

Is Cantonese Tone Perception Influenced by the Presence of a Simple Beat Gesture?

Yitian Hong¹, Patrick Louis Rohrer², Hans Rutger Bosker³

¹The Chinese University of Hong Kong, Hong Kong SAR, ²Serra Húnter fellow, Universitat Rovira i Virgili Tarragona, Spain, ³Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, The Netherlands

yitianhong@cuhk.edu.hk, patricklouis.rohrer@urv.cat, hansrutger.bosker@donders.ru.nl

Speech communication is multimodal. Speakers not only adopt acoustic cues to indicate a word's prominence but also employ gestures. For non-tone languages, producing a manual gesture causes lengthening of word duration and lowering of F2, effects that are similar to those of prosodic accents [1]. Recent studies on tone languages (Mandarin [2] and Cantonese [3]) revealed that producing a beat gesture significantly increased F0 in monosyllabic words. This raises the question whether listeners also use gesture information in lexical tone perception. Previous research has shown that listeners compensate for the presence of a beat gesture in speech perception. For example, Dutch listeners perceived the same vowel length as a short vowel when presented *with* a gesture, but as a long vowel if *without* a gesture [4]. In our study, we tested the effect of gesture on Cantonese lexical tone perception. Cantonese listeners have been shown to use contextual information in tone perception in a compensatory manner: higher F0 context elicited fewer high tone responses for subsequent targets [5]. If they likewise compensate for gestural cues in tone perception, we predicted that a tone contour would be perceived as relatively lower when co-occurring with a beat gesture (vs. without).

Cantonese native-speakers (N=65, mean age = 23.5 ± 5.2) participated in an online audiovisual perception experiment. They were presented with seven-step lexical tone continua on nine Cantonese word pairs, ranging from T3 (mid-level tone) to T1 (high-level tone). The continua were combined with videos of a speaker who was either producing a simple downward beat gesture (gesture condition) or resting hands at both sides of her thighs (no gesture condition). The speaker's face was masked to prevent articulatory cues to tone perception (Figure 1). The participants were asked to indicate which word they perceived: the word bearing T1 or the word bearing T3. Each participant completed 138 target trials (9 word pairs $\times$ 7 steps $\times$ 2 conditions) and 12 catch trials for monitoring their attention to the videos.

Sixty-two participants' data were analyzed (three excluded for failing catch trials). T1 responses were coded as 1, T3 as 0. A Generalized Linear Mixed Model (GLMM) was run with Step (centered), Gesture Condition (dummy coded; no-gesture as reference level) and their interaction as fixed effects; Participants and Items as random intercepts; and random slopes for Step by Participant and by Item. Contrary to our predictions, Gesture Condition did not show a main effect. However, we observed a significant interaction between Step and Gesture condition ($\beta = 0.076$, $SE = 0.038$, $z = 2.00$, $p = 0.0454^*$), suggesting that gesture modulates tone perception differently across continuum steps. As shown in Figure 2, only the lower continuum steps (e.g., step 1-2) shows the predicted pattern: listeners compensated for the gesture-induced expectation of a higher F0 and thus gave fewer T1 responses.

Overall, the lack of main effect of gesture is inconsistent with our predictions, suggesting that the effect of gestures is more likely to be local (emerging only for a subset of stimuli) than global. On the other hand, Cantonese production studies have shown that duration and intensity, rather than F0, are the primary acoustic correlates of prosodic prominence [6, 7], and F0 is especially unlikely to cue prominence for T1 words in perception [6]. Cantonese speakers may therefore have limited experience linking prominence with F0 raising, which could explain the absence of a main effect of gesture. These findings contribute to our understanding of speech-gesture relationships from a cross-linguistic perspective.

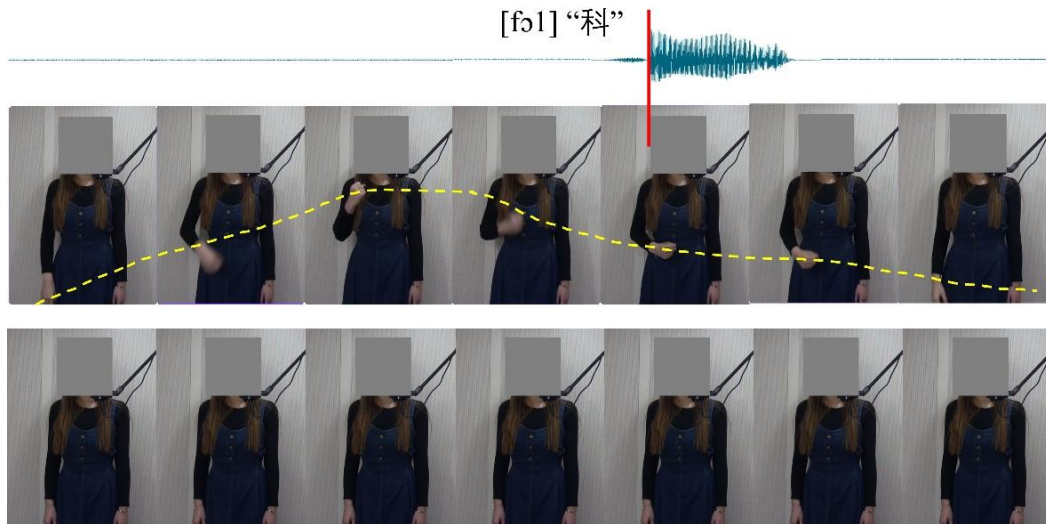


Fig. 1. Example of audiovisual stimuli. The vowel onset (red vertical line) was aligned to the apex of the gesture naturally produced by the speaker.

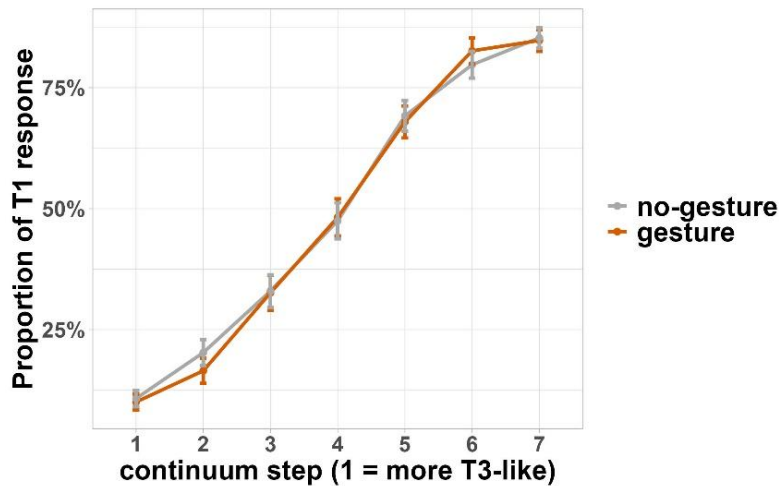


Fig. 2. Audiovisual perception of Cantonese tones

References

- [1] E. Krahmer and M. Swerts, “The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception,” *Journal of Memory and Language*, vol. 57, no. 3, pp. 396–414, Oct. 2007, doi: 10.1016/j.jml.2007.06.005.
- [2] P. L. Rohrer, Y. Hong, and H. R. Bosker, “Gestures time to vowel onset and change the acoustics of the word in Mandarin,” presented at the Speech Prosody 2024, 2024, pp. 866–870. doi: 10.21437/SpeechProsody.2024-175.
- [3] P. L. Rohrer, Y. Hong, and H. R. Bosker, “Beat gestures align with vowel onsets in word-initial syllables and affect speech acoustics in Cantonese monosyllabic utterances,” in *Book of Abstracts PaPE 2025. 6th Phonetics and Phonology in Europe*, Palma, Spain: Institut de Recerca i Innovació Educativa (IRIE).
- [4] H. R. Bosker and D. Peeters, “Beat gestures influence which speech sounds you hear,” *Proc Biol Sci*, vol. 288, no. 1943, p. 20202419, Jan. 2021, doi: 10.1098/rspb.2020.2419.
- [5] M. J. Sjerps, C. Zhang, and G. Peng, “Lexical tone is perceived relative to locally surrounding context, vowel quality to preceding context,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 44, no. 6, pp. 914–924, Jun. 2018, doi: 10.1037/xhp0000504.
- [6] W. L. Wu and Y. Xu, “Prosodic focus in Hong Kong Cantonese without post-focus compression,” in *Speech Prosody 2010*, ISCA, May 2010, p. paper 040-0. doi: 10.21437/SpeechProsody.2010-85.
- [7] H. S. H. Fung and P. P. K. Mok, “Temporal coordination between focus prosody and pointing gestures in Cantonese,” *Journal of Phonetics*, vol. 71, pp. 113–125, Nov. 2018, doi: 10.1016/j.wocn.2018.07.006.