PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Bosker HR, Peeters D. 2021 Beat gestures influence which speech sounds you hear. *Proc. R. Soc. B* **288**: 20202419. https://doi.org/10.1098/rspb.2020.2419

Received: 1 October 2020
Accepted: 10 December 2020

Subject Category:
Neuroscience and cognition

Subject Areas:
behaviour, cognition, neuroscience

Keywords:
audiovisual speech perception, manual McGurk effect, multimodal communication, beat gestures, lexical stress

Author for correspondence:
Hans Rutger Bosker
e-mail: hansrutger.bosker@mpi.nl

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5251460.

THE ROYAL SOCIETY PUBLISHING

Beat gestures influence which speech sounds you hear

Hans Rutger Bosker[1,2] and David Peeters[3]

[1]Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands
[3]Department of Communication and Cognition, TiCC Tilburg University, Tilburg, The Netherlands

HRB, 0000-0002-2628-7738

Beat gestures—spontaneously produced biphasic movements of the hand—are among the most frequently encountered co-speech gestures in human communication. They are closely temporally aligned to the prosodic characteristics of the speech signal, typically occurring on lexically stressed syllables. Despite their prevalence across speakers of the world's languages, how beat gestures impact spoken word recognition is unclear. Can these simple 'flicks of the hand' influence speech perception? Across a range of experiments, we demonstrate that beat gestures influence the explicit and implicit perception of lexical stress (e.g. distinguishing *OBject* from *objECT*), and in turn can influence what vowels listeners hear. Thus, we provide converging evidence for a manual McGurk effect: relatively simple and widely occurring hand movements influence which speech sounds we hear.

1. Background

The human capacity to communicate evolved primarily to allow for efficient face-to-face interaction [1]. As humans, we are indeed capable of translating our thoughts into communicative messages, which in everyday situations often comprise concurrent auditory (speech) and visual (lip movements, facial expressions, hand gestures, body posture) information [2–4]. Speakers thus make use of different channels (mouth, eyes, hands, torso) to get their messages across, while addressees have to quickly and efficiently integrate or *bind* the different ephemeral signals they concomitantly perceive [2,5]. This process is skilfully guided by top-down predictions [6,7] that exploit a lifetime of communicative experience, acquired world knowledge and personal common ground built up with the speaker [8,9].

The classic McGurk effect [10] is an influential illustration of how visual input influences the perception of speech sounds. In a seminal study [10], participants were made to repeatedly listen to the sound /ba/. At the same time, they watched the face of a speaker in a video whose mouth movements corresponded to the sound /ga/. When asked what they heard, a majority of participants reported actually perceiving the sound /da/. This robust finding (yet see [11,12]) indicates that during language comprehension our brains take both verbal (here: auditory) and non-verbal (here: visual) aspects of communication into account [13,14], providing the addressee with a best guess of what exact message a speaker intends to convey. We thus listen not only with our ears, but also with our eyes.

Visual aspects of everyday human communication are not restricted to the subtle mouth or lip movements. In close coordination with speech, the hand gestures we spontaneously and idiosyncratically make during conversations help us express our thoughts, emotions and intentions [15]. We manually point at things to guide the attention of our addressee to relevant aspects of our immediate environment [16], enrich our speech with iconic gestures that in handshape and movement resemble what we are talking about [17–19], and produce simple flicks of the hand aligned to the acoustic prominence of our spoken utterance to highlight relevant points in speech [20–23]. Over the

past decades, it has become clear that such co-speech manual gestures are often semantically aligned and closely temporally synchronized with the speech we produce [24,25]. It is an open question, however, whether the temporal synchrony between these hand gestures and the speech signal can actually influence which speech sounds we hear.

In this study, we test for the presence of such a potential *manual* McGurk effect. Do hand gestures, like observed lip movements, influence what exact speech sounds we perceive? We here focus on manual *beat gestures*—commonly defined as biphasic (e.g. up and down) movements of the hand that speakers spontaneously produce to highlight prominent aspects of the concurrent speech signal [15,20]. Beat gestures are among the most frequently encountered gestures in naturally occurring human communication [15] as the presence of an accompanying beat gesture allows for the enhancement of the perceived prominence of a word [20]. Not surprisingly, they hence also feature prominently in politicians' public speeches [26]. By highlighting concurrent parts of speech and enhancing its processing [27], beat gestures may even increase memory recall of words in both adults and children [28]. As such, they form a fundamental part of our communicative repertoire and also have direct practical relevance.

By contrast to iconic gestures, beat gestures are not so much related to the spoken message on a semantic level; rather, they are most prominently related to the speech signal on a temporal level: their apex is aligned to vowel onset in lexically stressed syllables carrying prosodic emphasis [15,22,29]. Neurobiological studies suggest that listeners may tune oscillatory activity in the alpha and theta bands upon observing the initiation of a beat gesture to anticipate processing an assumedly important upcoming word [26]. Bilateral temporal areas of the brain may in parallel be activated for efficient integration of beat gestures and concurrent speech [27]. However, evidence for behavioural effects of temporal gestural alignment on speech perception remains scarce. Given the close temporal alignment between beat gestures and lexically stressed syllables in speech production, ideal observers [30] could in principle use beat gestures as a cue to lexical stress in perception.

Lexical stress is indeed a critical speech cue in spoken word recognition. It is well established that listeners commonly use suprasegmental spoken cues to lexical stress, such as greater amplitude, higher fundamental frequency (F0) and longer syllable duration, to constrain online lexical access, speeding up word recognition and increasing perceptual accuracy [31,32]. Listeners also use visual articulatory cues on the face to perceive lexical stress when the auditory signal is muted [33]. Nevertheless, whether non-facial visual cues, such as beat gestures, also aid in lexical stress perception and might even do so in the presence of auditory cues to lexical stress—as in everyday conversations—is unknown. These questions are relevant for our understanding of face-to-face interaction, where multimodal cues to lexical stress might considerably increase the robustness of spoken communication, for instance in noisy listening conditions [34], and even when the auditory signal is clear [35].

Indeed, a recent theoretical account proposed that multimodal low-level integration must be a general cognitive principle fundamental to our human communicative capacities [2]. However, evidence showing that the integration of information from two separate communicative modalities (speech, gesture) modulates the mere perception of information derived from one of these two modalities (e.g. speech) is lacking. For instance, earlier studies have failed to find the effects of simulated pointing gestures on lexical stress perception, leading to the idea that manual gestures provide information perhaps only about sentential emphasis; not about the relative emphasis placed on individual syllables within words [36]. As a result, there is no formal account or model of how audiovisual prosody might help spoken word recognition [37–39]. Hence, demonstrating a manual McGurk effect would imply that existing models of speech perception need to be fundamentally expanded to account for how multimodal cues influence auditory word recognition.

Below, we report the results of a set of experiments (and two direct replications: experiments S1–S2 in the electronic supplementary material) in which we tested for the existence of a manual McGurk effect. The experiments each approach the theoretical issue at hand from a different methodological angle, varying the degree of explicitness of the participants' experimental task and their response modality (perception versus production). Experiments 1A–1B use an explicit task to test whether observing a beat gesture influences the perception of lexical stress in disyllabic pseudowords (1A) and real words (1B). Using the same stimuli as experiment 1A, experiment 2 investigates *implicit* lexical stress perception by means of a two-group design. First, participants in Group A were asked to vocally shadow (i.e. repeat back) the pseudowords the audiovisual speaker produced. Participants in Group B, in turn, listened to the shadowed productions from Group A, categorizing the auditory stimuli as having lexical stress on the first or second syllable. We predicted that Group B would be more likely to perceive a shadowed production from Group A to have lexical stress on the first syllable if the shadowed production itself had been produced in response to an audiovisual pseudoword with a beat gesture on the first syllable.

Finally, experiment 3 tests the strongest form of a manual McGurk effect: can beat gestures influence the perceived identity of individual speech sounds? The rationale behind experiment 3 is grounded in the observation that listeners take a syllable's perceived prominence into account when estimating its vowel's duration [40,41]. In speech production, vowels in stressed syllables typically have a longer duration than vowels in unstressed syllables. Hence, in perception, listeners expect a vowel in a stressed syllable to have a longer duration. Therefore, when an ambiguous vowel duration is presented in a *stressed* syllable, it mismatches this expectation, resulting in a bias to perceive it as unexpectedly short. Conversely, if the ambiguous vowel duration is presented in an *unstressed* syllable, it mismatches the expectation of unstressed vowels having relatively short vowels, inducing a bias to perceive it as relatively long. This effect of prosodic structure on vowel duration has been known since the 1980s [40], with implications for languages where vowel duration is a critical cue to phonemic contrasts (e.g. to coda stop voicing in English, e.g. 'coat—code' [41]; to vowel length contrasts in Dutch, e.g. *tak* /tɑk/ 'branch'—*taak* /ta:k/ 'task' [42]). Based on this literature, prosodic expectations should thus influence the perceived identity of speech sounds in a *contrastive* fashion: a Dutch vowel midway between /ɑ/ and /a:/ may be perceived as relatively short /ɑ/ when presented in a syllable carrying audiovisual prominence (i.e. when paired with a beat gesture). Yet, an anonymous reviewer pointed out that the effect of perceived prosodic structure could in principle also function in an *attractive* fashion. Specifically, seeing a beat gesture on the first syllable could in principle also bias

perception towards more long /a:/ responses. Therefore, experiment 3 assesses whether and how beat gestures influence the implicit perception of lexical stress in Dutch, which—in turn—may affect whether human listeners perceive a short or a long vowel.

2. Method

(a) Participants

Native speakers of Dutch were recruited from the Max Planck Institute's participant pool (experiment 1A: $n = 20$; 14 females (f), 6 males (m); $M_{age} = 25$, range = 20–34; experiment 1B: $n = 48$; 30f, 18 m; $M_{age} = 22$, range = 18–26; experiment 2, Group A: $n = 26$; 22f, 4 m; $M_{age} = 23$, range = 19–30; group B: $n = 26$; 25f, 1 m; $M_{age} = 24$, range = 20–28; experiment 3 was performed by the same individuals as those in Group B). All gave informed consent and all research was performed in accordance with relevant guidelines and regulations (see 'Ethics').

(b) Materials

(i) Experiment 1A and 2

For experiment 1A, we created 12 Dutch disyllabic pseudowords (e.g. *wasol* /ʋa.sɔl/; see electronic supplementary material, table). Following a detailed stimulus creation protocol (steps 1–4; https://doi.org/10.17504/protocols.io.bmw5k7g6), we recorded audiovisual stimuli of a speaker pronouncing a Dutch carrier sentence (*Nu zeg ik het woord …* 'Now say I the word …'), followed by a pseudoword, while producing three beat gestures (see figure 1*a* and example video stimuli at https://osf.io/b7kue). Pseudowords were manipulated to involve (i) 7-step F0 continua, ranging from a 'strong–weak' (SW; step 1) to a 'weak–strong' prosodic pattern (WS; step 7); (ii) two different temporal alignments of the sentence-final beat gesture: its apex was either aligned to the *first* or the *second* vowel onset, by modulating the silent interval preceding the pseudoword (cf. figure 1*a*). Note that the same audiovisual stimuli were used in experiment 1A with an explicit lexical stress categorization task as well as for Group A of experiment 2 with a shadowing task (which itself generated the auditory stimuli for Group B in experiment 2).

(ii) Experiment 1B

To test the generalizability of our findings with pseudowords to more naturalistic stimuli, experiment 1B was identical to experiment 1A except that we used five existing Dutch minimal word pairs that only differ in lexical stress (e.g. *Plato*: /'pla:to/ 'Plato' with stress on the first syllable versus *plateau* /pla:'to/ 'plateau' with stress on the second syllable; see electronic supplementary material, table S2 for the full list). Audiovisual stimuli were created following a similar protocol as in experiment 1A (steps 5–8; https://doi.org/10.17504/protocols.io.bmw5k7g6). Critically, they also involved 7-step F0 continua and the same two different temporal alignments of the sentence-final beat gesture (see example video stimuli at https://osf.io/b7kue).

(iii) Experiment 3

For experiment 3, we created 8 new disyllabic pseudowords, with either a short /ɑ/ or a long /a:/ as first vowel (e.g. *bagpif* /bɑx.pɪf/ versus *baagpif* /ba:x.pɪf/; cf. electronic supplementary material, table S2). Following an adjusted stimulus protocol (steps 9–13; https://doi.org/10.17504/protocols.io.bmw5k7g6), we created acoustic tokens that were ambiguous in their prosodic cues to lexical stress. Moreover, the length of the critical first vowel was set to a fixed ambiguous duration, while varying the second formant (F2) along a 5-step continuum (a secondary cue to the /ɑ-a:/ contrast in Dutch [42]). These tokens were spliced into the two audiovisual beat gesture conditions stimuli used before (see example video stimuli at https://osf.io/b7kue).

(c) Procedure

(i) Experiment 1A

Participants were tested individually in a sound-conditioning booth. They were seated at a distance of approximately 60 cm in front of a 50.8 cm by 28.6 cm computer screen and listened to stimuli at a comfortable volume through headphones. Stimulus presentation was controlled by Presentation software (v.16.5; Neurobehavioural Systems, Albany, CA, USA).

Participants were presented with two blocks of trials: an audiovisual block, in which the video and the audio content of the manipulated stimuli were concomitantly presented, and an audio-only control block, in which only the audio content but no video content was presented. Block order was counter-balanced across participants. On each trial, the participants' task was to indicate whether the sentence-final pseudoword was stressed on the first or the last syllable (2-alternative forced-choice task; 2AFC).

Before the experiment, participants were instructed to always look at the screen during trial presentations. In the audiovisual block, trials started with the presentation of a fixation cross. After 1000 ms, the audiovisual stimulus was presented (video dimensions: 960 × 864 pixels). At stimulus offset, a response screen was presented with two response options on either side of the screen, for instance: *WAsol* versus *waSOL*; positioning of response options was counter-balanced across participants. Participants were instructed that capitals indicated lexical stress, that they could enter their response by pressing the 'Z' key on a regular keyboard for the left option and the 'M' key for the right option, and that they had to give their response within 4 s after stimulus offset; otherwise, a missing response was recorded. After their response (or at time-out), the screen was replaced by an empty screen for 500 ms, after which the next trial was initiated automatically. The trial structure in the audio-only control block was identical to the audiovisual block, except that instead of the video content a static fixation cross was presented during stimulus presentation.

Participants were presented with 12 pseudowords, sampled from 7-step F0 continua, in two beat gesture conditions (onset of first or second vowel aligned to the apex of beat gesture), resulting in 168 unique items in a block. Within a block, the item order was randomized. Before the experiment, four practice trials were presented to participants to familiarize them with the materials and the task. Participants were given the opportunity to take a short break after the first block.

(ii) Experiment 1B

Experiment 1B was run online using Testable (https://testable.org [43]). Each participant was explicitly instructed to use headphones and to run the experiment with undivided attention in quiet surroundings. The procedure of experiment 1B was identical to the procedure used in experiment 1A: participants were presented with two blocks of trials, audiovisual and audio-only, with block order counter-balanced across participants. On each trial, the participants' task was to indicate whether the sentence-final word was stressed on the first or the last syllable (e.g. *PLAto* versus *plaTEAU*; 2AFC). Participants were presented with 5 minimal word pairs, sampled from 7-step F0 continua, in two beat gesture conditions (onset of first or second vowel aligned to the apex of beat gesture), resulting in 70 unique items in a block.

(iii) Experiment 2

Experiment 2 comprised two participant groups A and B, each with a different experimental task. Group A used a similar

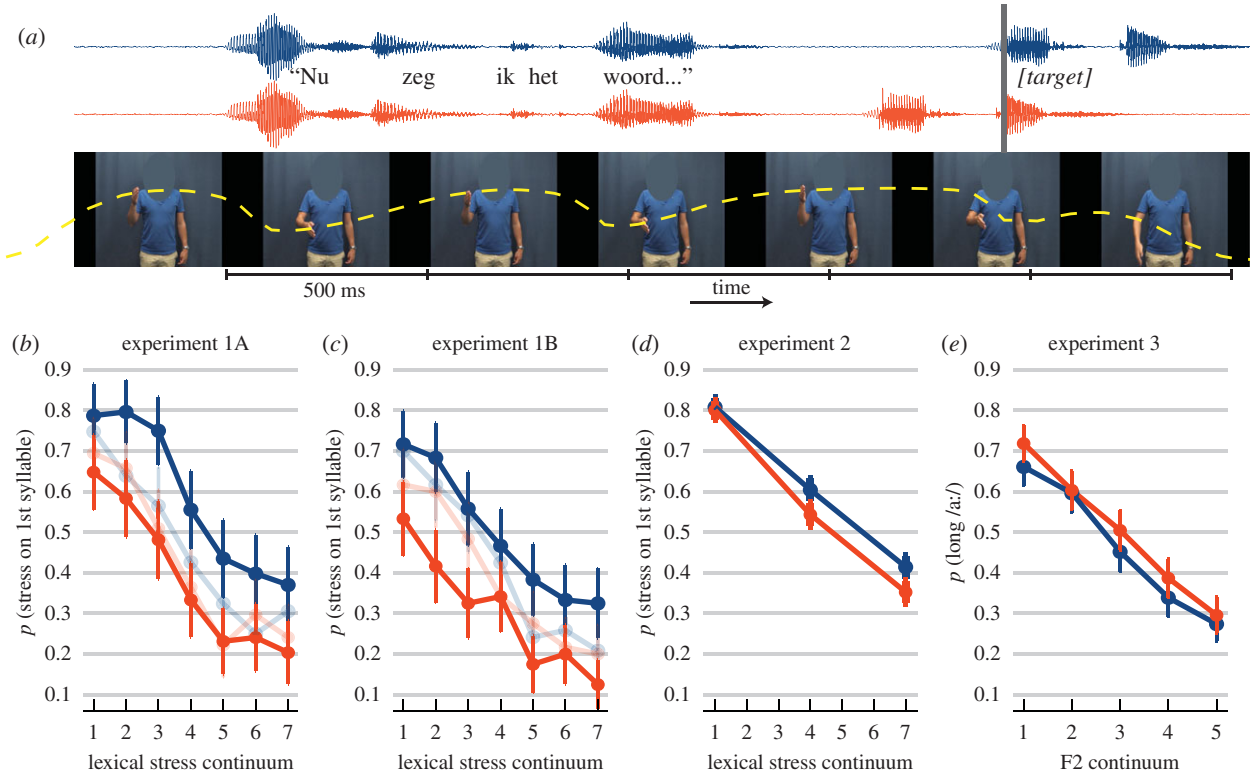

Figure 1. Example audiovisual stimulus and results from experiments 1A, 1B, 2 and 3. (*a*) In all experiments, participants were presented with an audiovisual speaker producing the Dutch carrier sentence *Nu zeg ik het woord* … 'Now say I the word … ' followed by a disyllabic target (e.g. pseudowords like *bagpif*; or real words like *plateau*). The speaker produced three beat gestures (vertical hand trajectory shown in yellow dashed line), with the last beat gesture's apex falling in the target time window. Two auditory conditions were created by aligning the onset of either the first (top waveform in blue) or the second vowel of the target (bottom waveform in orange) to the final gesture's apex (grey vertical line). (*b*) Participants in experiment 1A categorized audiovisual pseudowords as having lexical stress on either the first or the second syllable. Each target pseudoword was sampled from a 7-step continuum (on *x*-axis) varying F0 in opposite directions for the two syllables. In the audiovisual block (solid lines), participants were significantly more likely to report perceiving lexical stress on the first syllable if the speaker produced a beat gesture on the first syllable (in blue) versus the second syllable (in orange). No such difference was observed in the audio-only control block (transparent lines). (*c*) Similar effects were observed when the targets were replaced by real Dutch minimal word pairs in experiment 1B (e.g. *Plato* with stress on the first syllable versus *plateau* with stress on the second syllable) (*d*) Participants in Group B from experiment 2 categorized the audio-only shadowed productions, recorded from Group A, as having lexical stress on either the first or the second syllable. If the audiovisual stimulus used for the shadowing Group A involved a pseudoword with a beat gesture on the first syllable, the elicited shadowed production itself was more likely to be perceived as having lexical stress on the first syllable by Group B (in blue). Conversely, if the audiovisual stimulus for Group A involved a pseudoword with a beat gesture on the second syllable, the elicited shadowed production was more likely to be perceived as having lexical stress on the second syllable by Group B (in orange). (*e*) Participants in experiment 3 categorized audiovisual pseudowords as containing either short /ɑ/ or long /a:/ (e.g. *bagpif* versus *baagpif*). Each target pseudoword contained a first vowel that was sampled from a 5-step F2 continuum from long /a:/ to short /ɑ/ (on the *x*-axis). Moreover, prosodic cues to lexical stress (F0, amplitude, syllable duration) were set to ambiguous values. If the audiovisual speaker produced a beat gesture on the first syllable (in blue), participants were biased to perceive the first syllable as stressed, making the ambiguous first vowel relatively short for a stressed syllable, leading to a lower proportion of long /a:/ responses. Conversely, if the audiovisual speaker produced a beat gesture on the second syllable (in orange), participants were biased to perceive the initial syllable as unstressed, making the ambiguous first vowel relatively long for an unstressed syllable, leading to a higher proportion of long /a:/ responses. For all panels, error bars enclose 1.96 × s.e. on either side; that is, the 95% confidence intervals. (Online version in colour.)

procedure as experiment 1A, except that only audiovisual stimuli were used (i.e. no audio-only control block). Instead of making explicit 2AFC decisions about lexical stress, Group A was instructed to simply repeat back the sentence-final pseudoword after stimulus offset, and to hit the 'Enter' key to move to the next trial. The use of pseudowords (from experiment 1A) rather than real Dutch words (as in experiment 1B) was motivated by potentially interfering effects of word-specific lexical frequency on pronunciation when using real words (i.e. higher frequency words are typically produced with shorter durations and reduced spectral cues [44]). No mention of lexical stress was made in the instructions. Participants were instructed to always look at the screen during trial presentations.

At the beginning of the experiment, participants were seated behind a table with a computer screen inside a double-walled acoustically isolated booth. They wore a circum-aural Sennheiser Game Zero headset with an attached microphone throughout the experiment to ensure that amplitude measurements were made at a constant level. Participants were presented with 12 pseudowords, sampled from step 1, 4 and 7 of the F0 continua (only three steps to reduce the load on the participants), in two beat gesture conditions, resulting in 72 unique items. Each unique item was presented twice over the course of the experiment, leading to 144 trials per participants. Item order was random and item repetitions only occurred after all unique items had been presented. Before the experiment, one practice trial was presented to participants to familiarize them with the task.

Thereafter, Group B was presented with the auditory shadowed productions from Group A, categorizing them as having lexical stress on the first or second syllable. This was run online using PsyToolkit (v.2.6.1 [45]) because of limitations due to the coronavirus pandemic. Each participant was explicitly instructed

to use headphones and to run the experiment with undivided attention in quiet surroundings. Each participant was presented with all shadowed productions from two talkers from Group A, leading to 13 different lists, with two Group B participants assigned to each list. Stimuli from the same talker were grouped to facilitate talker and speech recognition, and presented in a random order. The task was to indicate after stimulus offset whether the pseudoword had stress on the first or the second syllable (2AFC). Two response options were presented, with capitals indicating stress (e.g. *WAsol* versus *waSOL*), and participants used the mouse to select their response. No practice items were presented and the experiment was self-timed. Note that the same participants that formed Group B in experiment 2 also participated in experiment 3. To avoid explicit awareness about lexical stress influencing their behaviour in experiment 3, these participants always first performed experiment 3, and only then took part in the categorization task for experiment 2.

(iv) Experiment 3

Experiment 3 was run online using Testable (https://testable.org [43]). Each participant was explicitly instructed to use headphones and to run the experiment with undivided attention in quiet surroundings. Before the experiment, participants were instructed to always look and listen carefully to the audiovisual talker and categorize the first vowel of the sentence-final target pseudoword (2AFC). Lexical stress was never mentioned in any part of the instructions. Specifically, at stimulus offset, a response screen was presented with two response options on either side of the screen, for instance: *bagpif* versus *baagpif*. Participants entered their response by pressing the 'Z' key on a regular keyboard for the left option and the 'M' key for the right option. The next trial was only presented after a response had been recorded (ITI = 500 ms). Participants were presented with eight pseudowords, sampled from 5-step F2 continua, in two beat gesture conditions (onset of first or second vowel aligned to the apex of beat gesture), resulting in 80 unique items in a block. Each participant received two identical blocks with a unique random order within blocks.

3. Results

(a) Experiments 1A–1B

Experiment 1A tested whether beat gestures influence how listeners explicitly categorize disyllabic spoken stimuli as having initial stress or final stress (e.g. *WAsol* versus *waSOL*). An audiovisual speaker produced beat gestures whose apex was manipulated to be either aligned to the onset of the first vowel of a sentence-final pseudoword, or its second vowel (cf. figure 1*a*). Moreover, experiment 1B tested whether the same effect of gestural alignment on lexical stress perception could be obtained with real Dutch words (e.g. *Plato* /'pla:to/ 'Plato' versus *plateau* /pla:'to/ 'plateau').

Trials with missing data ($n = 7$ from experiment 1A; 0.1%) were excluded from the analysis. Data from both experiments 1A and 1B were combined and statistically analysed using a generalized linear mixed model (GLMM) [46] with a logistic linking function as implemented in the lme4 library (v.1.0.5) in R. The binomial dependent variable was participants' categorization of the target as either having lexical stress on the first syllable (SW; e.g. *WAsol*; coded as 1) or the second syllable (WS; e.g. *waSOL*; coded as 0). Fixed effects were continuum step (continuous predictor; scaled and centred around the mean), beat condition (categorical predictor; deviation coding, with beat gesture on the second syllable coded as −0.5 and beat gesture on the first syllable coded as +0.5), modality (categorical predictor; deviation coding, with audiovisual coded as −0.5 and audio-only coded as +0.5), experiment (categorical predictor; deviation coding, with experiment 1A coded as −0.5 and experiment 1B coded as +0.5), and all interactions. The model also included participant and items as random factors, with by-participant and by-item random slopes for all fixed effects, except experiment (since it was not nested within participants or items).

The model showed a significant effect of continuum step ($\beta = -1.009$, s.e. = 0.124, $z = -8.124$, $p < 0.001$), indicating that higher continuum steps led to lower proportions of SW responses. It also showed an effect of beat condition ($\beta = 0.603$, s.e. = 0.088, $z = 6.831$, $p < 0.001$), indicating that—in line with our hypothesis—listeners were biased towards perceiving lexical stress on the first syllable if there was a beat gesture on that first syllable (cf. blue versus orange solid lines in figure 1*b*; the overall difference in proportions = 0.20). Moreover, an interaction was observed between beat condition and modality ($\beta = -0.748$, s.e. = 0.081, $z = -9.194$, $p < 0.001$), indicating that the effect of beat condition was drastically reduced in the audio-only control block. No other effects or interactions with the experiment were found, suggesting that the findings held equally (cf. figure 1*b*,*c*) across both pseudowords (1A) and real words (1B). Together, these findings demonstrate that the temporal alignment of beat gestures to the speech signal influences the explicit perception of lexical stress. Moreover, the interaction between beat condition and modality indicates that the effect observed in the audiovisual block cannot be attributed to the differential duration of the pre-target silent interval in the two beat conditions. Nevertheless, the task in experiments 1A–1B, asking participants to categorize the audiovisual stimuli as having stress on either the first or second syllable, is very explicit and presumably susceptible to response strategies.

(b) Experiment 2

Therefore, experiment 2 tested lexical stress perception *implicitly* and in another response modality by asking Group A to overtly shadow the audiovisual pseudowords from experiment 1A. Since lexical status did not seem to influence behaviour in experiment 1, we opted for pseudowords, not real words. In addition, we thus avoided potential item-specific lexical frequency effects on word realization (i.e. high-frequency words are typically produced in a more 'reduced' form [44]). Acoustic analysis of the shadowed productions from Group A are reported in detail in the electronic supplementary material. In summary, we observed that when the beat gesture fell on the second syllable, the second syllable was more likely to be perceived as stressed, resulting in louder (+0.5 dB) and longer (+9 ms) second syllables and shorter first syllables (−8 ms) in the shadowed productions, compared to when the beat gesture fell on the first syllable.

Still, the acoustic adjustments in response to the two beat gesture conditions were relatively modest. Therefore, new participants (Group B) assessed the perceptual relevance of these acoustic adjustments produced by Group A. Group B listened to the shadowed productions from Experiment 2 and indicated whether they perceived the shadowed productions to have stress on the first or second syllable (2AFC). Critically, this

involved audio-only stimuli and Group B was unaware of the beat gesture manipulation used for Group A. Similar to experiment 1, the categorization responses (SW = 1; WS = 0) were statistically analysed using a GLMM with a logistic linking function. Fixed effects were continuum step and beat condition (coding adopted from experiment 1), and their interaction. The model included random intercepts for participant, talker and item, with by-participant, by-talker and by-item random slopes for all fixed effects.

The model showed a significant effect of continuum step ($\beta = -0.985$, s.e. = 0.162, $z = -6.061$, $p < 0.001$), indicating that higher continuum steps led to lower proportions of SW responses. Critically, it showed a modest yet statistically significant effect of beat condition ($\beta = 0.232$, s.e. = 0.078, $z = 2.971$, $p = 0.003$), indicating that—in line with our hypothesis—listeners were biased towards perceiving lexical stress on the first syllable if the shadowed production was produced in response to an audiovisual pseudoword with a beat gesture on the first syllable (cf. blue versus orange lines in figure 1*d*; the overall difference in proportions = 0.05). The interaction between continuum step and beat condition was marginally significant ($\beta = 0.110$, s.e. = 0.059, $z = 1.886$, $p = 0.059$), suggesting a tendency for the more ambiguous steps to show a larger effect of the beat gesture.

Still, the effect of beat condition demonstrated by Group B in experiment 2 was relatively modest. Therefore, we ran an identical replication of Group B, namely experiment S1, presenting the same stimuli from Group A to a new participant sample, resulting in very similar findings (cf. electronic supplementary material). Together, these findings suggest that the acoustic adjustments produced by Group A in response to the different beat gesture alignments were perceptually salient enough for new participants to note them and use them as cues to lexical stress identification. Thus, it provides supportive evidence that beat gestures influence *implicit* lexical stress perception.

(c) Experiment 3

Finally, experiment 3 tested the strongest form of a manual McGurk effect: can beat gestures influence the perceived identity of individual speech sounds? We analysed the binomial categorization data (/a:/ coded as 1; /ɑ/ as 0) using a GLMM with a logistic linking function. Fixed effects were continuum step and beat condition (coding adopted from experiment 1), and their interaction. The model included random intercepts for participant and pseudoword, with by-participant and by-item random slopes for all fixed effects.

The model showed a significant effect of continuum step ($\beta = -0.922$, s.e. = 0.091, $z = -10.155$, $p < 0.001$), indicating that higher continuum steps led to lower proportions of long /a:/ responses. Critically, it showed a modest effect of beat condition ($\beta = -0.240$, s.e. = 0.105, $z = -2.280$, $p = 0.023$). This indicates that when the beat gesture was aligned to the first syllable (blue line in figure 1*e*), participants were more likely to perceive the first syllable as stressed, rendering the ambiguous vowel duration in the first syllable as relatively short for a stressed syllable, lowering the proportion of long /a:/ responses. Conversely, if the beat gesture was aligned to the second syllable (orange line in figure 1*e*), participants were more likely to perceive the first syllable as unstressed, making the ambiguous vowel duration in the first syllable relatively long for an unstressed syllable, leading to a small increase in the proportion of long /a:/ responses (overall difference in proportions = 0.04). The effect of the sentence-final beat gesture thus functioned in a *contrastive* fashion, in line with earlier literature on prosodic effects on vowel duration [40,41]. No interaction between continuum step and beat condition was observed ($p = 0.817$).

Still, the beat gesture effect, although statistically significant, was relatively modest. Therefore, we ran an identical replication (experiment S2 in the electronic supplementary material) with a new participant sample. Results from experiment S2 replicated the findings in experiment 3: seeing a beat gesture on the first syllable made participants more likely to report hearing the short vowel /ɑ/. Together, these findings showcase the strongest form of the manual McGurk effect: beat gestures influence which speech sounds we hear.

4. Discussion

To what extent do we listen with our eyes? The classic McGurk effect illustrates that listeners take both visual (lip movements) and auditory (speech) information into account to arrive at a best possible approximation of the exact message a speaker intends to convey [10]. Here, we provide evidence in favour of a *manual* McGurk effect: the beat gestures we see influence what exact speech sounds we hear. Specifically, we observed across a range of experiments that beat gestures influence both the explicit and implicit perception of lexical stress, and in turn, may influence the perception of individual speech sounds. To the best of our knowledge, this is the first demonstration of behavioural consequences of beat gestures on low-level speech perception. These findings support the idea that everyday language comprehension is primordially a multimodal undertaking in which the multiple streams of information that are concomitantly perceived through the different senses are quickly and efficiently integrated by the listener to make a well-informed best guess of what specific message a speaker is intending to communicate.

The experiments we presented made use of a combination of implicit and explicit paradigms. Not surprisingly, the *explicit* nature of the task used in experiments 1A–1B led to strong effects of the timing of a perceived beat gesture on the lexical stress pattern participants reported to hear. The *implicit* effects in subsequent experiments 2–3 were more modest, primarily due to targeting *indirect* effects. For instance, experiment 2 involved the assessment of implicit effects of beat gestures on participants' shadowing performance (Group A), which *in turn* affected the auditory perception of those shadowed productions (Group B). Similarly, in experiment 3, participants saw beat gestures that influenced implicit lexical stress perception, which then *in turn* raised expectations about vowel duration. Hence, effect sizes for such indirect effects are more likely to be modest.

Nevertheless, the effects obtained were robust and replicable (cf. experiments S1–S2 in the electronic supplementary material). Moreover, these beat gesture effects also surfaced in real words (experiment 1B) and even when the auditory cues were relatively clear (i.e. at the extreme ends of the acoustic continua), suggesting that these effects also prevail in naturalistic face-to-face communication outside the laboratory. In fact, the naturalistic nature of the audiovisual stimuli

would seem key to the effect, since the temporal alignment of artificial visual cues (i.e. an animated line drawing of a disembodied hand) does not influence lexical stress perception (see [36]). Perhaps the present study even underestimates the behavioural influence of beat gestures on speech perception in everyday conversations, since all auditory stimuli in the experiments involved 'clean' (i.e. noise-free) speech. However, in real-life situations, the speech of an attended talker often occurs in the presence of other auditory signals, such as background music, competing speech or environmental noise; in such listening conditions, the contribution of visual beat gestures to successful communication would conceivably be even greater. Likewise, presumably, the saliency of the visual signal might also modulate the size of the effect: perhaps even more pronounced (larger and/or faster) manual gestures could further enhance the manual McGurk effect.

The outcomes of the present study challenge current models of (multimodal) word recognition, which would need to be extended to account for how prosody, specifically multimodal prosody, supports and modulates word recognition. Audiovisual prosody is implemented in the fuzzy logical model of perception [47], but how this information affects lexical processing is underspecified. Moreover, experiment 3 implicates that the multimodal suprasegmental and segmental 'analysers' would have to allow for interaction, since one can influence the other. Therefore, existing models of audiovisual speech perception and word recognition thus need to be expanded to account for (i) how suprasegmental prosodic information is processed [48]; (ii) how suprasegmental information interacts with and modulates segmental information [49]; and (iii) how multimodal cues influence auditory word recognition [13,33].

A recent, broader theoretical account of human communication (i.e. beyond speech alone) proposed that two domain-general mechanisms form the core of our ability to understand language: multimodal low-level integration and top-down prediction [2]. The present results are indeed in line with the idea that *multimodal low-level integration* is a general cognitive principle supporting language understanding. More precisely, we here show that this mechanism operates across different yet tightly coupled communicative modalities. The cognitive and perceptual apparatus that we have at our disposal apparently not just binds auditory and visual information that derive from the same communicative channel (i.e. articulation), but also combines visual information derived from the hands with concurrent auditory information—crucially to such an extent that what we hear is modulated by the timing of the beat gestures we see. As such, what we perceive is the model of reality that our brains provide us by binding visual and auditory communicative input, and not reality itself.

In addition to multimodal low-level integration, earlier research has indicated that also *top-down predictions* play a pivotal role in the perception of co-speech beat gestures. It seems reasonable to assume that listeners throughout life build up the implicit knowledge that beat gestures are commonly tightly paired with acoustically prominent verbal information. As the initiation of a beat gesture typically precedes acoustic prominence cues, these statistical regularities may indeed be used in a top-down manner to allocate additional attentional resources to an upcoming word [26].

The current first demonstration of the influence of simple 'flicks of the hands' on low-level speech perception should be considered an initial step in our understanding of how bodily gestures, in general, may affect speech perception. In fact, we explicitly do not claim that the manual McGurk effect observed presently is limited to beat gestures only; instead, we would argue that it is driven by the specific temporal alignment of visual bodily cues to the speech signal, which may also apply to other types of non-manual gestures, such as head nods and eyebrow movements—as long as they are communicatively relevant (cf. the null effect in [36]). Moreover, other behavioural consequences beyond lexical stress perception could likely be implicated, ranging from phase-resetting neural oscillations [26] with consequences for speech intelligibility and phoneme identification [50,51], to facilitating selective attention in 'cocktail party' listening, word segmentation and word learning [52–55].

Finally, the current findings have several practical implications. Speech recognition systems may be improved when including multimodal signals, even beyond articulatory cues on the face. Presumably beat gestures are particularly valuable multimodal cues, because, while cross-language differences exist in the use and weighting of acoustic suprasegmental prosody (e.g. amplitude, F0, duration [56]), beat gestures could presumably function as a language-universal prosodic cue. Also, the field of human–computer interaction could benefit from the finding that beat gestures support spoken communication, for instance by enriching virtual agents and avatars with human-like gesture-to-speech alignment. The use of such agents in immersive virtual reality environments will in turn allow for taking the study of language comprehension into naturalistic, visually rich and dynamic settings while retaining the required degree of experimental control [57]. By acknowledging that spoken language comprehension involves the integration of auditory sounds, visual articulatory cues and even the simplest of manual gestures, it will be possible to significantly further our understanding of the human capacity to communicate.

Ethics. All participants gave informed consent as approved by the Ethics Committee of the Social Sciences department of Radboud University (project code: ECSW2014-1003-196). All research was performed in accordance with relevant guidelines and regulations.

Data accessibility. Example stimuli and all experimental data of this study are publicly available for download from https://osf.io/b7kue/ under a CC-By Attribution 4.0 International licence. The protocol used to create the audiovisual stimuli is available from: https://doi.org/10.17504/protocols.io.bmw5k7g6. A preprint of this paper was published on *bioRxiv* at https://doi.org/10.1101/2020.07.13.200543 [58].

Authors' contributions. Both authors conceived of the study, participated in the design of the study, supervised the research assistants in charge of recruiting and testing the participants and drafted and revised the manuscript. H.R.B. created the stimuli, built the experiments and ran the statistical analyses. Both authors gave final approval for publication and agree to be held accountable for the work performed therein.

Competing interests. We declare we have no competing interests.

Acknowledgements. This research was supported by the Max Planck Society for the Advancement of Science, Munich, Germany (H.R.B.; D.P.) and the Netherlands Organisation for Scientific Research (D.P.; Veni grant no. 275-89-037). We would like to thank Giulio Severijnen for help in creating the pseudowords of Experiment 1A, Birgit Knudsen for annotating the video recordings, and Nora Kennis, Esther de Kerf and Myriam Weiss for their help in testing participants and annotating the shadowing recordings.

References

1. Levinson SC. 1983 *Pragmatics*. Cambridge, UK: Cambridge University Press.
2. Holler J, Levinson SC. 2019 Multimodal language processing in human communication. *Trends Cogn. Sci.* **23**, 639–652. (doi:10.1016/j.tics.2019.05.006)
3. Vigliocco G, Perniss P, Vinson D. 2014 Language as a multimodal phenomenon: implications for language learning, processing and evolution. *Phil. Trans. R. Soc. B* **369**, 20130292. (doi:10.1098/rstb.2013.0292)
4. Levelt WJM. 1989 *Speaking: from intention to articulation*. Cambridge, MA: MIT Press.
5. Hagoort P. 2019 The neurobiology of language beyond single-word processing. *Science* **366**, 55–58. (doi:10.1126/science.aax0289)
6. Kaufeld G, Ravenschlag A, Meyer AS, Martin AE, Bosker HR. 2020 Knowledge-based and signal-based cues are weighted flexibly during spoken language comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* **46**, 549–562. (doi:10.1037/xlm0000744)
7. Kuperberg GR, Jaeger TF. 2016 What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* **31**, 32–59. (doi:10.1080/23273798.2015.1102299)
8. Clark HH. 1996 *Using language*. Cambridge, UK: Cambridge University Press.
9. Hagoort P, Hald L, Bastiaansen M, Petersson KM. 2004 Integration of word meaning and world knowledge in language comprehension. *Science* **304**, 438–441. (doi:10.1126/science.1095455)
10. McGurk H, MacDonald J. 1976 Hearing lips and seeing voices. *Nature* **264**, 746–748. (doi:10.1038/264746a0)
11. Alsius A, Paré M, Munhall KG. 2018 Forty years after hearing lips and seeing voices: the McGurk effect revisited. *Multisensory Res.* **31**, 111–144. (doi:10.1163/22134808-00002565)
12. Rosenblum LD. 2019 Audiovisual speech perception and the McGurk effect. In *Oxford research encyclopedia of linguistics*. See https://oxfordre.com/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-420.
13. Bosker HR, Peeters D, Holler J. 2020 How visual cues to speech rate influence speech perception. *Q. J. Exp. Psychol.* **73**, 1523–1536. (doi:10.1177/1747021820914564)
14. Rosenblum LD, Dias JW, Dorsi J. 2017 The supramodal brain: implications for auditory perception. *J. Cogn. Psychol.* **29**, 65–87. (doi:10.1080/20445911.2016.1181691)
15. McNeill D. 1992 *Hand and mind: what gestures reveal about thought*. Chicago, IL: University of Chicago Press.
16. Peeters D, Chu M, Holler J, Hagoort P, Özyürek A. 2015 Electrophysiological and kinematic correlates of communicative intent in the planning and production of pointing gestures and speech. *J. Cogn. Neurosci.* **27**, 2352–2368. (doi:10.1162/jocn_a_00865)
17. Kelly SD, Özyürek A, Maris E. 2010 Two sides of the same coin: speech and gesture mutually interact to enhance comprehension. *Psychol. Sci.* **21**, 260–267. (doi:10.1177/0956797609357327)
18. Özyürek A, Willems RM, Kita S, Hagoort P. 2007 On-line integration of semantic information from speech and gesture: insights from event-related brain potentials. *J. Cogn. Neurosci.* **19**, 605–616. (doi:10.1162/jocn.2007.19.4.605)
19. Drijvers L, Özyürek A, Jensen O. 2018 Hearing and seeing meaning in noise: alpha, beta, and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Hum. Brain Mapp.* **39**, 2075–2087. (doi:10.1002/hbm.23987)
20. Krahmer E, Swerts M. 2007 The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. *J. Mem. Lang.* **57**, 396–414. (doi:10.1016/j.jml.2007.06.005)
21. Leonard T, Cummins F. 2011 The temporal relation between beat gestures and speech. *Lang. Cogn. Process.* **26**, 1457–1471. (doi:10.1080/01690965.2010.500218)
22. Pouw W, Harrison SJ, Dixon JA. 2020 Gesture–speech physics: the biomechanical basis for the emergence of gesture–speech synchrony. *J. Exp. Psychol. Gen.* **149**, 391–404. (doi:10.1037/xge0000646)
23. Biau E, Soto-Faraco S. 2013 Beat gestures modulate auditory integration in speech perception. *Brain Lang.* **124**, 143–152. (doi:10.1016/j.bandl.2012.10.008)
24. Özyürek A. 2014 Hearing and seeing meaning in speech and gesture: insights from brain and behaviour. *Phil. Trans. R. Soc. B* **369**, 20130296. (doi:10.1098/rstb.2013.0296)
25. Pouw W, Paxton A, Harrison SJ, Dixon JA. 2020 Acoustic information about upper limb movement in voicing. *Proc. Natl Acad. Sci. USA* **117**, 11 364–11 367. (doi:10.1073/pnas.2004163117) [cited 2020 May 12]
26. Biau E, Torralba M, Fuentemilla L, de Diego Balaguer R, Soto-Faraco S. 2015 Speaker's hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. *Cortex* **68**, 76–85. (doi:10.1016/j.cortex.2014.11.018)
27. Hubbard AL, Wilson SM, Callan DE, Dapretto M. 2009 Giving speech a hand: gesture modulates activity in auditory cortex during speech perception. *Hum. Brain Mapp.* **30**, 1028–1037. (doi:10.1002/hbm.20565)
28. So WC, Chen-Hui CS, Wei-Shan JL. 2012 Mnemonic effect of iconic gesture and beat gesture in adults and children: is meaning in gesture important for memory recall? *Lang. Cogn. Process.* **27**, 665–681. (doi:10.1080/01690965.2011.573220)
29. Kendon A. 2004 *Gesture: visible action as utterance*. Cambridge, UK: Cambridge University Press.
30. Kleinschmidt DF, Jaeger TF. 2015 Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol. Rev.* **122**, 148. (doi:10.1037/a0038695)
31. Jesse A, Poellmann K, Kong Y. 2017 English listeners use suprasegmental cues to lexical stress early during spoken-word recognition. *J. Speech Lang. Hear Res.* **60**, 190–198. (doi:10.1044/2016_JSLHR-H-15-0340)
32. Reinisch E, Jesse A, McQueen JM. 2010 Early use of phonetic information in spoken word recognition: lexical stress drives eye movements immediately. *Q. J. Exp. Psychol.* **63**, 772–783. (doi:10.1080/17470210903104412)
33. Jesse A, McQueen JM. 2014 Suprasegmental lexical stress cues in visual speech can guide spoken-word recognition. *Q. J. Exp. Psychol.* **67**, 793–808. (doi:10.1080/17470218.2013.834371)
34. Sumby WH, Pollack I. 1954 Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* **26**, 212–215. (doi:10.1121/1.1907309)
35. Golumbic EMZ, Cogan GB, Schroeder CE, Poeppel D. 2013 Visual input enhances selective speech envelope tracking in auditory cortex at a 'cocktail party'. *J. Neurosci.* **33**, 1417–1426. (doi:10.1523/JNEUROSCI.3675-12.2013)
36. Jesse A, Mitterer H. 2011 Pointing gestures do not influence the perception of lexical stress. See https://www.isca-speech.org/archive/interspeech_2011/i11_2445.html.
37. Norris D, McQueen JM. 2008 Shortlist B: a Bayesian model of continuous speech recognition. *Psychol. Rev.* **115**, 357–395. (doi:10.1037/0033-295X.115.2.357)
38. Peelle JE, Sommers MS. 2015 Prediction and constraint in audiovisual speech perception. *Cortex* **68**, 169–181. (doi:10.1016/j.cortex.2015.03.006)
39. McClelland JL, Elman JL. 1986 The TRACE model of speech perception. *Cognit. Psychol.* **18**, 1–86. (doi:10.1016/0010-0285(86)90015-0)
40. Nooteboom SG, Doodeman GJN. 1980 Production and perception of vowel length in spoken sentences. *J. Acoust. Soc. Am.* **67**, 276–287. (doi:10.1121/1.383737)
41. Steffman J. 2019 Intonational structure mediates speech rate normalization in the perception of segmental categories. *J. Phon.* **74**, 114–129. (doi:10.1016/j.wocn.2019.03.002)
42. Bosker HR. 2017 Accounting for rate-dependent category boundary shifts in speech perception. *Atten. Percept. Psychophys.* **79**, 333–343. (doi:10.3758/s13414-016-1206-4)
43. Rezlescu C, Danaila I, Miron A, Amariei C. 2020 More time for science: using testable to create and share behavioral experiments faster, recruit better participants, and engage students in hands-on research. *Prog. Brain Res.* **253**, 243–262.
44. Pluymaekers M, Ernestus M, Baayen RH. 2005 Lexical frequency and acoustic reduction in spoken

Dutch. *J. Acoust. Soc. Am.* **118**, 2561–2569. (doi:10.1121/1.2011150)

45. Stoet G. 2017 PsyToolkit: a novel web-based method for running online questionnaires and reaction-time experiments. *Teach. Psychol.* **44**, 24–31. (doi:10.1177/0098628316677643)
46. Quené H, Van den Bergh H. 2008 Examples of mixed-effects modeling with crossed random effects and with binomial data. *J. Mem. Lang.* **59**, 413–425. (doi:10.1016/j.jml.2008.02.002)
47. Massaro DW. 1998 *Perceiving talking faces: from speech perception to a behavioral principle*, 524 p. MIT Press.
48. Cho T, McQueen JM, Cox EA. 2007 Prosodically driven phonetic detail in speech processing: the case of domain-initial strengthening in English. *J. Phon.* **35**, 210–243. (doi:10.1016/j.wocn.2006.03.003)
49. Maslowski M, Meyer AS, Bosker HR. 2019 Listeners normalize speech for contextual speech rate even without an explicit recognition task. *J. Acoust. Soc. Am.* **146**, 179–188. (doi:10.1121/1.5116004)
50. Kösem A, Bosker HR, Takashima A, Jensen O, Meyer A, Hagoort P. 2018 Neural entrainment determines the words we hear. *Curr. Biol.* **28**, 2867–2875. (doi:10.1016/j.cub.2018.07.023)
51. Doelling KB, Arnal LH, Ghitza O, Poeppel D. 2014 Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage* **85**, 761–768. (doi:10.1016/j.neuroimage.2013.06.035)
52. Bosker HR, Sjerps MJ, Reinisch E. 2020 Spectral contrast effects are modulated by selective attention in 'cocktail party' settings. *Atten. Percept. Psychophys.* **82**, 1318–1332. (doi:10.3758/s13414-019-01824-2)
53. Bosker HR, Sjerps MJ, Reinisch E. In press. Temporal contrast effects in human speech perception are immune to selective attention. *Sci. Rep.* **10**. (doi:10.1038/s41598-020-62613-8)
54. Jesse A, Johnson EK. 2012 Prosodic temporal alignment of co-speech gestures to speech facilitates referent resolution. *J. Exp. Psychol. Hum. Percept. Perform.* **38**, 1567–1581. (doi:10.1037/a0027921)
55. Golumbic EMZ *et al.* 2013 Mechanisms underlying selective neuronal tracking of attended speech at a 'cocktail party'. *Neuron* **77**, 980–991. (doi:10.1016/j.neuron.2012.12.037)
56. Tyler MD, Cutler A. 2009 Cross-language differences in cue use for speech segmentation. *J. Acoust. Soc. Am.* **126**, 367–376. (doi:10.1121/1.3129127)
57. Peeters D. 2019 Virtual reality: a game-changing method for the language sciences. *Psychon. Bull. Rev.* **26**, 894–900. (doi:10.3758/s13423-019-01571-3)
58. Bosker HR, Peeters D. 2020 Beat gestures influence which speech sounds you hear. *bioRxiv*. (doi:10.1101/2020.07.13.200543)