Seeing a beat gesture can change what speech sounds you hear

Hans Rutger Bosker *(Max Planck Institute for Psycholinguistics)* and David Peeters *(Max Planck Institute for Psycholinguistics; Tilburg University).*
HansRutger.Bosker@mpi.nl

Natural languages are evolutionarily designed for face-to-face communication involving both the *auditory* (speech) and the *visual* modality (e.g., a speaker's face and hand gestures). Intriguingly, facial visual information contributes to auditory processing. A seminal example is the McGurk effect: hearing /ba/ but seeing a talker's lips say /ga/ makes listeners perceive the 'intermediate' /da/ [1]. What we hear is hence influenced by what we see.

The current study introduces the manual McGurk effect. In two experiments, we tested whether what listeners hear is directly influenced by the manual beat gestures a speaker makes while speaking. Beat gestures are commonly observed spontaneous biphasic movements of the hand that speakers produce to highlight relevant points in speech [2]. In contrast to iconic gestures, beat gestures are not related to the spoken message on a semantic level, but only on a temporal level: they tend to occur on lexically stressed syllables carrying prosodic emphasis [3]. Empirical evidence that these beat gestures actually influence speech perception is scarce. Could these simple 'flicks of the hand' influence what words we hear?

Experiment 1 tested the *explicit* perception of lexical stress. Twenty native speakers of Dutch were presented with an audiovisual speaker (cf. Figure 1; facial features masked) who produced a carrier sentence (*Nu zeg ik het woord...* "Now I say the word...") followed by 1 of 12 target pseudowords (e.g., *bagpif*). Target pseudowords were sampled from a 7-step acoustic continuum varying F0 independently for the two syllables (e.g., step 1 is most *BAGpif*-like; step 7 is most *bagPIF*-like). Crucially, in the audiovisual block, the speaker produced a manual beat gesture that was either aligned to the onset of the 1st vs. 2nd vowel (cf. Figure 1). The participants' task was to select on each trial (168 trials per block) which pseudoword they heard (e.g., *BAGpif* or *bagPIF*; 2-alternative forced choice task; 2AFC). In the auditory-only control block, only auditory stimuli and no video was presented (block order counter-balanced). Results indicated a manual McGurk effect (see Figure 2A): in the audiovisual block (solid lines), a linear mixed model with logistic linking function (glmer in lme4) on the binomial categorization data showed that participants were significantly more likely to select the pseudoword *BAGpif* (vs. *bagPIF*) when the beat gesture occurred on the 1st (vs. 2nd) syllable, and *vice versa* (β=1.142, p<0.001). No difference was observed between the two auditory conditions in the auditory-only control block (transparent lines).

Experiment 2 tested the *implicit* perception of lexical stress by asking listeners to categorize the first vowel of the target pseudowords. Stressed syllables typically have longer vowels than unstressed syllables. We hypothesized that listeners might use this knowledge in their perception of vowel length. In Dutch, the /ɑ-a:/ vowel contrast is cued by both duration and the second formant (F2): /ɑ/ is shorter and has a lower F2 than /a:/. We manipulated the pseudowords to contain a first vowel that was ambiguous between /ɑ/ and /a:/ by setting the duration to a fixed ambiguous value and varying the F2 on a 5-step continuum. Moreover, prosodic cues to lexical stress (F0, amplitude, syllable durations) were set to ambiguous values. These manipulated pseudowords were spliced into the position of the targets in the audiovisual stimuli from Experiment 1. Participants categorized the sentence-final targets as having either a short /ɑ/ or a long /a:/ as first vowel (e.g., *bagpif* vs. *baagpif*). Results indicated that when the speaker produced a beat gesture on the 1st syllable, participants *implicitly* perceived lexical stress on the 1st syllable, rendering the ambiguous vowel in the 1st vowel as relatively short for a stressed syllable, leading to a small but robust decrease in the proportion of long /a:/ responses (vs. beat on 2nd syllable; β=-0.240, p=0.023).

Our findings thus introduce the manual McGurk effect: beat gestures influence the explicit and implicit perception of lexical stress, in turn changing what vowels people hear. These findings further confirm that listeners integrate visual (e.g., hand gestures) and auditory (speech) cues from multiple modalities during online language comprehension.

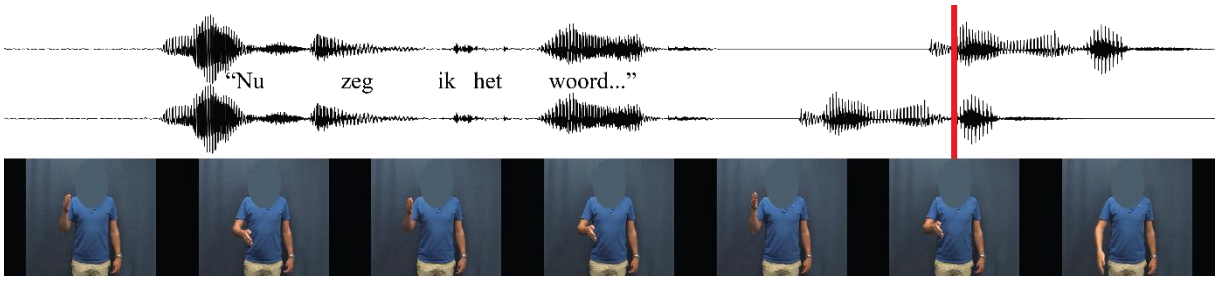

Figure 1. Example of audiovisual stimuli used in Experiment 1 and 2. On each trial, participants were presented with an audiovisual speaker (facial features masked) producing the Dutch carrier sentence *Nu zeg ik het woord...* "Now I say the word...", followed by a disyllabic target pseudoword (e.g., *bagpif*). Critically, the speaker performed two beat gestures in the carrier sentence (apex aligned to onsets of underlined vowels) and one in the target time window. Two auditory conditions were created by aligning the onset of either the first (top waveform) or the second vowel of the target pseudoword (bottom waveform) to the gesture apex (red vertical line). In Experiment 1, participants indicated where they perceived the lexical stress in the sentence-final target word. In Experiment 2, participants indicated whether they perceived the target's first vowel as short /ɑ/ or as long /a:/

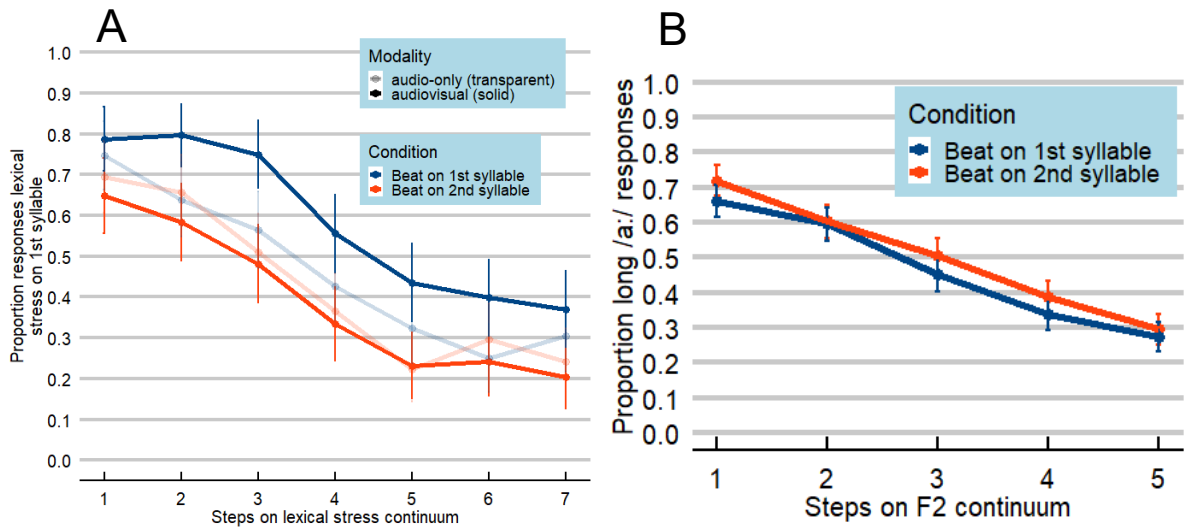

Figure 2. Results from Experiment 1 and 2. A. Proportion of trials in which participants reported perceiving lexical stress on the 1st syllable (e.g., *BAGpif* vs. *bagPIF*). Each target pseudoword was sampled from a 7-step continuum varying F0 independently for the two syllables. In the audiovisual block (solid lines), participants were significantly more likely to report perceiving lexical stress on the 1st syllable if the speaker produced a beat gesture on the 1st syllable (in blue) vs. 2nd syllable (in red). No such difference was observed in the auditory-only block (transparent lines). **B.** Proportion of long /a:/ reports (e.g., *baagpif* vs. *bagpif*). Each target word contained a first vowel that was ambiguous between /ɑ/ and /a:/ (fixed vowel duration, varying F2 on a 5-step continuum). Moreover, prosodic cues to lexical stress (F0, amplitude, syllable duration) were set to ambiguous values. Participants were more likely to report perceiving a long /a:/ as the first target vowel if the speaker produced a beat gesture on the 2nd syllable (in red) vs. 1st syllable (in blue).

REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, p. 746, Dec. 1976, doi: 10.1038/264746a0.

[2] D. McNeill, *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press, 1992.

[3] T. Leonard and F. Cummins, "The temporal relation between beat gestures and speech," *Language and Cognitive Processes*, vol. 26, no. 10, pp. 1457–1471, Dec. 2011, doi: 10.1080/01690965.2010.500218.