
The role of gestural timing in audiovisual speech perception

Hans Rutger Bosker

Donders Center for Cognition, Radboud University, Nijmegen

hansrutger.bosker@donders.ru.nl

Spoken communication typically involves much more than just sound: we move our heads, hands, and bodies when we speak. The vast majority of these movements do not carry substantial semantic value (Shattuck-Hufnagel & Ren 2018). For instance, we may flick our hands in a lively conversation, without any particular meaning being discernible from that movement alone.

Even though such non-referential co-speech gestures come in many forms (e.g., a head nod, eyebrow raise, finger flick, or beat gesture), they all typically show a temporal relationship with speech prosody. For instance, the apex of beat gestures tends to be aligned to the pitch peak of stressed syllables (Leonard & Cummins 2011). However, although the *average* asynchrony between apex and pitch peak approximates zero, the *standard deviation* is reported to be greater than 400 ms (Pouw & Dixon 2019), reflecting substantial variation in gesture-speech synchrony, possibly within as well as between speakers.

This raises the question (i) how gesture-speech synchrony surfaces in languages that do not have stress, such as tone languages; (ii) how humans perceive variation in gesture-speech synchrony; (iii) what is its role in speech perception. In this talk, I will present recent work from both Mandarin and Cantonese, where gestures *do not* time to pitch peaks, but to vowel onsets instead. Furthermore, I will show evidence for between-speaker differences in gesture-speech synchrony (e.g., Talker A is generally ‘late’, but B is generally ‘early’). In turn, listeners can be shown to track these talker-specific timing signatures, even changing which words listeners think the talker is saying. That is, the same video stimulus can be perceived differently, depending on what you know about how this talker tends to time their gestures.

Together, these findings reveal that the rhythmic movements we make when speaking can actually help our audience, for instance in word recognition and speech segmentation. Therefore, even though non-referential co-speech gestures carry minimal semantic meaning, their timing contributes *meaningfully* to audiovisual spoken language comprehension.

References: • Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10), 1457–1471, doi:10.1080/01690965.2010.500218. • Pouw, W., & Dixon, J. A. (2019). Quantifying gesture-speech synchrony. *Proceedings of the 6th Gesture and Speech in Interaction [GESPIN] Conference*, 75–80, doi:10.17619/UNIPB/1-815. • Shattuck-Hufnagel, S., & Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology*, 9, doi:10.3389/fpsyg.2018.01514.