University of Pennsylvania Working Papers in Linguistics

Volume 26
Issue 1 *Proceedings of the 43rd Annual Penn Linguistics Conference*

Article 12

10-1-2020

Perceptual Confusion of Mandarin Tone 3 and Tone 4

Chao Han

Irene Vogel

Yue Yuan

Angeliki Athanasopoulou

Perceptual Confusion of Mandarin Tone 3 and Tone 4

Abstract

In connected speech, the acoustic properties of Mandarin tones undergo modifications not observed in isolation. The current study investigated the perceptual distinction between Mandarin tones in connected speech, focusing on Tone 3 and Tone 4, which have been reported to share a similar initial falling contour. The current study also tested whether syllables produced with focus and / or in certain syllable positions affect the tonal perception. In a forced choice perception task, participants heard syllables extracted from three syllable words previously recorded in short dialogues, and were instructed to select one of four characters representing corresponding monosyllabic words differing only in tone. The accuracy results showed that Tone 4 was much more successfully identified than Tone 3. Nonetheless, after using a d-prime analysis to control for an observed T4 response bias, we found the same level of perceptibility of T3 and T4. Furthermore, the two tones were better perceived when a tone was produced in a focus context or at the edge of a word, confirming the effect of prosodic structure on tonal perception.

Perceptual Confusion of Mandarin Tone 3 and Tone 4

Chao Han, Irene Vogel, Yue Yuan, and Angeliki Athanasopoulou

1 Introduction

Mandarin Chinese is described as having four contrastive tones with distinct pitch properties: Tone 1 (T1 = high), Tone 2 (T2 = mid-rising), Tone 3 (T3 = low-falling-rising), and Tone 4 (T4 = falling). The meaning of a Mandarin word depends crucially on its tone since the same syllable could have multiple meanings depending on which tone appears with it. The prescribed pitch properties of the four tones are primarily based on monosyllabic words uttered in isolation; however, it has been observed that words in connected speech undergo modifications of these properties. For example, it has been noted that in connected speech, the pitch contours of the tones may vary extensively in different contexts (Xu 1994, 1997), and that the duration of the tones becomes shorter, resulting in minimal duration differences among them (Yang, et al. 2017).

Mandarin speakers do not, however, have trouble understanding connected speech. This could mean that there remain adequate acoustic cues to identify the tones, or if not, that speakers rely on other information (i.e., context) to interpret the intended tones. In this paper, we examine the extent to which native Mandarin speakers are able to distinguish between the tones of syllables produced in connected speech when the syllables are extracted from their context, and the only information about their meaning is thus the tonal pattern itself. We focus on the distinction between T3 and T4, and additionally consider whether two prosodic properties, syllable position in a word and focus, contribute to the clarity of the tones, and thus the perceptibility of the tonal contrast.

In Section 2, we discuss our research question and hypotheses. Section 3 introduces our methodology, followed by the results in Section 4. Sections 5 and 6 present our discussion and conclusions, respectively.

2 Mandarin Tone Confusion

Previous studies of tonal confusion in Mandarin have tended to focus on T2 and T3 (e.g., Shen and Lin 1999, Cao 2012), noting their physical similarity: both tones are said to have a concave shape, and both start with an F0 value that falls in the middle of speaker's pitch range (Shen and Lin, 1991, Huang 2001). The former similarity is often lost in connected speech, however, where the rising part of T3 tends to be reduced, leaving a falling contour, more similar to that of T4 (Gårding 1987). Since we are concerned here with connected speech properties, we therefore expect that words with T3 and T4 are the ones that would be most susceptible to confusion with each other in the absence of contextual information.

We thus first test the following hypothesis regarding the perceptibility of T3 versus T4 in general:

Hypothesis 1: T3 tends to be confused with T4 and vice versa.

Since certain prosodic conditions have been found to affect the acoustic manifestation of tones, we also investigate whether their effect is observable in the perception of the tones. Focus often affects the prosodic structure of a sentence, introducing a strong boundary following the focused item (Nespor and Vogel 1986), and it has been reported that tones in focused positions often show increased duration and pitch range (Chen and Braun, 2006, Chen and Gussenhoven 2008, Ouyang and Kaise 2015, Lee, Wang, and Liberman 2018). We thus assess the effect of focus on the perceptibility of T3 versus T4. That is, we test the following hypothesis regarding the clarity of the tonal properties produced with and without focus:

Hypothesis 2: T3 and T4 are more perceptually distinct when they have been produced in a focus context than in a non-focus context.

In connected speech, the position of the syllable in a sentence may also affect the acoustic manifestation of tones (e.g., Zhang et al. 2018). We thus test the following additional hypothesis:

Hypothesis 3: T3 and T4 are more perceptually distinct when they have been produced at the edge (beginning and end) of a word, as opposed to the middle of a word.

Although our focus here is on the perceptual patterns for T3 and T4, we also consider T1 and T2 by way of comparison. That is, since these tones are expected not to exhibit much confusability, they serve as a type of baseline for evaluating the possible confusion between T3 and T4.

3 Methodology

3.1 Stimuli

The stimuli were CV words extracted from a corpus of Mandarin connected speech previously collected for acoustic analysis (Athanasopoulou and Vogel, In preparation). In that corpus, a total of 4320 target vowels appeared in real three-syllable compounds produced by ten native speakers (18-28 years old). All four tones appeared in six syllables with the vowels /i, u, a/ in all three syllable positions. The same items appeared in both a focus and a non-focus context. To minimize tonal co-articulation, each target syllable was flanked by syllables carrying congruent tones[1]. Two adjacent syllables with T3 were not permitted to avoid the application of tone sandhi.

For the present perception experiment, 108 CV words were extracted from the recordings of four male and four female speakers (total = 864). These stimuli included two items with each of the vowels /i, u, a/ and with T3 and T4 in all three syllable positions. For T1 and T2, only one item with each vowel appeared in all three syllable positions. All of the items were produced in both focus and non-focus contexts.

2.2 Participants

Seventeen native Mandarin speakers (11 females) between 18 and 30 years old (mean age: 22 years) participated in the perception study. The experiment was conducted at the University of Delaware.

2.3 Procedure

The experiment was presented using E-Prime, and the task required that the participants select a character corresponding to each monosyllabic word they heard. Each participant heard four blocks of stimuli, that is, sets of 108 words produced by four speakers (two female). The blocks, as well as the items within each block, were presented in random order. Each speaker voice was heard approximately the same number of times across the participants in the perception study.

Since tone height is relative and to some extent speaker-dependent, the participants in the perception study were familiarized with the voices they heard in the experiment. That is, they listened to two dialogues produced by the speaker of each block prior to the experimental trials in that block.

Each trial started with a fixation cross for 500 ms. Then a syllable carrying one of four tones was played twice with an interval of 250 ms. At the onset of the second repetition, four Chinese characters corresponding to the segments of the syllable, but carrying different tones, appeared on the screen. Participants were instructed to press one of the four keys associated with the position of the four characters to indicate which word (character) they thought they heard. The positions of the characters were randomized. The experiment took about 50 minutes to complete.

[1] We considered the tone of an adjacent syllable to be congruent to that of the target, if the syllable's onset/offset F0 value agrees, to the extent possible, with the offset/onset F0 value of the tone of the target syllable. For instance, a target syllable with T1 (i.e., beginning and ending with a high F0) could be preceded by T1 or T2, both of which end with a high F0 value, and it could be followed by T1 or T4, both of which begin with a high F0. Analogously, T2 could be preceded by T4 and followed by T1 or T4; T3 could be preceded by T1 or T2 and followed by T1 or T4; T4 could be preceded by T1 or T2 and followed by T2. Given the complexity of T3, it did not appear adjacent to any of the target syllables.

The accuracy of each response was recorded. Trials without a response input were excluded from the analysis. The excluded trials constitute 0.1% of the total trials.

4 Results

4.1 Accuracy Measure

To measure how well T3 and T4 were perceived, we first determined the percentage of correct responses. As can be seen in Figure 1, the overall accuracy (i.e., correct selection) for T3 is only 49%, while for T4 it is 72%; (chance = 25%). The accuracy rates for T1 and T2 selection are also shown, to provide perspective on the T3 and T4 results.

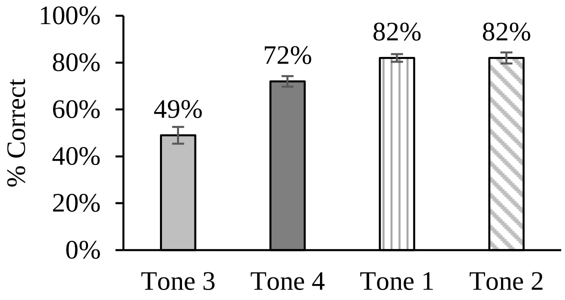

Figure 1: Percentage of correct responses for each tone.

We constructed a generalized linear mixed effects model to analyze the response accuracy, using the function *glmer()* from the *lme4* package (Bates et al. 2015) in R. Models were compared with the *anova()* function. The only fixed factor of the current analysis is Tone (T1, T2, T3, and T4). Starting with the maximal random effects structure, the model converged when it included Participant, Block, and Syllable as random intercepts. The converged model was then compared to a baseline model where the fixed factor Tone was removed. The results suggested a main effect of Tone on the accuracy [χ^2 (3) = 364.62, p < .001]. Planned contrasts revealed a significantly better overall performance for T1 and T2 than for T3 and T4 (β = −1.82, SE = 0.14, z = −12.87, p < .001). There was also a significant difference in accuracy between T3 and T4 (β = 0.86, SE = 0.07, z = 12.59, p < .001). These results suggested that although both T3 and T4 were more prone to confusion than T1 and T2, correctly identifying T3 was especially problematic.

4.2 Distribution of Incorrect Responses

To gain further insight into the tonal perception, we examined the distribution of incorrect responses for each tone. As shown in Figure 2, T4 syllables were mistakenly identified as T3 only 8% of the time, suggesting that T4 was not easily confused with T3. By contrast, more than one-third of T3 syllables were perceived as T4. This pattern suggested that although there was a confusion between T3 and T4, the confusion was mostly driven by T3 being perceived as T4, but not vice versa. More importantly, the pattern indicated a bias in favor of selecting T4 when T3 was not correctly identified. Thus, the higher overall accuracy rate we observed for T4 than T3 could be attributed to a T4 bias rather than a better perception of T4 than T3. To control for the bias, we additionally conducted d' analyses of the responses for T3 and T4, and for the comparison tones, T1 and T2.

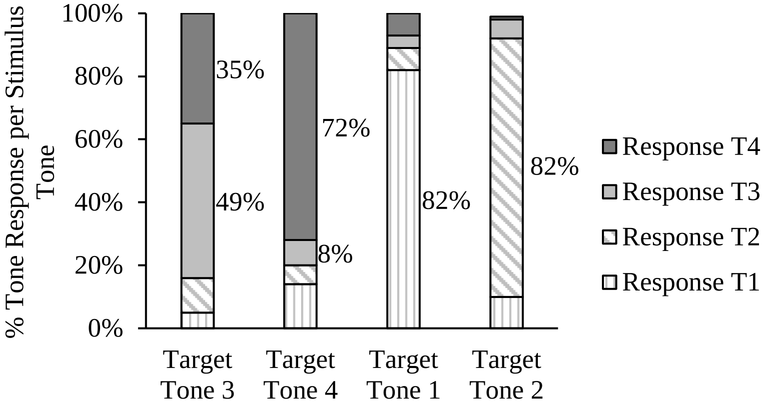

Figure 2: Distribution of response for each tone.

4.3 d’ Measure

The d’ statistic measures the sensitivity to a signal, in this case, taking into account not only how well participants identified a tone when the tone was present (Hit), but also how well participants determined that a tone was not present (Correct rejection). Table 1 shows the scheme we used to calculate the d-prime scores for T3 and T4 separately, as well as for Tones 1 and 2. The d’ scores for T3 and T4 are presented in Figure 1, along with the scores for T1 and T2, for comparison.

	Respond as target tone	Respond as non-target tones
Target Tone	Hit	Miss
Non-target tone	False alarm	Correct rejection

Table 1: Calculation scheme for d’ scores for each target tone.

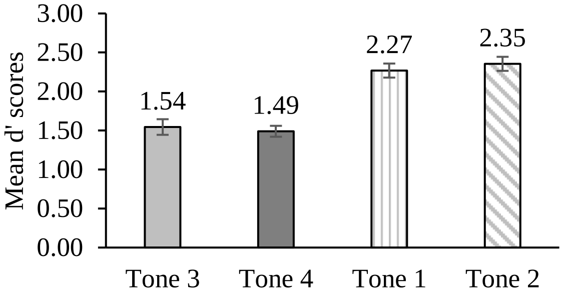

Figure 3: Mean d’ scores for each tone.

We also conducted a linear mixed effects model to analyze d’ scores. The model included Tone as a fixed factor and Participant as a random intercept. The model was then compared to a baseline model with the fixed factor Tone removed. Again, the result suggested a main effect of Tone on d’ scores [χ^2 (3) = 81.08, p < .001]. Planned contrasts also revealed a significantly better overall performance for T1 and T2 than for T3 and T4 (β = −1.62, SE = 0.17, t = −9.30, p < .001). By contrast, no difference was found between T3 and T4 (p = .14), a pattern that differs from the pattern that emerged when we measured the response accuracy. This result indicated that after the T4 bias was controlled for, the perception of T4 was no longer any better than the perception of T3.

4.4 Effects of Focus and Syllable Position

To assess whether certain prosodic structures affected the clarity of the distinction between T3 and T4, we compared the perception of syllables extracted from the focus and non-focus contexts, and also from each of the three syllable positions in words recorded for the original production corpus. We thus added Focus (focus and non-focus) and Syllable Position (Syllable 1, Syllable 2, and Syllable 3) as two fixed factors to the model we previously built for analyzing the effect of Tone on d' scores, resulting in a model with Tone, Focus, and Syllable Position as fixed factors. The model converged when we included Participant as the random intercept. No interaction was significant, nor was the effect of tone. Figure 5 shows the similar patterns that emerged for T3 and T4.

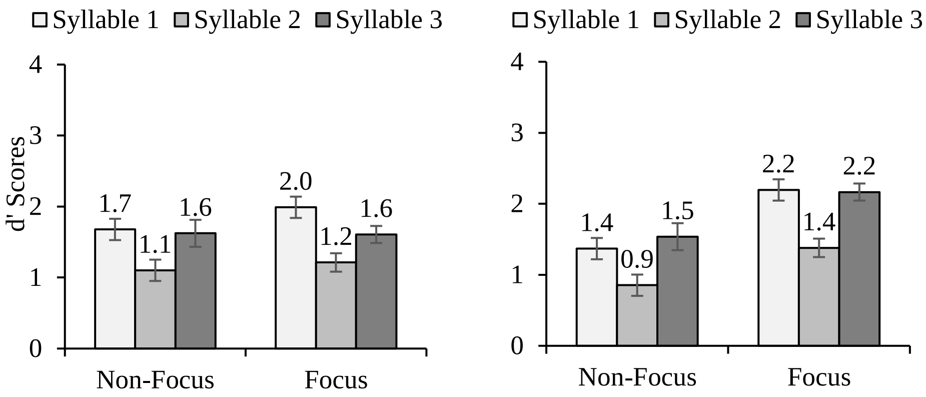

Figure 4: T3 (left panel) and T4 (right panel) d' scores by Syllable and Focus.

Specifically, we observed a main effect of Focus [χ^2 (1) = 65.47, p < .001], with higher d' scores in the focus condition than in the non-focus condition. There was also a main effect of Syllable Position [χ^2 (2) = 96.58, p < .001]. Planned contrasts revealed lower d' scores in Syllable 2 than in Syllable 1 and Syllable 3 (β = −0.90, SE = 0.14, t = −6.35, p < .001), but no difference was found between Syllable 1 and Syllable 3 (p = .38).

5 Discussion

Although the descriptions of T3 and T4, when produced in isolation, are quite distinct (i.e., dipping vs. falling), connected speech is known to affect their production, so they are no longer necessarily as distinct. To assess the effects of different aspects of connected speech (i.e., focus and syllable position in a word) on the perceptual distinction between T3 and T4, the present study tested the perception of these tones, and T1 and T2 for comparison, in a forced-choice paradigm. Both percent accuracy of tone identification and d' scores were used as the dependent measure, and we found that the two different measures provided different insights into the perception of the four Mandarin tones, and in particular, T3 and T4. In terms of the percent accuracy, T3 exhibited more perceptual errors than T4. The error distribution pattern of the two tones, however, revealed that more than one-third of the syllables with T3 were perceived as having T4, while T4 syllables were rarely perceived as having T3. Since the confusability between T3 and T4 was not reciprocal, Hypothesis 1 (T3 tends to be confused with T4, and vice versa.) was only partially confirmed. Given previous observations that the rising part of T3 tends to be reduced in connected speech (e.g., Gårding 1987), it seems likely that the similarity of the falling pitch contour at the beginning of both T3 and T4 is the acoustic property that the participants most readily detected.

We also found that both T3 and T4 exhibited more errors than the comparison tones, T1 and T2. This finding was not consistent with Xu 1994, where the identification of both T2 and T4 was highly accurate. One reason for the difference could be that Xu 1994 used stimuli extracted from three-syllable compounds produced in isolation, while the current stimuli were extracted from three-syllable compounds embedded in carrier dialogues, where the items may have been somewhat less carefully realized by the speakers. This difference also draws attention to the possibility that testing the perception of tones on syllables, compounds, or even phrases, produced in isolation may not accurately reflect how native speakers perceive tones in connected speech.

Although the percent accuracy and the error distribution pattern revealed confusability between T3 and T4, the strong preference for a T4 response when T3 was not correctly perceived suggested that a sensitivity measure was more appropriate than just correct responses in interpreting the perception patterns of T3 and T4. In fact, a d' analysis revealed a different perceptual pattern. That is, the participants' sensitivity to T3 and T4 was essentially at the same level. Moreover, since the comparison tones, T1 and T2, showed higher d-prime scores, indicating that their manifestations are clear, we suggest that listeners must make additional use of contextual information more in differentiating T3 and T4 in connected speech.

With regard to the effect of prosodic structure on the perceptibility, our continued use of the d' analysis revealed a significant effect of both Syllable Position and Focus. In support of Hypothesis 2 (better perception in a focus context than in a non-focus context.), the sensitivity to both T3 and T4 was higher in the focus condition than in the non-focus condition. This finding is consistent with the frequently observed acoustic enhancing effect of focus (e.g., Chen and Braun 2006, Chen and Gussenhoven 2008, Ouyang and Kaiser 2015, Lee, Wang, and Liberman 2018), which in turn, leads to better perception. We also found support for Hypothesis 3 (better perception at the edge of a word than the middle of a word), given the greater sensitivity for both T3 and T4 in the peripheral syllables (Syllable 1 and Syllable 3), as opposed to the middle syllable (Syllable 2). This finding is consistent with Liu 1989's results of automatic recognition of the tones in trisyllabic words produced by native Mandarin speakers. Specifically, it was found that the recognition model based on multi-syllabic words yielded higher recognition rates for Syllable 1 and 3 than for Syllable 2. Along the same line, Zhang et al. 2018 observed that speakers spent more effort to clearly produce tones at prosodic word/phrase boundary, with the consequence that a tone produced at the prosodic boundary exhibited an F0 contour more similar to that of the same tone produced in isolation, compared to the tones produced in a non-boundary position.

Finally, it must be noted that we did not find a main effect of Tone or an interaction between Tone and Focus, which means that the conclusion that T3 and T4 exhibited the same level of sensitivity holds in both the focus and non-focus conditions. This appears to contradict Lee, Wang, and Liberman 2016's finding that T3 was less well identified than T4 in the focus condition. The difference could be attributed to the use of different dependent measures. While Lee, Wang, and Liberman 2016 measured accuracy, the current study measured d' scores, which took into account the error patterns of T3. In fact, even in the focus condition, there were still 37% of T3 perceived as T4. Thus, although the acoustic enhancing effect of focus led to better a perception of a focused T3 than a non-focused T3 (main effect of Focus), it barely attenuated the bias favoring a T4 response. We suggest that the limited effect of focus on T3 may be explained by T3's relatively small capacity for pitch range expansion. That is, since the pitch contour of T3 already extends to the lowest portion of the pitch range (Cao 2012, Lee, Wang, and Liberman, 2018), there is minimal room for further expansion. Indeed, the acoustic analysis of the current stimuli revealed a T4-like falling contour of T3, with minimal trough and rising contour, even in the focus condition (Athanasopoulou et al. 2019, Vogel et al. 2019).

6 Conclusion

The current study investigated the perceptibility of Mandarin tones produced in connected speech, focusing primarily on T3 and T4. When we considered percent accuracy (i.e., successful tone identification), we found that T4 was much more successfully identified than T3, the latter being perceived as T4 more than one-third of the time. Thus, our first hypothesis about the mutual confusability of T3 and T4 was partially confirmed, since T3 was frequently perceived as T4, but not vice versa. When we took into account the bias favoring a T4 response using a d' analysis, the advantage for T4 was no longer observed, suggesting that perceiving T4 was just as difficult as perceiving T3. Moreover, we found that both focus and syllable position affected the tonal perception. Improved perceptibility of both T3 and T4 was observed when a tone was produced in a focus context or at the edge of a word. In sum, the current findings confirmed our hypotheses about the confusability of T3 and T4 and the mitigating effects of the prosodic factors of focus and syllable position in a word. We conclude, furthermore, that caution must be exercised in interpreting accuracy rates in perception studies of Mandarin tones, given the observed presence of a response bias.

References

Athanasopoulou, Angeliki, and Irene Vogel. In preparation. The acoustic properties of Mandarin tones: effects of focus and syllable position.

Athanasopoulou, Angeliki, lrene Vogel, Chao Han, and Yue Yuan. 2019. Confusability of Mandarin tone 3 and tone 4: Effects of focus and syllable position. In *Proceedings of the 19th International Congress of Phonetic Sciences,* ed. S. Calhoun, P. Escudero, M. Tabain, and P. Warren, 442-446

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48.

Cao, Rui. 2012. Perception of Mandarin Chinese Tone 2/Tone 3 and the Role of Creaky Voice. Doctoral dissertation, University of Florida.

Chen, Yiya, and Bettina Braun. 2006. Prosodic realization of information structure categories in standard Chinese. *Proceedings of Speech Prosody* 54: 5–8.

Chen, Yiya, and Carlos Gussenhoven. 2008. Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics* 36: 724–746.

Gårding, Eva, Paul Kratochvil, Jan-Olof Svantesson, and Jialu Zhang. 1986. Tone 4 and tone 3 discrimination in modern Standard Chinese. *Language and Speech* 29: 281–293.

Gårding, Eva. 1987. Speech act and tonal pattern in Standard Chinese: constancy and variation. *Phonetica* 44: 13-29.

Lee, Yong-Cheol, Ting Wang, and Mark Liberman. 2016. Production and perception of tone 3 focus in Mandarin Chinese. *Frontiers in Psychology* 7: 1–13.

Liu, Lih-Cherng, Wu-ji Yang, Hsiao-Chuan Wang, and Yueh-Chin Chang. 1989. Tone recognition of polysyllabic words in Mandarin speech. *Computer Speech and Language* 3: 253–264.

Nespor, Marina, and Irene Vogel. 1986. *Prosodic phonology*. Dordrecht: Foris.

Shen, Xiaonan Susan, and Lin Maocan. 1991. A perceptual study of Mandarin tones 2 and 3. *Language and Speech,* 34: 145-156.

Shih, Chilin. 2005. Understanding phonology by phonetic implementation. In *INTERSPEECH 2005*, 2469-2472 .

Ouyang, Iris Chuoying, and Elsi Kaiser. 2015. Prosody and information structure in a tone language: an investigation of Mandarin Chinese. *Language, Cognition and Neuroscience* 30: 57–72.

Vogel, Irene, Angeliki Athanasopoulou, Chao Han, and Yue Yuan. 2019. How perceptible is the difference between tone 3 and tone 4 Mandarin? In *Proceedings of the 19th International Congress of Phonetic Sciences,* ed. S. Calhoun, P. Escudero, M. Tabain, and P. Warren, 2027-2031.

Xu, Yi. 1994. Production and perception of coarticulated tones. *Journal of the Acoustical Society of America* 95: 2240–2253.

Xu, Yi. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics* 25: 61–83.

Yang, Jing, Zhang Yu, Li Aijun, and Li Xu. 2017. On the duration of Mandarin tones. In *INTERSPEECH 2017*, 1407–1411.

Yang, Bei. 2015. *Perception and production of Mandarin tones by native speakers and L2 learners*. Springer Berlin Heidelberg.

Yang, Chunsheng. 2016. *The acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice*. John Benjamins Publishing Company.

Zhang, Wei, Lixia Hao, Yanlu Xie, and Jinsong Zhang. 2017. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2017*, 926–930.

Department of Linguistics
University of Delaware
Newark, DE 19716
hanchao@udel.edu
ivogel@udel.edu
yueyuan@udel.edu
angeliki.athanasopou@ucalgary.ca