Syllable perception depends on tone perception

Iris Chuoying Ouyang, Khalil Iskarous

Department of Linguistics, University of Southern California, Los Angeles, USA

chuoyino@usc.edu, kiskarou@usc.edu

Abstract

We conducted a perception study on Mandarin, a tone language where pitch carries contrastive information, to investigate whether pitch changes interfere with spectral information in determining the number of syllables in a utterance. We generated F0 contours and simulated tonal coarticulation using the qTA model [1]. The perception of syllable numbers depended on the perception of tones, and this effect held across speech rate. Combining with prior work [2], the results indicate that laryngeal and supralaryngeal events interact in syllable perception in tone languages. We discuss how our findings support the notion of language-specific perception.

Index Terms: syllable perception, tonal coarticulation, speech rate, Mandarin

1. Introduction

The number of syllables in a word is generally thought to be determined by segmental information, after taking phonotactics into account. In other words, syllable perception in a given language is usually believed to depend on two factors: what kinds of segments there are in a word, and what kinds of syllable structure are possible in the language. Pitch patterns are, in contrast, not considered to be involved in syllable perception, since pitch is associated with laryngeal movements, which are largely independent from supralaryngeal movements that decide the segmental content of an utterance.

However, do pitch patterns never substantially affect the perception of syllables? The mappings between phonetic representations and phonological categories are known to be language-dependent [3]; what count as phonetic cues for syllabification might vary with the characteristics of a given language. In tone languages, where pitch plays a crucial role in syllabification, it is reasonable to consider F0 as a potential factor of syllable perception. The contrastive pitch patterns, referred to as lexical tones, have been suggested to align with syllable edges or syllable constituents in East Asian tone languages such as Mandarin [4][5]. Could this distinctive use of highly local pitch patterns yield interactions between segmental material and F0 material in the perception of syllable numbers?

Degradation of segmental information has been shown to influence the perception of tones and syllables in Mandarin. When high-frequency spectral information was removed from a sequence of [ma] syllables, listeners could not accurately identify the number of syllables and the categories of tones in the sequence, despite the presence of intact F0 information [2]. What remains unclear is, on the other hand, whether the perception of segments and syllables is affected by tonal information in Mandarin – could F0 changes override formant patterns in determining the number of syllables?

1.1. Background: Mandarin tones

To investigate whether the number of syllables depends on pitch dynamics in a tone language, we conducted a perception study on Mandarin. We asked: does change in F0 contours, without change in formant patterns, alter the count of syllables in a word?

There are five lexical tones in Mandarin: high (Tone 1), rising (Tone 2), low (Tone 3), falling (Tone 4), and a neutral tone 1. Every syllable carries a tone, which distinguishes between lexical items, as shown in (1).

\[
\begin{align*}
\text{Tone 1} & \quad \text{ma [High]} \quad \text{‘mother’} \\
\text{Tone 2} & \quad \text{ma [Rising]} \quad \text{‘hemp’} \\
\text{Tone 3} & \quad \text{ma [Low]} \quad \text{‘horse’} \\
\text{Tone 4} & \quad \text{ma [Falling]} \quad \text{‘scold’}
\end{align*}
\]

The simplicity of the tone system in Mandarin allows us to examine how 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 is similar to a contour tone (i.e. Rising or Falling) in terms of the shape, but typically different from the a contour tone in F0 onset and offset, F0 ranges, the turning point of F0 movement, etc. [6][7]. This provides the circumstances where two words are composed of different numbers of syllables, yet minimally differ from each other in both 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] and [CVVC]. Thus, we were able to use real words in the experiment, where the stimuli only differed in F0 while retaining a chance of being identified as either bisyllabic or monosyllabic.

2. Method

Participants performed an identification task, where they heard a word in isolation or in a carrier sentence, saw pictures on the computer screen, and determined which picture matched the word. We manipulated the tones and controlled the segments in a word and the speech rate of its carrier sentence. The use of pictures allowed us to avoid presenting participants with written words that convey phonetic information in addition to the audible stimuli.

---

1 The neutral tone in Mandarin is critically unspecified. Its actual F0 contour depends on the adjacent tones, yet a neutral tone is contrastive to other lexical tones, e.g. ‘ʃɤ [Rising]’ou [Rising]’ ‘tongue’ vs. ‘ʃɤ [Rising]’ou [Neutral]’ ‘head of snake’.

2.1. Design and stimuli

A male speaker of Beijing Mandarin recorded the target words in isolation at a medium speech rate (254 ms/syllable), and the sentence frame, as illustrated in (2), at three different speech rates: fast (131 ms/syllable), medium (259.5 ms/syllable), and slow (441.5 ms/syllable). Carrier sentences serve for two major purposes. First, it enables us to examine the effect of F0 changes across different speech rates while keeping constant the duration of target words. In other words, we manipulated the perception of overall speech rate by embedding the same target item in a sentence frame that was recorded at different speech rates. Second, it also allows us to vary the F0 excursion of target words in a wider range, because target words become the discourse focus in a sentence when the sentence frame were the same throughout the experiment. The existing work shows that new focus in a sentence when the sentence frame were the same target word to be perceived as monosyllabic, (i) the two [u][uV], in which V refers to a vowel other than [u]. Thus, for a Three types of vocalic sequences were used: [u.uV], [Vu.u], and [u.uV], appeared at both sides of the syllable boundary in every word. The same vowel [u] as the perceptual consequences of F0 changes are listed in Table 2.

A repeated-measured design with two independent variables was used: (i) ‘monotone-ness’ of a F0 contour (with 6 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 4 levels: fast, medium, slow, and no carrier). Every participant responded to six repetitions of an item. Thus, there were 144 items (6 target words * 6 steps of F0 contours * 4 levels of speech rate) and 864 trials in total.

2.2. Participants and procedures

Native speakers of Mandarin, ten women who were born in China, participated in the study. All of them were students at University of Southern California who left China no longer than 3.5 years before.

The experiment was a two-alternative forced-choice identification test (2AFC). Participants saw two pictures side by side while hearing a target word in isolation or flanked by a carrier sentence. One of the pictures represented the bisyllabic target word they heard, and the other represented a monosyllabic word which they might have ‘misperceived’. Participants were asked to listen to what the speaker said and choose the picture that matched the word he said. They were told that the study was interested in how people recognize words.

Table 1. Target words and their tonal and vocalic properties

<table>
<thead>
<tr>
<th>Tone</th>
<th>Vocalic Structure</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>HL</td>
<td>u.u</td>
<td>(a) tʂhu.u  ‘first dance’</td>
</tr>
<tr>
<td>HL</td>
<td>Vu.u</td>
<td>(b) tʂou.u  ‘Friday’</td>
</tr>
<tr>
<td>HL</td>
<td>u.uV</td>
<td>(c) u.ua     ‘roof tile’</td>
</tr>
<tr>
<td>LH</td>
<td>u.u</td>
<td>(d) tʂu.u   ‘ancestral house’</td>
</tr>
<tr>
<td>LH</td>
<td>Vu.u</td>
<td>(e) pau.u   ‘treasure valley’</td>
</tr>
<tr>
<td>LH</td>
<td>u.uV</td>
<td>(f) huo.u   ‘tiger’s lair’</td>
</tr>
</tbody>
</table>

Table 2. Monosyllabic alternatives of the target words

<table>
<thead>
<tr>
<th>Tone</th>
<th>Vocalic Structure</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>u</td>
<td>(a) tʂhu  ‘touch’</td>
</tr>
<tr>
<td>F</td>
<td>Vu</td>
<td>(b) tʂou  ‘wrinkle’</td>
</tr>
<tr>
<td>F</td>
<td>uV</td>
<td>(c) ua   ‘sock’</td>
</tr>
<tr>
<td>R</td>
<td>u</td>
<td>(d) tʂu  ‘foot’</td>
</tr>
<tr>
<td>R</td>
<td>Vu</td>
<td>(e) pau  ‘thin’</td>
</tr>
<tr>
<td>R</td>
<td>uV</td>
<td>(f) huo  ‘alive’</td>
</tr>
</tbody>
</table>

2.3. F0 synthesis

In order to systematically vary pitch patterns, we generated F0 contours based on the qTA model [1] and synthesized with natural words using the PSOLA (Pitch Synchronous Overlap Add) method implemented in the Praat software. The synthetic F0 contours were on a continuum between the most two-tone-like pattern and the most single-tone-like pattern (i.e. LH-R and HL-F). All the parameters in the qTA model were fixed but the rate at which F0 changed in the second syllable of a target word (i.e. Low in HL and High in LH) was manipulated. Fast F0 movements suggest that there is only one tone (i.e. black lines in Figure 1-2), whereas slow F0 movements indicate the existence of two tones (i.e. red lines in Figure 1-2). In the F0 contours which were more bitonal, this simulated the carryover effect that a previous tone has on a following tone in Mandarin [8].

2.4. Hypothesis and predictions

Measurements in this study are the counts of monosyllabic-word choices, namely, the numbers of trials where participants chose the picture representing the monosyllabic alternative (e.g. [huo] instead of the bisyllabic one (e.g. [hu.uo])). Our general hypothesis is that syllables, as important linguistic units, of which the perception is determined by the phonetic cues which are crucial in a given language. As F0 contours provide crucial cues for tone events, we predict that the change of F0 contours substantially impacts the perception of syllable numbers in Mandarin, regardless of segmental information, and that such effect should appear across speech rate. Specifically, we expect that the counts of monosyllabic-word choices increase as the F0 contours indicate a single tone event, despite that the intact formant patterns originally produce bisyllabic words. We also expect the counts of monosyllabic-word choices increase as the speak rate becomes slower, since speech rate has been shown to affect the length of segmental units [10].
3. Results

Overall, the results confirm our predictions: Both F0 contours and speech rate influenced participants' judgments about the number of syllables in a target word. As can be seen in Figure 3, the percentages of monosyllabic-word responses increase when the F0 contours are more monotone-like and when the carrier sentences are slower. In the presence of original, bisyllabic segmental material, the participants still chose the monosyllabic word for 44% of the time (157 out of 360 trials) in the conditions with the most monotonal F0 contours and slowest carrier sentences. In general, the most monotonal F0 contours (Step 6 in Figure 3) induce 20% more of the monosyllabic choices than the most bitonal F0 contours (Step 1 in Figure 3); the slow carrier sentences yield 15% more of the monosyllabic choices than the fast carrier sentences.

Mixed-effect models were conducted with F0 contours and speech rate as two fixed effects and subjects as the random effect. The percentages of monosyllabic-word choices were transformed into arcsine data and tested at a significant level of p<.05. Both F0 contours and speech rate show significant differences between the two ends: The most monotonal contours induce significantly more monosyllabic-word responses than the most bitonal contours in all four conditions of speech rate; the slow carrier sentences induce significantly more monosyllabic-word responses than the fast carrier sentences in all six steps of F0 contours. With regard to the intermediate conditions, namely the middle steps of F0 contours and the medium-rate carrier sentences, their response biases depend on items. Interaction between F0 contours and speech rate is not found. Words presented in isolation are not significantly different from those embedded in medium-rate carrier sentences, in terms of the response tendency in any step of F0 contour.

Recall that the target words were controlled along two other dimensions: lexical tone categories, HL or LH, and vocalic structures, [u/u], [Vu.u], and [u.uV]. We did T-tests on tone categories for every step of F0 contours in each condition of speech rate, and found no significant difference in the percentages of monosyllabic-word responses between HL and LH sequences. In contrast, different vocalic structures appear to affect the response tendencies. In general, as can be seen in Figure 4, the [u/u] sequences induce more monosyllabic-word choices than the [Vu.u] sequences, which in turn yield more monosyllabic choices than the [u.uV] sequences. In other words, it is more likely for [u/u] to be identified as [u], than for [Vu,u] to be identified as [Vu], than for [uV.u] to be identified as [uV]. The largest difference in response tendencies between vocalic sequences emerges in the conditions of most monotonal F0 contours: The [u/u] sequences lead to 23% more of the monosyllabic choices than the [u.uV] sequences; this difference is even larger than the difference between the most monotonal and the most bitonal F0 contours (20%, as described above). T-test statistics suggest the magnitude of response biases induced by vocalic structures varies across different steps of F0 contours and conditions of speech rate, yet the overall percentage of monosyllabic-word choices do differ between vocalic structures.
4. Discussion and Conclusions

This study aims at a better understanding of the perception of syllables. Based on the long-developed notion of language-specific perception [3], we believe that pitch patterns play an important role in syllable perception in tone languages, where pitch provides the major cue for lexical tones. Therefore, we conducted a perception study to investigate whether pitch information can override segmental information in determining the number of syllables in a tone language, Mandarin. Our results clearly show that the perception of syllable numbers depends on the perception of tones, and this effect holds across speech rate. When the segmental content indicates a bisyllabic word while the tonal content favors the interpretation of a monosyllabic word, listeners identify the word as monosyllabic for a considerable percent of time, e.g. 35%, in a medium-rate carrier sentence with the most monotonal F0 contour. Note that the segmental content of an item in this study was taken from a naturally-produced token of the bisyllabic word; the formant pattern in an item strongly encouraged participants to choose the bisyllabic word. If segmental information were sufficient for syllable perception, we should have obtained equally low counts of monosyllabic-word choices, no matter what kinds of pitch information were present. The fact that the change in F0 contours changed the amount of monosyllabic interpretations reflects the importance of pitch information in syllable perception. Furthermore, when the segmental material and the tonal material provide conflicting cues, for the language perception system to comply with the cues from tonal material, it must somewhat ignore the cues from segmental material – in our study, the two [u] across the syllable boundary have to be perceived as one. Thus, our findings suggest that pitch information interferes with segmental information in the perception of syllable numbers.

There is also evidence in our results that shows the importance of segmental information in syllable number perception. The percentages of monosyllabic choices differ between vocalic structures, which implies that the effect of a conflicting tonal cue is weaker in some segmental contexts than others. Among the three vocalic structures, [u] allows the highest percentages of monosyllabic-word responses, [Vu.u] the second, and [u.uV] the lowest. A possible explanation for these differences is the degree of perceptual difficulty in merging two [u] into one. When [u.u] are identified as monosyllabic, the two [u] are perceived as one full vowel; whereas in the cases of [Vu.u] and [u.uV], the two [u] need to be perceived as part of a diphthong. Existing work has shown that vowels and glides are perceptually distinguished by duration: Two vocalic segments that include a [j] or [w] are identified as a diphthong, rather than two vowels, only when the [j] or [w] is sufficiently short [11][12]. Therefore, the reason why the [u] sequences obtain more monosyllabic-word responses might be that, the combined length of two [u] in the [Vu.u] and [u.uV] sequences is too long for them to be easily perceived as part of a diphthong. With regard to the different response tendencies between the [Vu.u] and [u.uV] sequences, a speculation based on the same argument is that glides in rising diphthongs might be longer than those in falling diphthongs; it might be easier for the language perception system to accept an extraordinarily long [u] as the second member of [Vu] than as the first member of [uV]. While this phenomenon deserves further investigation, the effect of vocalic sequences clearly indicate the influence of segmental cues on the perception of syllable numbers.

Prior work has shown that degradation of spectral information impacts the perception of syllable numbers and hence the perception of tones [2]. Combining our findings with theirs, it appears that the cues for segmental events, whether formant patterns or duration, and the cues for pitch events interact in perceiving syllables in Mandarin. In tone languages where the temporal relations between segmental and pitch events are crucial to word recognition, syllable perception involves the integration of information from pitch as well as segments.

5. References