Speech segmentation and tone sandhi in Mandarin Chinese

Sijia Zhang[1] and Michael Wagner[2]
[1]University of British Columbia, [2]McGill University

When listening to speech, we effortlessly segment the incoming acoustic stream into words. How we achieve this is still not fully understood. The so-called Iambic-Trochaic Law (ITL) states that listeners parse the signal into iambs when long and short sounds alternate, and trochees when loud and soft sounds alternate, both in sequences of tones [1] and in sequences of speech syllables [2]. However, this ‘law’ has been found to vary crosslinguistically [3; 4]. ITL effects can arguably better be thought of as the result of the cue distribution for the orthogonal dimensions of prominence (prominent syllables are louder and longer) and of word-segmentation/phrasing (final syllables in words/phrases consequence are longer and softer), as shown for English [5]. When looked at in this way, greater crosslinguistic consistency is observed than based on the ITL (see [6] for a study of 7 languages): duration is crosslinguistically a robust cue for prominence; intensity for grouping. The present perception study explores parsing along two dimensions in more detail in Mandarin, and asks two new research questions: (i) Are the cues to prominence and segmentation in sound sequences similarly interpreted when the stimuli are actual words rather than nonce syllables? (ii) How do acoustic cues compare and interact with phonological cues to segmentation and prominence, in particular, the cue for tone sandhi?

Methods. In Mandarin, when a word carrying the third tone (falling-rising) is followed by another third tone, its tone can turn into Tone 2 (rising) [7]. This process, referred to as third tone (T3) sandhi, only applies when the two words are syntactically and prosodically close to each other [8]. Tone sandhi therefore potentially provides an effective cue for segmentation, and for prominence (cf. [9] on Southern Min). We selected disyllabic Mandarin words with underlying Tone 3 on each syllable that were ‘reversible’, in which swapping the syllables still formed an existing word, albeit with a different meaning (e.g., *kʰow-T3.kʰɤ-T3* ‘thirsty’ and *kʰɤ-T3.kʰow-T3* ‘delicious’). We used four such reversible word pairs, roughly matched for lexical frequency, and created ambiguous sequences of repetitions of each word by alternating the two syllables, varying whether they were pronounced with sandhi (e.g. *kʰɤ-T2.kʰow-T3.kʰɤ-T2.kʰow-T3…*, or *kʰɤ-T3.kʰow-T2.kʰɤ-T3.kʰow-T2…*; the latter sequence corresponds to sandhi applied in the word with a reversed syllable order), or without (e.g. *kʰɤ-T3.kʰow-T3.kʰɤ-T3.kʰow-T3…*). For the acoustic manipulation, we manipulated the sequence such that alternate syllables differed in duration or intensity, in addition to a baseline condition where their acoustic features were matched. Underlying order was fully counterbalanced. Twenty native speakers of Mandarin took part remotely. On a given trial, participants listened to the generated sequences and were asked (1) to choose which of the two words they heard; (2) which of the two syllables they heard as carrying greater emphasis.

Results. We fitted mixed-effects logistic regression models for each decision using the package *lme4* in R [10]. Both intensity and duration cued prominence ($p < 0.001$ for both), but only intensity was a robust cue to grouping ($p < 0.001$). Furthermore, each decision depended on the other (prominence was a significant predictor in the model for grouping, and vice versa, $p < 0.01$). These results replicate the previous findings for Mandarin based on nonsense syllables [5]. The data also shows that tone sandhi was indeed a very strong cue for grouping ($\beta = 6.38$, $p < 0.001$), and not for prominence ($\beta = -0.93$, $p = 0.08$). Furthermore, the acoustic effects were modulated by tone sandhi: acoustic differences were attributed entirely to prominence if tone sandhi, as the stronger grouping cue, already settled the phrasing (in the model, tone sandhi significantly interacts with acoustic cues in predicting the prominence decision, $p < 0.05$). These results make sense if listeners parse the signal along two, in principle orthogonal dimensions, grouping and prominence, which make use of overlapping cues. They also illustrate that to understand one decision (say, grouping), the other (prominence) needs to be taken into account.

References

[1] Bolton, T. L. (1894). Rhythm. *The American Journal of Psychology, 6*(2), 145–238.
[2] Hay, J. S., & Diehl, R. L. (2007). Perception of rhythmic grouping: Testing the iambic/trochaic law. *Perception & psychophysics*, *69*(1), 113-22.
[3] Bhatara, A., Boll-Avetisyan, N., Unger, A., Nazzi, T., & Höhle, B. (2013). Native language affects rhythmic grouping of speech. *The Journal of the Acoustical Society of America*, *134*(5), 3828-43.
[4] Iversen, J. R., Patel, A. D., & Ohgushi, K. (2008). Perception of rhythmic grouping depends on auditory experience. *The Journal of the Acoustical Society of America*, *124*(4), 2263-71.
[5] Wagner, M. (2021). Two-dimensional parsing explains the iambic-trochaic law. *Psychological Review.*
[6] Wagner, M., Iturralde Zurita, A., and Zhang, S. (2021). Parsing speech for grouping and prominence, and the typology of rhythm. *Proceedings of Interspeech*, 2656–2660.
[7] Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, *27*(1), 153-201.
[8] Chen, M. (2000). *Tone sandhi: Patterns across Chinese dialects.* Cambridge Univeristy Press.
[9] Yiu, S. (2019). Iambic and trochaic rhythm in jieyang (southern min). *Proceedings of the Annual Meetings on Phonology*, *7.*
[10] Bates, D., Machler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67 (1), 1–48.