angeles.

[14] Kuang, J. 2017. Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *J. Acoust. Soc. Am.* 142, 1693-1706.

[15] Lai, W., Kuang, J. 2016. Prosodic grouping in Chinese trisyllabic structures by multiple cues - tone coarticulation, tone sandhi and consonant lenition. *TAL*.

[16] Lai, W., Liberman, M., Yuan, J., Xu, X. 2016. Prosodic Strength Intrinsic to Lexical Items: A Corpus Study on Tone Reduction in Tone4+Tone4 Words in Mandarin Chinese. *ISCSLP*.

[17] Lee, C.-Y., Tao, L., Bond, Z.S. 2009. Speaker variability and context in the identification of fragmented Mandarin tones by native and non-native listeners. *J. Phonetics* 37, 1-15.

[18] Lee, Y.-C., Wang, T., Liberman, M. 2016. Production and Perception of Tone 3 Focus in Mandarin Chinese. *Front. Psychol.* 7, 1058.

[19] Li, Y., Tao, J., Hirose, K., Xu, X., Lai, W. 2015. Hierarchical stress modelling and generation inMandarin for expressive Text-to-Speech. *Speech Communication* 72, 59-73.

[20] Li, Y., Tao, J., Zhang, M., Pan, S., Xu, X. 2010. Text-based unstressed syllable prediction in Mandarin. *Interspeech*, 1752-5.

[21] Liu, F. 2009. *Intonation Systems of Mandarin and English: A Functional Approach*. University of Chicago.

[22] Liu, L.-C., Yang, W.-J., Wang, H.-C., Chang, Y.-C. 1989. Tone recognition of polysyllabic words in Mandarin speech. *Computer Speech and Language* 3, 253-64.

[23] Liu, M., Shi, S., Zhang, J. 2014. A preliminary study on acoustic correlates of tone2+tone2 disyllabic word stress in Mandarin. *Interspeech*, 179-83.

[24] Massaro, D., Cohen, M., Tseng, C. 1985. The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *J. Chinese Linguistics* 13, 267-89.

[25] Minli, C., XinMin, C., Li, Z. 2003. HMM based recognition of Chinese tones in continuous speech. *Int. Conf. Neural Networks and Signal Processing*, 916-9.

[26] Nespor, M., Vogel, I. 1986. *Prosodic phonology*. Dordrecht: Foris.

[27] Ouyang, I., Kaiser, E. 2015. Prosody and information structure in a tone language: An investigation of Mandarin Chinese. *Language, Cognition, and Neuroscience* 30, 57-72.

[28] Peng, S., Chan, M., Tseng, C., Huang, T., Lee, O., Beckman, M. 2005. Towards a Pan-Mandarin system for prosodic transcription. In: Jun, S.-A. (ed.), *Prosodic typology: The phonology of intonation and phrasing*. Oxford: Oxford University Press, 230-70.

[29] Shih, C. 2005. Understanding phonology by phonetic implementation. *Interspeech*, 2469-72.

[30] Shue, Y.-L., Keating, P., Vicenik, C., Yu, K. (2011) VoiceSauce: A program for voice analysis. *ICPhS*, 1846-9.

[31] Wu, Y., Hemmi, K., Inoue, K. 1991. A tone recognition of polysyllabic Chinese words using an approximation model of four tone pitch patterns. *IECON*.

[32] Xu, Y. 1994. Production and perception of coarticulated tones. *J. Acoust. Soc. Am.* 95, 2240-53.

[33] Xu, Y. 1999. Effects of tone and focus on the formation and alignment of f0 contours. *J. Phonetics* 27, 55-105.

[34] Yang, J., Zhang, Y., Li, A., Xu, L. 2017. On the Duration of Mandarin Tones. *Interspeech*, 1408-11.

[35] Yang, R. 2015. The role of phonation cues in Mandarin tonal perception. *J. Chinese Ling.* 43, 453-72.

[36] Yuan, J. 2004. *Intonation in Mandarin Chinese: Acoustics, Perception, and Computational Modelling*. Cornell University.

[37] Zhang, W., Hao, L., Xie, Y., Zhang, J. 2017. A study on quantitative computation for prosodic strength of Mandarin speech. *APSIPA ASC*, 926-30.