

Prominence-lending hand movements can guide word segmentation downstream

Hans Rutger Bosker¹

¹ *Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands*
hansrutger.bosker@donders.ru.nl

Prominence in spoken language relies - in part - on speech prosody. Prosody in spoken communication generally surfaces in the suprasegmental characteristics of speech (e.g., modulations in fundamental frequency (F0), duration, and intensity), conveying such phenomena as intonation, stress, and rhythm. These phenomena play an important role in spoken language comprehension, influencing word recognition, perceived segmental distinctions, and also word segmentation. For instance, prominent syllables are typically perceived as word-initial by English listeners, thus guiding their segmentation of a novel syllable stream (Johnson & Jusczyk, 2001; Tyler & Cutler, 2009). That is, prominence patterns that surface *locally* on a given syllable shape listeners' segmentation strategies.

Moreover, Dilley and colleagues have demonstrated that prominence patterns *preceding* a critical target sequence can change listeners' segmentation strategies downstream (Dilley et al., 2010). They presented listeners with word lists, ending in lexically ambiguous structures (e.g., “banker, helpful, [tie, murder, bee/timer, derby]”). If the precursor followed a H-L H-L prominence pattern, the final sequence was more likely to be perceived as ending in a monosyllabic word (“bee”; indexing a “tie, murder, bee” segmentation). However, if the precursor followed a L-H L-H prominence pattern, the acoustically identical sequence was more likely to be perceived as ending in a disyllabic word (“derby”; indexing a “timer, derby” segmentation). This *distal rhythm effect* thus demonstrates that distal rhythmic prominence patterns are a powerful factor in word segmentation and lexical access.

Nevertheless, spoken language is a multimodal phenomenon: most of our everyday spoken communication involves speech with concurrent visual cues such as lip movements, head nods, and hand gestures. These visual cues support speech perception and also convey multimodal prominence. For instance, the timing of simple up-and-down hand movements can – like spoken prosody – guide word recognition (Rohrer et al., 2025), mediate vowel perception (Bosker & Peeters, 2021), and facilitate lexical access (Bujok et al., in press). However, no evidence to date exists that human manual gestures can guide word segmentation.

This study tested whether simple up-and-down hand movements in a precursor can influence segmentation of a syllable sequence downstream. Following Dilley et al. (2010)'s distal rhythm paradigm, 74 native speakers of US English watched videos of a female avatar saying short word lists while producing simple up-and-down hand movements. On experimental trials, word lists ended in lexically ambiguous structures that were (pretested to be) acoustically ambiguous between ending in a monosyllabic word (“...tie, murder, bee”) or a disyllabic word (“...timer, derby”). The avatar produced beat gestures in two possible conditions (counter-balanced across experimental items): either lending prominence to word-initial syllables in the precursor (banker, helpful...), mirroring a H-L H-L pattern, or to word-final syllables in the precursor, mirroring a L-H L-H pattern (see Figure 1). Note that the two gestural prominence patterns were created by aligning gesture apexes to exactly those syllables that had carried *acoustic* prominence in the two conditions in Dilley et al. (2010), mirroring their manipulation in the precursor as well as the target windows. Hence, our critical manipulation forms a direct gestural counterpart of the acoustic distal rhythms in Dilley et al. (2010). In each session, 30 experimental trials were interspersed between 90 filler trials without any lexical ambiguities. Participants' task was to freely type out the last word in each word list. Word responses indicated that participants were more likely to perceive a monosyllabic list-final word (“bee”) in the H-L H-L condition, but a disyllabic word (“derby”) in the L-H L-H condition (see Figure 2). Note that the acoustic speech signal, the avatar's lip movements, and the avatar's gestures *in the final syllables* were identical in the two conditions. They only differed in the timing of the gestures in the distal precursor, thus demonstrating a *gestural distal rhythm effect*.

The current results are the first to present effects of manual movements on *word segmentation*. Moreover, the present finding invites future research into other kinds of visual cues, such as head nods (de la Cruz-Pavía et al., 2019), eyebrow raises, or even non-human signals, like a bouncing disc (Maran et al., 2025). Thus, this study emphasizes the impact of the multimodal nature of prominence in spoken language.

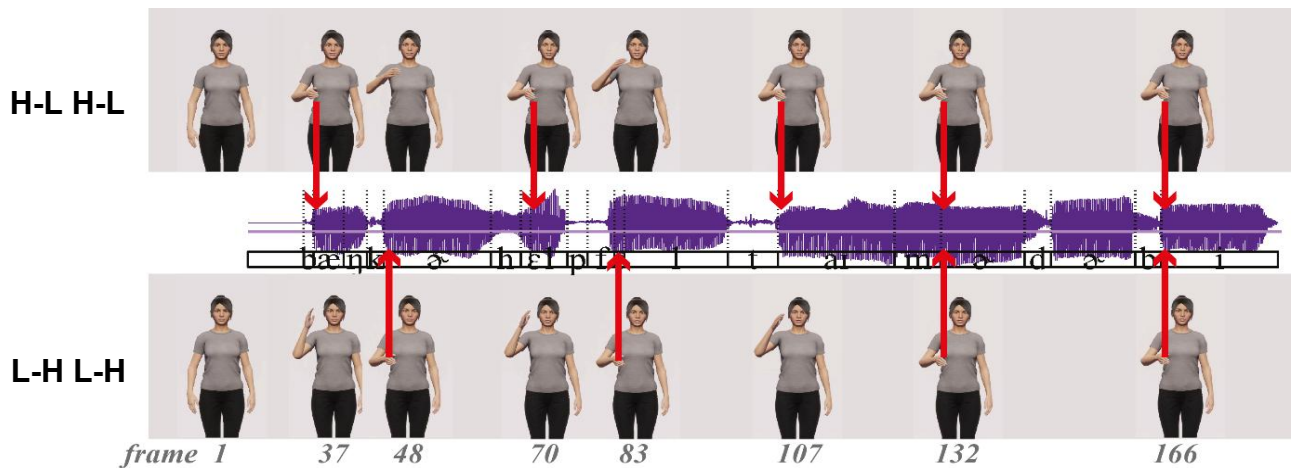


Figure 1. Example of audiovisual avatar in two conditions. A virtual avatar was animated to gesture and speak along with spoken word lists (here: banker, helpful, [tie, murder, bee/timer, derby]) that were acoustically ambiguous between two possible segmentations. The two Beat conditions critically differed in their gesture-speech alignment (H-L H-L vs. L-H L-H), with beat gestures aligned to those syllables that had carried acoustic prominence in the two conditions in Dilley et al. (2010). Critically, the beat gestures in the final target sequence window (e.g., /mæ.də.bi/) were identical between the two conditions. The lip sync was identical across conditions throughout the entire stimulus. Hence, any difference in segmentation of the list-final syllable sequence can only be attributed to differences between conditions in the distal visual prominence patterns.

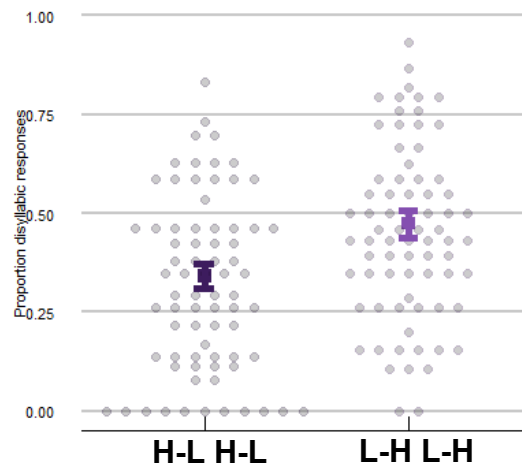


Figure 2. Results illustrated in proportion of disyllabic responses, split by condition. The H-L H-L condition (dark purple) has a significantly lower proportion of disyllabic responses (e.g., fewer “derby”, more “bee”) than the L-H L-H condition (light purple). That is, the same audio combined with the same gestures during the target sequence induced different speech segmentation depending on the distal gestural rhythm. Gray points are individual participants; error bars give 95% CI.

References:

- Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B*, 288, 20202419. <https://doi.org/10.1098/rspb.2020.2419>
- Bujok, R., Maran, M., Meyer, A., & Bosker, H. R. (in press). Beat gestures facilitate lexical access in constraining sentence contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- de la Cruz-Pavía, I., Werker, J. F., Vatikiotis-Bateson, E., & Gervain, J. (2019). Finding Phrases: The Interplay of Word Frequency, Phrasal Prosody and Co-speech Visual Information in Chunking Speech by Monolingual and Bilingual Adults. *Language and Speech*, 002383091984235.
- Dilley, L. C., Mattys, S. L., & Vinke, L. (2010). Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Memory and Language*, 63, 274–294.
- Johnson, E. K., & Jusczyk, P. W. (2001). Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics. *Journal of Memory and Language*, 44(4), 548–567.
- Maran, M., Rötjes, R., Schreurs, A., & Bosker, H. R. (2025). Beat gestures made by human-like avatars affect speech perception. *Proceedings of Interspeech 2025*, 5038–5042.
- Rohrer, P. L., Bujok, R., Van Maastricht, L., & Bosker, H. R. (2025). From “I dance” to “she danced” with a flick of the hands: Audiovisual stress perception in Spanish. *Psychonomic Bulletin & Review*, 32, 2136–2145.
- Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *The Journal of the Acoustical Society of America*, 126(1), 367–376. <https://doi.org/10.1121/1.3129127>