

# **Non-segmental Cues for Syllable Perception: the Role of Local Tonal $f_0$ and Global Speech Rate in Syllabification<sup>1</sup>**

**Iris Chuoying Ouyang\***

## **Abstract**

This study is aimed at a better understanding of the perception of syllables. As the traditional view seems to associate syllable perception with segmental cues that result from local (*i.e.* present only within or adjacent to the syllable) supralaryngeal events, we are particularly interested in whether non-segmental and non-local laryngeal information contribute to syllable perception as well. Existing works on Indo-European languages show that local stress patterns and global (*i.e.* non-local) speech rates provide perceptual cues to words and phonemes. While we believe that the effects of the global speech rate hold across languages, based on the long-developed notion of language-specific perception, we expect that lexical tones, rather than stress patterns, serve as an important local non-segmental cue in tonal languages. We conducted a perception study on Mandarin to investigate whether tonal  $f_0$  patterns and speech rates interfere with spectral information in determining the number of syllables in an utterance.  $F_0$  contours were generated using the qTA model (Prom-on, Xu & Thipakorn, 2009). Our results show that the perceptual number of syllables depends on the perception of tonal  $f_0$  patterns and speech rates to a substantial extent. Combining our findings with prior claims (Olsberg, Xu & Green, 2007), it appears that a variety of cues – segments, lexical tones, and speech rate – compete in perceiving Mandarin syllables. In relating this study to the existing works on word segmentation, lexical access, and phoneme identification,

---

<sup>1</sup> Earlier versions of this work were presented at LabPhon (Conference on Lab Phonology) 13 and InterSpeech (Conference of the International Speech Communication Association) 2012; I would like to thank the audience members for the valuable comments and suggestions. Thanks also go to the USC Phonetics Lab group for feedback during the development of this project. I would also like to thank the editors of 'International Journal of Computational Linguistics and Chinese Language Processing' and two anonymous reviewers for their helpful comments.

\* Department of Linguistics, University of Southern California, U.S.A.  
E-mail: chuoyino@usc.edu

we find that the language comprehension system integrates local with global, supralaryngeal with laryngeal information, in perceiving linguistic units – not only words and phonemes, but also syllables.

**Keywords:** Cue Integration, Syllable Perception, Tone Perception, Speech Rate, Mandarin Chinese

## 1. Introduction

The phonetic cues that determine the perception of phonological entities constitute one of the central issues in the research on speech processing. The syllable, a fundamental phonological unit across languages, is generally thought to be perceived via segmental cues and phonotactics. In other words, syllable identification and syllabification in a given language usually are believed to be determined by supralaryngeal *and* local (*i.e.* temporally only present within or adjacent to the syllable in question) information. Local yet laryngeal information such as rhythmic patterns, or global (*i.e.* non-local, temporally present throughout a larger utterance that includes the syllable in question) cues such as speech rate, have not been considered important factors in syllable perception to our knowledge. Nevertheless, existing studies on word perception have shown that listeners use both local and global prosodic cues in word segmentation and/or lexical access (*e.g.* local rhythmic pattern: Nakatani & Schaffer, 1978; Mattys, Jusczyk, Luce & Morgan, 1999; local prosodic boundary: Christophe, Peperkamp, Pallier, Block & Mehler, 2004; Gout, Christophe & Morgan, 2004; global rhythmic grouping: Dilley & McAuley, 2008; Brown, Salverda, Dilley, Laura & Tanenhaus, 2011). Moreover, phoneme identification also has been found to be sensitive to both local and distant rhythmic patterns (*e.g.* Shields, McHugh & Martin, 1974; Pitt & Samuel, 1990), as well as global speech rate (*e.g.* Summerfield, 1975; Miller & Grosjean, 1981). Based on these findings, it is reasonable to speculate that local and global laryngeal information are involved in the perception of syllables, just as they are in the perception of phonemes and words.

Mandarin Chinese may be one of the languages where the syllable plays a particularly important role. In addition to the argument that it may be a syllable-timed language (*e.g.* Lin & Wang, 2007), Mandarin is known for its close relationship between syllables, morphemes and lexical tones, as most morphemes consist of one syllable and *vice-versa* (*e.g.* DeFrancis, 1984) and every syllable is temporally aligned with a lexical tone and *vice-versa* (*e.g.* Xu, 1998; Gao, 2008). As the syllable appears to have a special status in Mandarin, we conducted this study on Mandarin to investigate non-segmental cues for syllables and hoped to improve our understanding of syllable perception in general.

The next question that naturally arises is what kinds of non-segmental information might affect syllable perception. Prior work on word perception extensively has examined rhythmic representations, such as stress and stress-based prosodic grouping in Indo-European languages

(see the beginning of Section 1 for relevant citations). Nevertheless, the mappings between phonetic representations and phonological categories are language-dependent (*e.g.* Caramazza & Yeni-Komshian, 1974); what counts as perceptual cues for syllables may vary depending on the characteristics of a given language. For tonal languages, where pitch patterns provide contrastive information that differentiates one word from another, we consider pitch as a potential factor in segmenting and identifying linguistic units. Given the fixed timing relationship between syllables and tones in Mandarin (*e.g.* Xu, 1998; Gao, 2008), we expect lexical pitch patterns to serve as a local cue and interact with another local cue – segments – in perceiving Mandarin syllables.

Degradation of segmental information has been shown to influence the perception of tones and syllables in Mandarin. Olsberg, Xu, and Green (2007) found that, when high-frequency spectral information was removed from a sequence of [ma] syllables, listeners could not accurately identify either the number of syllables or the categories of tones in the sequence, despite the presence of intact f0 information. What remains unclear is whether the effect goes both ways: does tonal information impact the perception of segments and syllables as well? Particularly, can lexical f0 cues override formant patterns in determining the number of syllables?

Thus far, we have discussed tonal f0 as providing local laryngeal cues for syllable perception. Let us now consider global laryngeal cues. Prior work on phoneme identification has examined speech rate and found that it influences the expectation of VOT length (*e.g.* Summerfield, 1975; Miller & Grosjean, 1981). Since speech rate is not a language-specific phenomenon, we expect to see a similar effect in Mandarin. Global speech rate should affect expected length of linguistic units – in our case, the expected duration of syllables – which should in turn affect the perceptual syllable count within a given time window.

## **1.1 Aims of this Study**

In this paper, we report a perception study that investigates whether local and global non-segmental information influence syllable perception. As discussed in the preceding paragraphs, existing studies that have looked into the role of non-segmental information in segmenting and identifying speech units mostly have focused on phonemes and words. The syllable, another fundamental speech unit, has not been brought into this conversation. To shed light on this issue, we examined the effects of lexical pitch contours<sup>2</sup> and overall speech

---

<sup>2</sup> In this paper, the terms ‘lexical pitch/f0 patterns’, ‘lexical pitch/f0 contours’, ‘tonal pitch/f0 patterns’, and ‘tonal pitch/f0 contours’ are used interchangeably. We avoid using simply ‘tonal (cues)’ to refer to lexical f0 cues, because tones also involve other acoustic dimensions, such as amplitude (*e.g.*

rate on the perception of syllable numbers. We asked if changes in either lexical pitch contours or overall speech rate, without changes in formant patterns, alter the count of syllables in an utterance. If so, to what extent can they impact syllabification?

## 1.2 Background: Mandarin Tones

There are four lexical tones in Mandarin: high (Tone 1), rising (Tone 2), low (Tone 3), and falling (Tone 4). They distinguish lexical items from one another, as illustrated in (1).

(1) Tone 1	ma [High]	‘mother’	
	Tone 2	ma [Rising]	‘hemp’
	Tone 3	ma [Low]	‘horse’
	Tone 4	ma [Falling]	‘scold’

The simplicity of the tone system in Mandarin allows us to examine how lexical pitch patterns affect a listener’s judgment of syllable numbers in a natural way. Crucially, when two level tones (*i.e.* High and Low) are adjacent to each other in continuous speech, they form a pitch pattern that, shape-wise, looks similar to a contour tone (*i.e.* Rising or Falling), although typically different from the contour tone in f0 onsets, offsets, ranges, the turning points of f0 movement, *etc.* (*e.g.* Shih, 1986; Shen, 1990). This opens up possibilities where two words consist of different numbers of syllables yet minimally differ from each other in segments and tones to an extent that they may ‘sound similar’. For example, a bisyllabic word with two different level tones (*i.e.* High-Low or Low-High) out of context can potentially be perceived as a monosyllabic word with a contour tone, if there is no consonant at the syllable boundary, *e.g.* [CV.VC] → [CVVC]. Thus, we were able to use real words in the experiment, where the stimuli only differed in tonal f0 contours while retaining a chance of being identified as either bisyllabic or monosyllabic.

## 2. Perception Study: Method

Participants performed an identification task, where they heard a target word in isolation or embedded in a carrier sentence, saw pictures on a computer screen, and determined which picture matched the word they heard. Using pictures allowed us to avoid presenting participants with written words, which might carry additional phonetic information that is not in the aural stimuli.

---

Whalen & Xu, 1992, “Information for Mandarin tones in the amplitude contour and in brief segments”, *Phonetica*, 49, 25-47) and duration, in addition to f0.

We focused on two factors that were expected to influence the perceptual syllable count of a word: the local tonal pitch contour and the global speech rate. In this study, the term ‘local’ refers to cues that only occur within the target word region of a sentence, such as the tonal pitch contour carried by the target word; whereas ‘global’ refers to cues that occur throughout a sentence that includes a target word, such as the overall speech rate of a sentence. In the following subsections, we will first go over the experiment design and procedures before discussing the hypothesis and predictions.

## 2.1 Design and Stimuli

Participants heard target words or sentences containing a target word one at a time. To investigate whether syllabification depends on local tonal pitch contours and global speech rate, we manipulated the pitch contour in a target word and the speech rate of its carrier sentence while controlling the segments. Specifically, a repeated-measures within-subjects design with two independent variables was used: (i) **‘monotone-ness’ of the  $f_0$  contour in a word** (with six steps, on a continuum between a bitonal sequence to a monotonal sequence) and (ii) **absence or the average syllable length of a carrier sentence** (with four levels: fast, medium, slow, and no carrier).

### 2.1.1 Target Words

All target words originally consisted of two syllables, as words are predominately bisyllabic in Mandarin. The options of words were limited to the sequences whose tonal and vocalic properties allowed ambiguity about the number of syllables in a word, as discussed in Section 1.1. We used bisyllabic words of which the two syllables had different level tones, namely, Low-High or High-Low, given that the  $f_0$  contour of a Low-High tonal sequence (LH) appears similar to a Rising tone (R), and the contour of a High-Low tonal sequence (HL) appears similar to a Falling tone (F). To examine syllable perception across different types of syllable structure, we used bisyllabic words that contained one of the three vocalic sequences, with the same vowel [u] at both sides of the syllable boundary: [u.u], [Vu.u], and [u.uV] (in which V refers to a vowel other than [u]). Thus, for a target word to be identified as monosyllabic, besides perception as a single tone (*i.e.* R or F), the two [u] segments had to be ‘misperceived’ as one segment. In addition, for the words with a diphthong [Vu] or [uV], the perceptually merged [u] had to be ‘misperceived’ as part of the diphthong. There were six target words, each in one of the two tonal sequences and one of the three vocalic sequences. (See Appendix A for the list of bisyllabic target words and Appendix B for the list of their monosyllabic alternatives.) Due to the limitations regarding tonal and vocalic properties on the selection of target words, the identity of segments in the target words was not controlled across tonal and vocalic sequences. These limitations also made it difficult to control lexical properties of

target words, such as word frequency and word class, which might raise concerns about whether such kinds of differences between bisyllabic target words and their monosyllabic alternatives could have impacted our data. Nevertheless, as the goal of this study was not to compare utterances of different numbers of syllables, but rather to compare segmental material occurring with different tonal f<sub>0</sub> at different speech rates, an imbalance in lexical properties between the two members of a bisyllabic-monosyllabic word pair (e.g. [tsu.u] - [tsu]) should not distort our results. For example, even if the differences in word frequency or word class between [tsu.u] and [tsu] had biased participants towards [tsu], such bias had probably existed across different levels of tonal f<sub>0</sub> and speech rate for this word pair. In other words, response tendencies associated with particular word pairs should not affect data patterns with respect to the effects of tonal f<sub>0</sub> and speech rate. Thus, although it would have been ideal for lexical properties to be controlled, we do not regard this as a problem for our findings.

### 2.1.2 Carrier Sentences and Speech Rate

In one-fourth of the trials, target words were played in isolation; in the other three-fourths of the trials, target words were embedded in a sentence frame, as illustrated in (2). Sentence frames were at one of three different speech rates: fast, medium, or slow. The use of carrier sentences served two major purposes. First, it enabled us to examine the effect of tonal f<sub>0</sub> across different speech rates while keeping the duration of target words constant. In other words, we manipulated global speech rate by presenting the same target item with sentence frames that differed in speech rate. Second, it also allowed us to vary the f<sub>0</sub> excursion of target words over a wider range and test more levels of f<sub>0</sub> contours. As the content of sentence frames were identical throughout the experiment, target words were the only new information in a sentence. Existing works on information structure and discourse-level intonation show that new information is produced with larger f<sub>0</sub> ranges than given information in Mandarin (Chen & Braun, 2006; Ouyang & Kaiser, 2013). By placing target words in sentence frames, we were able to use more extreme f<sub>0</sub> values in target words, *i.e.* higher f<sub>0</sub> for high tone targets and lower f<sub>0</sub> for low tone targets, without compromising the naturalness of sentence intonations.

- (2) wo            ba        TARGET shuo-le            san-ci  
       PRO.1st.sg BA        TARGET say-PERP        three-time  
       *'I said TARGET three times'*

A male speaker of Beijing Mandarin recorded target words in isolation at a medium speech rate (254 ms/syllable) and sentence frames at three different speech rates: fast (131

*the Role of Local Tonal f0 and Global Speech Rate in Syllabification*

ms/syllable), medium (259.5 ms/syllable), and slow (441.5 ms/syllable). To obtain natural-sounding tonal coarticulation between a target word and its adjacent words for the sentence stimuli, sentence frames for LH and HL tonal sequences were recorded separately, with a target word embedded as a placeholder. Among the target words in each tonal sequence, the one with the simplest vocalic structure, *i.e.* [tsu.u] in LH and [tʂ<sup>h</sup>u.u] in HL, was used for recording sentence frames. The speaker was first told to produce all materials at the rate he naturally spoke at, and then to produce the sentences at a rate faster and another rate slower than the first one. Sentence stimuli were later made by replacing the placeholder word in each recording with target words whose f0 had been altered synthetically. There was no apparent pause between a placeholder word and its adjacent words in the original recordings, and no pause was added during the sentence synthesis.

To make sure that embedding f0-altered words in naturally-produced sentence frames did not create artifact effects, we included the conditions of isolated words to be compared against the conditions of medium speech rate. Since target words and medium-rate sentence frames were both recorded at the speaker's natural speaking rate, target words presented in isolation should not yield different results from those flanked by the medium-rate sentence frames.

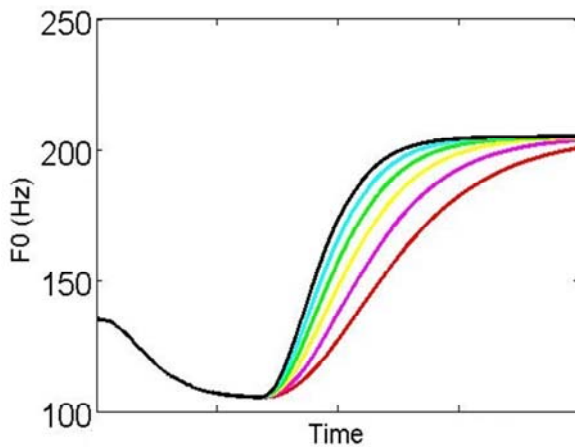
### 2.1.3 Tonal f0 Synthesis

To systematically vary tonal f0 patterns in target words, we generated f0 contours via the qTA model (Prom-on, Xu & Thipakorn, 2009), and synthesized them with natural words using the PSOLA (Pitch Synchronous Overlap Add) method implemented in the Praat software. Synthesized pitch patterns were on a continuum between an f0 contour that indicated a two-tone sequence and one that suggested a single tone (*i.e.* LH-R and HL-F). The qTA model was configured as a third-order linear system, where the level, velocity, and acceleration of f0 at a given time point depend on those at the previous time point. There were three parameters in the model: f0 target slope  $m$ , f0 target height  $b$ , and the rate at which f0 changes  $\lambda$ . In addition, there was an initial f0 state consisting of an initial f0 level  $f_0(0)$ , initial f0 velocity  $f'_0(0)$ , and initial f0 acceleration  $f''_0(0)$ . F0 contours were generated using formulae for one of the two-tone sequences: LH and HL. We kept all parameters in the qTA model fixed and only manipulated  $\lambda$  in the second tone of a sequence, namely the High tone in LH and the Low tone in HL, as shown in Table 1. Since  $\lambda$  corresponds to the speed at which f0 rises or falls, it determines whether and when an f0 contour arrives at its target: the f0 target may be reached later or eventually undershot when  $\lambda$  is low. As  $\lambda$  in the first tone of a sequence was set at a value such that the first tone always reached its (fixed) target, we effectively only varied the f0 contour during the second tone. This choice was motivated by two reasons. *First*, since f0 offsets have been suggested to play an important role in the perception of tones (see Xu, 1997 for relevant discussion), altering the latter part of an f0 contour might be a way of evoking

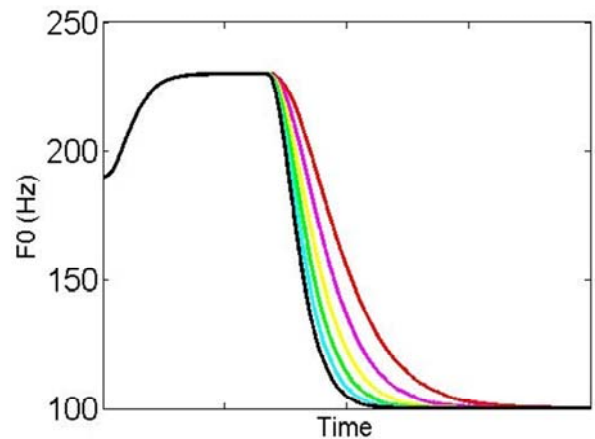
different tone perceptions with minimal manipulation. Pre-empting our findings somewhat, the  $f_0$  contour in the second tone of a two-tone sequence did influence the perceptual number of tones that listeners heard. *Second*, for tonal  $f_0$  steps that favored a two-tone perception, this manipulation simulated the unequal variability of  $f_0$  contours between two consecutive tones in real data. Prior work on Mandarin tones indicates that the  $f_0$  contour of a two-tone sequence varies substantially more in the first tone than in the second tone, due to an asymmetry between carryover and anticipatory tonal coarticulation (Xu, 1997).

**Table 1. Parameter settings of  $f_0$  synthesis using the qTA model**

	Low-High		High-Low	
	First tone	Second tone	First tone	Second tone
$m$	0	0	0	0
$b$	105	205	230	100
$\lambda$	80	65, 80, 95, 110, 125, 140	200	120, 150, 180, 210, 240, 270
$f_0$	135	-	190	-
$f'_0$	0	-	0	-
$f''_0$	0	-	0	-



**Figure 1.  $f_0$  contours for the Low-High target words**



**Figure 2.  $f_0$  contours for the High-Low target words**

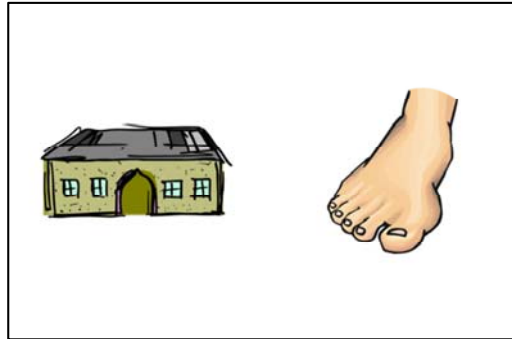


It is important to note that, since the tonal sequences used in this study contained two level tones rather than including any contour tones, the tone boundary (*i.e.* where the second tone began) specified in a formula matched the turning point of the f0 contour (*i.e.* the point where f0 started rising or falling for the second tone) it generated. This, to a large extent, resembled real f0 data in Mandarin: the f0 turning point in a two-tone sequence was near the tone boundary when the two tones were in an LH or HL combination (Xu, 1997, Figure 3, Tone 3-1 and Tone 1-3). Essentially, by changing the  $\lambda$  value of the second tone in a formula, we manipulated the part of an f0 contour after the turning point. With a high  $\lambda$  value after the turning point, *i.e.* the higher lines in Figure 1 and the lower ones in Figure 2, an f0 contour sounded more like a sequence of two tones; whereas, when the  $\lambda$  value was low after the turning point, *i.e.* the lower lines in Figure 1 and the higher ones in Figure 2, the entire f0 contour sounded more like one single tone. We aligned the turning point of a synthesized f0 contour with the turning point of the original f0 contour that was produced by the speaker in each bisyllabic target word. An original turning point was defined as f0 minimum for an LH contour and f0 maximum for an HL contour in natural words. Thus, the timing relationship between segments and pitch patterns was preserved to some extent.

A number of six  $\lambda$ 's (hence, six tonal f0 steps) was used for each tonal sequence due to reasons of statistical power and experiment length. F0 target heights  $b$ 's were determined such that f0 values fell between 105-205 Hz in LH target words and 100-230 Hz in HL target words. These tonal f0 targets were reasonable given the f0 ranges of sentence frames produced by the speaker: the average f0 minimum and maximum of sentence frames were 119 and 204 Hz for the LH conditions and 120 and 207 Hz for the HL conditions.

### **2.1.4 Identification Task**

The main experiment was a two-alternative forced-choice identification test (2AFC). Participants saw two pictures side by side on a computer screen while listening to a target word in isolation or flanked by a carrier sentence through headphones. As illustrated in Figure 3, one of the pictures represented the bisyllabic target word that was being played (*e.g.* [tsu.u] in the LH tone), and the other represented a monosyllabic word that participants could potentially hear in the case of 'misperception' induced by our manipulation of tonal f0 and speech rate (*e.g.* [tsu] in the R tone). Participants were asked to choose the picture that matched what they heard. (See Appendix A and B for the pictures used for each word.) They were told that the study was interested in how people recognize words.



**Figure 3. Sample display**

Before the main experiment, participants completed a familiarization phase where they learned the names of the pictures (*i.e.* the words represented by the pictures) used in the main experiment. The familiarization phase consisted of two parts. In the first part, each picture was shown on a computer screen one by one with the word it represented underneath. Participants read aloud and memorized the names of the pictures. In the second part, each picture was shown with two words on top of it, one being the word it represented and the other being its monosyllabic or bisyllabic counterpart, *e.g.* the picture of a foot with the words for ‘foot’ ([tsu]) and ‘ancestral house’ ([tsu.u]). Participants were asked to pick the name of the picture from the two words; none of them made any mistakes in this task.

Every participant responded to four repetitions of an item. There were 144 items (6 target words \* 6 levels of f<sub>0</sub> contours \* 4 levels of speech rate) and 576 trials in total. The pictures were counter-balanced for the left and right positions across trials. The dependent variable was the percentage of monosyllabic-word responses, namely, the percentage of trials where participants chose the picture representing the monosyllabic alternative rather than the one representing the bisyllabic word (*e.g.* [tsu] instead of [tsu.u]).

## 2.2 Participants

Twenty-seven native speakers of Mandarin who were born in China participated in the study. The experiment was conducted in Los Angeles, California, USA. All of the participants were between 24 to 44 years old and had left China no earlier than the age of 22. Each participant received 10 U.S. dollars for their participation.

## 2.3 Hypothesis and Predictions

As discussed in Section 1, the perception of linguistic entities, such as words and phonemes, has been shown to depend on local and on global laryngeal information. Along this line, we believe that syllables should not be an exception; syllabification should also be sensitive to both local and global laryngeal cues. This hypothesis was tested with Mandarin Chinese, a

tonal language where lexical pitch patterns and syllables have a fixed timing relationship. We investigated whether Mandarin syllabification is affected by lexical pitch patterns – a local laryngeal factor that indicates the number of syllables in Mandarin – and speech rate – a global laryngeal factor that may crosslinguistically suggest average syllable duration. Our general prediction was that the change of lexical pitch patterns and speech rates both would impact the perception of the number of syllables in Mandarin substantially, regardless of segmental information. Specifically, we expected that the percentages of monosyllabic-word responses would depend on the tonal f0 contour that a target word bears and the rate of the carrier sentences where a target word occurs. Monosyllabic-word responses would increase as the tonal f0 contours became more like a single tone event and the carrier sentences became slower, despite the intact formant patterns that originally produced bisyllabic words. Interaction between speech rates and tonal f0 contours, however, was not expected, as there is no existing evidence that would lead us to believe so. In other words, we predicted that the effect of speech rate would hold across tonal f0 contours, and the effect of tonal f0 contours would hold across different speech rates. Finally, we expected the conditions of isolated words not to differ from the conditions of the medium speech rate, since the target words were originally produced at the medium rate, as mentioned in Section 2.1.2.

### **3. Results**

Overall, the predictions outlined in Section 2.3 were borne out in our results: Tonal f0 contours and speech rate each considerably influenced participants' judgments about the number of syllables in a target word. In this section, the results of LH and HL target words will be analyzed separately, as the segmental properties were not fully controlled across the tonal combinations.

As can be seen in Figures 4-5, target words are perceived as monosyllabic more often when their f0 contours are more monotone-like and when their carrier sentences are slower. In the presence of original, bisyllabic segmental material, participants still chose the monosyllabic alternatives for some percentage of the time, depending on the conditions of tonal f0 and speech rate. In terms of tonal f0, the most monotonal f0 contours on average had 30% more of the monosyllabic choices than the most bitonal f0 contours for the LH words and 21% for the HL words. In the conditions of fast carriers, medium-rate carriers, slow carriers, and isolated words, the mean percentage of monosyllabic-word responses in Step 1 of tonal f0 was higher than that in Step 6 by 27, 35, 33, and 27, respectively, for the LH words and 19, 27, 19, and 19, respectively, for the HL words. With regard to speech rate, the slow carrier sentences on average had 28% more of the monosyllabic choices than the fast carrier sentences for the LH words and 22% for the HL words. From Steps 1 to 6 of tonal f0 contours respectively, the mean percentage of monosyllabic-word responses at the slow rate was higher

than that at the fast rate by 27, 33, 27, 32, 31, and 21 for the LH words and 23, 21, 29, 19, 21, and 23 for the HL words. Furthermore, by comparing the two ‘extreme’ conditions with each other, namely the condition of tonal f0 Step 1 at the slow rate with the condition of tonal f0 Step 6 at the fast rate, we see that tonal f0 contours and speech rate together increased the chance for target words to be identified as monosyllabic by 54% in the LH words and 42% in the HL words.

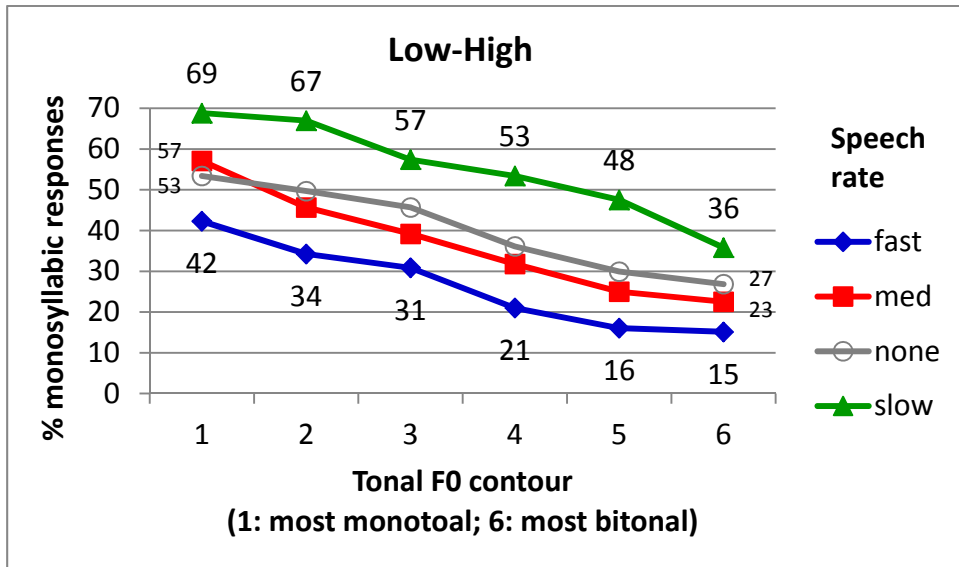


Figure 4. Percentage of monosyllabic-word responses in each condition of the Low-High tonal sequences

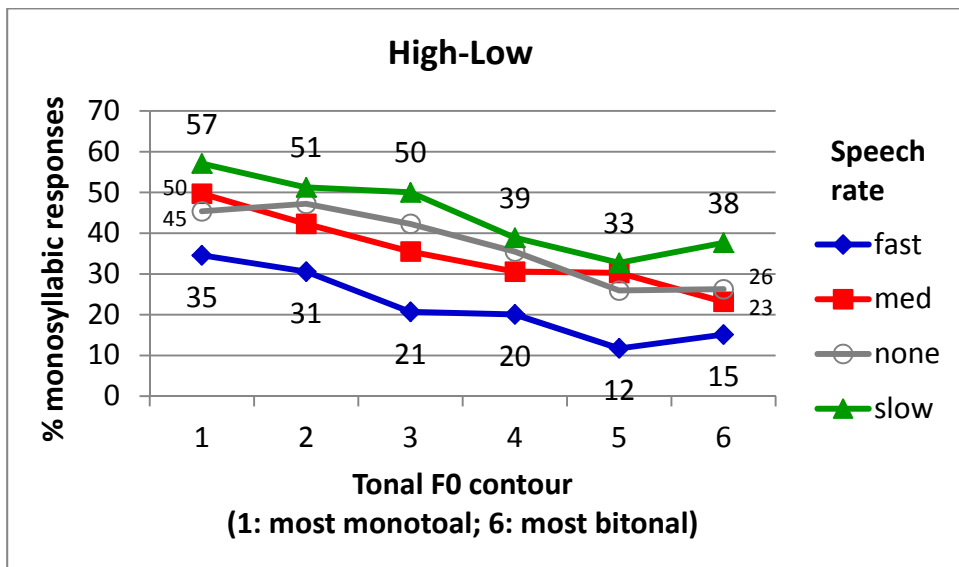


Figure 5. Percentage of monosyllabic-word responses in each condition of the High-Low tonal sequences

*the Role of Local Tonal f0 and Global Speech Rate in Syllabification*

Our observations based on visual inspection are confirmed by statistical analysis. The percentages of monosyllabic-word choices were transformed into arcsine data and analyzed using R: A Language and Environment for Statistical Computing (R Core Team, 2013) and the R packages *lmerTest* (Kuznetsova, Brockhoff & Christensen, 2012) and *multcomp* (Hothorn, Bretz & Westfall, 2008). In all of the analyses presented in this paper, the hypotheses were tested at a significant level of  $\alpha = 0.05$  and a marginally significant level of  $\alpha = 0.1$ . Mixed effect models were conducted using the *lmer* function in the *lme4* package. We used tonal f0 contours and speech rate as fixed effects and used subjects and items (*i.e.* target word pairs) as random effects. Models with different structures of fixed effects and random effects were compared using the *anova* function in the standard R distribution. We rejected models including a given effect or interaction when they did not differ significantly from the models excluding it. The results show the main effects of tonal f0 contours and speech rate in both LH and HL sequences (tonal f0 contours:  $F(5,1891.5)$ 's  $\geq 38.498$ ,  $p$ 's  $< 0.001$ ; speech rate:  $F(3,1891.5)$ 's  $\geq 72.141$ ,  $p$ 's  $< 0.001$ ). No interaction between tonal f0 contours and speech rates was found in either LH or HL sequences ( $F(15,1891.5)$ 's  $\leq 1.193$ ,  $p$ 's  $\geq 0.269$ ), as predicted. Planned comparisons were conducted using the *glht* function in the *multcomp* package. For both LH and HL sequences, tonal f0 contours and speech rate each robustly differed between the two ends on their scale. In terms of tonal f0, bisyllabic words with the most monotonal f0 contours were misidentified significantly more often as their monosyllabic alternatives than those with the most bitonal f0 contours at all four levels of speech rate ( $z$ 's  $\geq 4.576$ ,  $p$ 's  $< 0.001$ ). With regard to speech rate, bisyllabic words with the slow carrier sentences were misidentified significantly more often as monosyllabic than those with the fast carrier sentences in all six steps of tonal f0 contours ( $z$ 's  $\geq 4.692$ ,  $p$ 's  $< 0.001$ ). Words presented in isolation did not differ from those flanked by the medium-rate carrier sentences ( $|z|$ 's  $\leq 2.201$ ,  $p$ 's  $> 0.664$ ), as we expected (see Section 2.1.2 for relevant discussion).

Let us now move on to the intermediate levels of tonal f0 contours and speech rate, namely words with an f0 contour between Steps 2-5 and those in the medium-rate carrier sentences. Although some of the conditions did not significantly differ from each other, the patterns discussed in the preceding paragraph emerge in the intermediate levels as well. In fact, when we look at the pairs of conditions that have a significant difference, all of these response tendencies fit our predictions (*i.e.* monosyllabic-word responses increased when f0 was more monotonal and speech rate was slower), as shown in Tables 2-3. Thus, the hypothesis discussed in Section 2.3 is largely supported by our results. In what follows, we will first look at the middle steps of tonal f0 contours, *i.e.* comparisons between those less than five steps apart, and move on to comparing the medium rate with the slow and fast rate.

*Adjacent steps of tonal f0 contours:* There was no significant difference between any adjacent steps of f0 contours ( $z$ 's  $\leq 2.937$ ,  $p$ 's  $> 0.153$ , except Step 5 vs. Step 6 in HL was marginal:  $z = 3.128$ ,  $p = 0.090$ ). *Tonal f0 contours two steps apart (Step 1 vs. Step 3, Step 2 vs. Step 4, Step 3 vs. Step 5, Step 4 vs. Step 6):* For LH words, monosyllabic responses occurred significantly more in Step 1 than Step 3 at the medium rate, in Step 3 than Step 5 at the fast and the medium rate, and in Step 4 than Step 6 at the slow rate. For HL words, such a difference was significant between Steps 1 and 3 at the medium rate, and Steps 3 and 5 at the slow rate (Step 1 vs. Steps 3 at the medium rate in LH and HL:  $z$ 's  $\geq 3.533$ ,  $p$ 's  $< 0.05$ ; Step 3 vs. Steps 5 at the fast and medium rate in LH:  $z$ 's  $\geq 3.372$ ,  $p$ 's  $< 0.05$ ; Step 3 vs. Steps 5 at the slow rate in HL:  $z = 4.634$ ,  $p < 0.001$ ; Step 4 vs. Step 6 at the slow rate in LH:  $z = 4.406$ ,  $p < 0.001$ ; the remainder had  $z$ 's  $\leq 3.046$ ,  $p$ 's  $\geq 0.114$ , except these pairs were marginal: Step 1 vs. Step 3 in HL at the fast rate,  $z = 3.244$ ,  $p = 0.062$ , and Step 2 vs. Step 4 in LH at all three rates,  $z$ 's  $\geq 3.155$ ,  $p$ 's  $\leq 0.084$ ). *Tonal f0 contours three steps apart (Step 1 vs. Step 4, Step 2 vs. Step 5, Step 3 vs. Step 6):* In LH words with f0 contours that were three steps away from each other, the more monotonal contours yielded significantly more monosyllabic responses than the more bitonal contours at all three speech rates. In HL words, only the slow rate conditions showed significant differences between all of the pairs of f0 contours that were three steps apart. Among the conditions of the other two speech rates, significant differences only appeared between Steps 1 and 4 at the medium and the fast rate and Steps 2 and 5 at the fast rate (all three pairs at all three rates in LH:  $z$ 's  $\geq 3.699$ ,  $p$ 's  $< 0.05$ ; all three pairs at the slow rate in HL:  $z$ 's  $\geq 3.359$ ,  $p$ 's  $< 0.05$ ; Step 1 vs. Step 4 at the medium rate in HL:  $z = 4.344$ ,  $p < 0.001$ ; Step 2 vs. Step 5 at the fast rate in HL:  $z = 5.213$ ,  $p < 0.001$ ; the remainder had  $z$ 's  $\leq 2.896$ ,  $p$ 's  $\geq 0.170$ , except Step 1 vs. Step 3 at fast the rate in HL was marginal,  $z = 3.128$ ,  $p = 0.090$ ). *Tonal f0 contours four steps apart (Step 1 vs. Step 5, Step 2 vs. Step 6):* Words with more monotonal contours were significantly more often perceived as monosyllabic than words with more bitonal contours at all three speech rates ( $z$ 's  $\geq 4.170$ ,  $p$ 's  $< 0.01$ ), except in HL words at the slow rate where Steps 2 and 6 only marginally differ ( $z = 3.128$ ,  $p = 0.089$ ). *Medium vs. fast rate:* Monosyllabic responses occurred significantly more in the medium rate than in the fast rate in Step 1 of the f0 contours for LH words, and Steps 1, 3, and 5 of the f0 contours for HL words (Step 1 in LH:  $z = 3.427$ ,  $p < 0.05$ ; Steps 1, 3, and 5 in HL:  $z$ 's  $\geq 3.707$ ,  $p$ 's  $< 0.05$ ; the remainder:  $z$ 's  $\leq 2.780$ ,  $p$ 's  $\geq 0.231$ ). *Slow vs. medium rate:* Words in slow carrier sentences were significantly more often perceived as single syllables than words in medium-rate carrier sentences for LH words with all of the steps of f0 contours, except only marginally so in Step 1. For HL words, such a difference was only significant in Steps 3 and 6 of the f0 contours (Steps 2-5 in LH:  $z$ 's  $\geq 3.427$ ,  $p$ 's  $< 0.05$ ; Steps 3 and 6 in HL:  $z$ 's  $\geq 3.765$ ,  $p$ 's  $< 0.05$ ; the remainder:  $z$ 's  $\leq 2.259$ ,  $p$ 's  $\geq 0.617$ , except Step 1 in LH had  $z = 3.100$ ,  $p = 0.098$ ).

**Table 2. Effects of tonal f0 at each speech rate**

Distance in tonal f0 steps	Contrast	LH	HL
1 step apart	Step 1 vs. Step 2	F M S	F M S
	Step 2 vs. Step 3	F M S	F M S
	Step 3 vs. Step 4	F M S	F M S
	Step 4 vs. Step 5	F M S	F M S
	Step 5 vs. Step 6	F M S	F M S <sup>•</sup>
2 steps apart	Step 1 vs. Step 3	F M <sup>*</sup> S	F <sup>•</sup> M <sup>*</sup> S
	Step 2 vs. Step 4	F <sup>•</sup> M <sup>•</sup> S <sup>•</sup>	F M S
	Step 3 vs. Step 5	F <sup>*</sup> M <sup>*</sup> S	F M S <sup>*</sup>
	Step 4 vs. Step 6	F M S <sup>*</sup>	F M S
3 steps apart	Step 1 vs. Step 4	F <sup>*</sup> M <sup>*</sup> S <sup>*</sup>	F <sup>•</sup> M <sup>*</sup> S <sup>*</sup>
	Step 2 vs. Step 5	F <sup>*</sup> M <sup>*</sup> S <sup>*</sup>	F <sup>*</sup> M S <sup>*</sup>
	Step 3 vs. Step 6	F <sup>*</sup> M <sup>*</sup> S <sup>*</sup>	F M S <sup>*</sup>
4 steps apart	Step 1 vs. Step 5	F <sup>*</sup> M <sup>*</sup> S <sup>*</sup>	F <sup>*</sup> M <sup>*</sup> S <sup>*</sup>
	Step 2 vs. Step 6	F <sup>*</sup> M <sup>*</sup> S <sup>*</sup>	F <sup>*</sup> M <sup>*</sup> S <sup>•</sup>
5 steps apart	Step 1 vs. Step 6	F <sup>*</sup> M <sup>*</sup> S <sup>*</sup>	F <sup>*</sup> M <sup>*</sup> S <sup>*</sup>

(F: fast rate; M: medium rate; S: slow rate; \*: significantly more monosyllabic-word responses in the more monotonal step than in the more bitonal step; <sup>•</sup>: marginally more monosyllabic-word responses in the more monotonal step than in the more bitonal step)

**Table 3. Effects of speech rate in each tonal f0 step**

	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
<b>Medium vs. Fast</b>	LH <sup>*</sup> HL <sup>*</sup>	LH HL	LH HL <sup>*</sup>	LH HL	LH HL <sup>*</sup>	LH HL
<b>Slow vs. Medium</b>	LH <sup>•</sup> HL	LH <sup>*</sup> HL	LH <sup>*</sup> HL <sup>*</sup>	LH <sup>*</sup> HL	LH <sup>*</sup> HL	LH <sup>*</sup> HL <sup>*</sup>
<b>Fast vs. Slow</b>	LH <sup>*</sup> HL <sup>*</sup>	LH <sup>*</sup> HL <sup>*</sup>	LH <sup>*</sup> HL <sup>*</sup>	LH <sup>*</sup> HL <sup>*</sup>	LH <sup>*</sup> HL <sup>*</sup>	LH <sup>*</sup> HL <sup>*</sup>

(\*: for the two speech rates compared in a given cell, there are significantly more monosyllabic-word responses at the slower rate than at the faster rate; <sup>•</sup>: for the two speech rates compared in a given cell, there are marginally more monosyllabic-word responses at the slower rate than at the faster rate)

#### 4. Discussion

The study presented in this paper investigates two kinds of perceptual cues for syllabification in Mandarin: lexical pitch patterns that are locally associated with words and the overall speech rate of a sentence that is determined by global measures (*e.g.* average syllable length). As the traditional view seems to connect syllable perception with segmental cues that result from local supralaryngeal events, we are particularly interested in whether non-segmental and non-local laryngeal information contribute to syllable perception as well. A better understanding of these factors is important because they figure fundamentally in how the language perception system integrates different sources of information. In the following subsections, we will discuss how our results correspond to our research questions, sketched out in Section 1.1, as well as their broader implications.

Our key aims were to investigate whether and to what extent the syllable count of a Mandarin word depends on the tonal  $f_0$  contour that the word carries and the speech rate of the sentence that includes the word. Our results show that local tonal  $f_0$  and global speech rate highly impact the perception of the number of syllables. Across the two types of tonal sequences examined in this study (LH and HL), a change in tonal  $f_0$  contours increases the chance of bisyllabic words being perceived as monosyllabic by as much as 35%, and a change in speech rate does so by as much as 33%; changes in tonal  $f_0$  and speech rate together impact the likelihood by as much as 54%. These findings indicate that non-segmental and non-local information can contribute substantially to the perception of the number of syllables.

Note that the segmental content of an item in this study was taken from a naturally-produced, bisyllabic word token; its formant patterns strongly encouraged participants to choose the bisyllabic word. If segmental information had dominated syllable perception, we would have obtained equally low percentages of monosyllabic-word responses, no matter what kinds of tonal  $f_0$  contours or speech rate had been presented to the participants. The fact that the changes in tonal  $f_0$  contours and speech rate changed the likelihood of whether an item was interpreted as monosyllabic or bisyllabic tells the importance of non-segmental information in syllabification. Moreover, when segmental material and non-segmental material provide conflicting cues, for the language perception system to comply with the cues provided by non-segmental material, it must somewhat ignore the cues provided by segmental material – in our study, the two [u] vowels across the syllable boundary needed to be perceived as one. Results of our study suggest that tonal  $f_0$  contours and speech rate interfere with segments in perceiving the number of syllables.

Can the cues provided by tonal  $f_0$  and speech rate override segmental information in determining syllable counts? This strong claim is not supported by our results. When the tonal  $f_0$  content is ambiguous with regard to the number of syllables in a word, *i.e.* in the conditions



with the middle steps of tonal f0 contours (Steps 3 and 4), changes in speech rate increase the chance that listeners are ‘mistaken’ about syllable numbers by only 19-32%. Similarly, changes in tonal f0 raise the likelihood by only 27-35% when the speech rate is neutral, *i.e.* under the condition of medium-rate carrier sentences. In other words, when one of the non-segmental cues (*i.e.* tonal f0 or speech rate) conflicts with segmental information and the other stays neutral, perception of syllable numbers is at best altered to an extent where the likelihood of misperception increases by 35%. It does not seem that either of these two factors decisively influences the syllable count of a word. One might ask, however, can a combination of non-segmental cues ‘beat’ segmental information? If this were the case, we would have observed a big boost in the percentage of monosyllabic-word responses when both tonal f0 and speech rate strongly favored a monosyllabic interpretation. In fact, our manipulation of tonal f0 and speech rate only produced maximally 54% greater likelihood for bisyllabic words to be identified as monosyllabic. It seems that, even when these two sources of non-segmental information both conflict with the segmental cues, they do not dominate listeners’ perception but rather only create confusion about the number of syllables in a word. On the other hand, prior work has shown that degradation of spectral information impacts the perception of the number of syllables and the perception of tone categories (Olsberg *et al.*, 2007). Combining our findings with theirs, it appears that syllable perception in a tonal language involves integration of spectral patterns, tonal f0 contours, as well as the durational information provided by the speech rate.

As we saw in Section 1, existing studies have shown that listeners are sensitive to both local and global non-segmental cues in word segmentation, lexical access, and phoneme identification. Our study adds to the body of literature on the perception of linguistic units by providing evidence for non-segmental and non-local cues in syllable perception. Listeners are capable of using a vast range of information – local or global, supralaryngeal or laryngeal – to segment and identify linguistic units of different levels. Although segmental material alone might be sufficient for syllabification, the presence of cues from other sources impacts the perceptual number of syllables. This highlights cue integration (see Repp, 1982 for an early review on this topic) as the nature of the language comprehension system to utilize multiple domains of cues when available.

## **5. Conclusion**

Based on the perception study reported in this paper, we conclude that the perception of syllable numbers in Mandarin depends on the global speech rate of the entire utterance and local tonal f0 contours that are associated with lexical tones, as well as local segmental material (see also Olsberg *et al.*, 2007). Against the traditional view that syllable perception by and large involves the segmental content of an utterance, we found that the lexical pitch

patterns also play an important role in syllabification. Furthermore, not only local information, such as segments and tonal  $f_0$ , but also global information, such as speech rate, serve as cues for syllable counts. Our findings provide further evidence for cue integration as the nature of the language comprehension system. Even though syllabification and syllable identification supposedly can be accomplished based on local segmental information alone, the presence of non-segmental and non-local information impacts the perception of syllables.







## References

- Brown, M., Salverda, A. P., Dilley, L. C. & Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin & Review*, 18 (6), 1189-1196.
- Caramazza, A. & Yeni-Komshian, G. H. (1974). Voice onset time in two French dialects. *Journal of Phonetics*, 2, 239-245.
- Chen, Y. & Braun, B. (2006). Prosodic realization in information structure categories in standard Chinese. In R. Hoffmann & H. Mixdorff (Eds.), *Speech Prosody 2006*. Dresden, Germany: TUD Press.
- Christophe, A., Peperkamp, S., Pallier, C., Block, E. & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access: I. Adult data. *Journal of Memory and Language*, 51, 523-547.
- DeFrancis, John. (1984). *The Chinese language: Fact and fantasy*. Honolulu, HI: University of Hawaii Press.
- Dilley, L. C. & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, 59, 294-311.
- Gao, M. (2008). *Tonal Alignment in Mandarin Chinese: An Articulatory Phonology Account*. (Doctoral dissertation). Yale University, New Haven, CT.
- Gout, A., Christophe, A. & Morgan, J. (2004). Phonological phrase boundaries constrain lexical access: II. Infant data. *Journal of Memory and Language*, 51, 547-567.
- Hothorn, T., Bretz, F. & Westfall, F. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, 50 (3), 346-363.
- Lin, H. & Wang, Q. (2007). Mandarin rhythm: An acoustic study. *Journal of Chinese Linguistics and Computing*, 17 (3), 127-140.
- Kuznetsova, A., Brockho, P. B. & Christensen, R. H. B. (2012). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). URL <http://www.cran.r-project.org/package=lmerTest/>.
- Mattys, S. L., Jusczyk, P. W., Luce, P. A. & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38, 465-494.
- Miller, J. & Grosjean, F. (1981). How the components of speaking rate influence the perception of phonetic segments. *Journal of Experimental Psychology: Human Perception and Performance*, 7, 208-215.






*the Role of Local Tonal  $f_0$  and Global Speech Rate in Syllabification*

- Nakatani, L. H. & Schaffer, J. A. (1978). Hearing 'words' without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America*, 63, 234-245.
- Olsberg, M., Xu, Y. & Green, G. (2007). Dependence of tone perception on syllable perception. *Proceedings of Interspeech 2007*, Antwerp, Belgium, 2649-2652.
- Ouyang, I. C. & Kaiser, E. (2013). Prosody and information structure in a tone language: an investigation of Mandarin Chinese. *Language and Cognitive Processes*. DOI:10.1080/01690965.2013.805795.
- Prom-on, S., Xu, Y. & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America*, 125, 405-424.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Repp, B. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, 92 (1), 81-110.
- Shen, X. S. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics*, 18, 281-295.
- Shields, J. L., McHugh, A. & Martin, J. G. (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*, 102, 250-255.
- Shih, C. (1986). The phonetics of the Chinese tonal system. Technical memorandum, AT & T Bell Laboratories.
- Summerfield, A. Q. (1975). Aerodynamics versus mechanics in the control of voicing onset in consonant-vowel syllables. *Speech Perception*, 2 (4), 61-72. Belfast, Northern Ireland: Queen's University.
- Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25, 61-83.
- Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica*, 55, 179-203.

## Appendix A: Target words

Tone	Vocalic structure	Word	Image
LH	u.u	(a) tsu.u      ‘ancestral house’	
	Vu.u	(b) pau.u      ‘treasure valley’	
	u.uV	(c) hu.uo      ‘tiger’s lair’	
HL	u.u	(d) tʂ <sup>h</sup> u.u      ‘first dance’	
	Vu.u	(e) tʂou.u      ‘Friday’	
	u.uV	(f) ku.ua      ‘single roof tile’	

**Appendix B: Monosyllabic alternative of each target word**

Tone	Vocalic structure	Word	Image
R	u	(a) tsu 'foot'	
	Vu	(b) pau 'thin'	
	uV	(c) huo 'alive'	
F	u	(d) tʂ <sup>h</sup> u 'touch'	
	Vu	(e) tʂou 'wrinkle'	
	uV	(f) kua 'Chinese hexagram'	