Beat gestures can influence on-line spoken word recognition

Ronny Bujok [1,2], Antje S. Meyer [1,3] and Hans Rutger Bosker [1,3]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, NL

[2]International Max Planck Research School for Language Sciences, MPI for Psycholinguistics, Max Planck Society, Nijmegen, NL

[3]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, NL

Ronny Bujok: https://orcid.org/0000-0001-6557-9808

Antje S. Meyer: https://orcid.org/0000-0002-7735-9025

Hans Rutger Bosker: https://orcid.org/0000-0002-2628-7738

Corresponding Author: Ronny Bujok (Ronny.Bujok@mpi.nl)

Abstract

Beat gestures are common co-speech gestures, and closely coupled with lexical stress. Hence, seeing a beat gesture on an initial syllable can bias people to report hearing a word with initial stress. This study tested whether these effects reflect task-effects or genuine word recognition processes, using the visual world paradigm in eye-tracking. Participants heard auditory stimuli of disyllabic Dutch stress pairs (e.g., *VOORnaam* vs. *voorNAAM*) with clear or ambiguous stress cues, while seeing videos of a talker in the center of the screen producing a beat gesture on the first or second syllable (or no gesture). Results showed that on stress-ambiguous trials, beat gestures guided word recognition in an on-line manner, similar to acoustic stress cues. For instance, a beat gesture on the first syllable biased looks towards *VOORnaam* before word offset. These results suggest that beat gestures – despite their limited inherent meaning – can affect on-line lexical access.

Speech is made up of segmental information, like speech sounds such as vowels and consonants, but also includes prosody in the form of suprasegmental information. For example, changes in pitch can indicate a question rather than a statement (for review see Xie et al., 2021). But even on the word-level, prosody can be a strong factor in speech perception. For example, lexical stress can be very useful for spoken word recognition in free-stress languages like Dutch, which is studied here, because it can be lexically contrastive. That means that the placement of lexical stress can be the only difference between two segmentally identical words (e.g., Dutch *VOORnaam* [first name] vs. *voorNAAM* [respectable]; /vo:r.na:m/; capitals indicate stress). As such considering stress in speech comprehension can reduce lexical competition and facilitate word recognition and disambiguation (for review see Cutler & Jesse, 2021). In fact, recent studies on Dutch (Reinisch et al., 2010) and English (Jesse et al., 2017) have found that acoustic stress cues are used in real-time, thus facilitating fast and efficient speech perception. This study tests whether visual cues to stress, namely co-speech beat gestures, are also used in an on-line fashion in spoken word recognition.

Specifically, using a visual world paradigm (VWP) with eye-tracking, Reinisch and colleagues (2010) tested 24 participants and measured their eye movements while they heard audio recordings of Dutch words (e.g., *OC-to-pus* vs. *ok-TO-ber)* that were segmentally identical in the first two syllables (e.g., /ɔkto/) and only differed in the placement of lexical stress. Written response options were presented in each corner of the screen and participants were asked to click on the word they heard as fast as possible. The researchers found that participants were able to look at the correct word well before it was segmentally disambiguated. That is, participants already fixated the target (e.g., “OCtopus”) immediately after hearing the stressed syllable (e.g., “OC”) and before hearing the segmentally disambiguating final syllable (e.g., “pus”). The authors concluded that listeners used the auditory lexical stress cues in real-time as soon as they became available (well before word offset and before the segmental point of disambiguation) to resolve lexical competition and choose the correct word. This finding has since been replicated in studies of different free-stress languages such

as English (Connell et al., 2018) and Italian (Sulpizio & McQueen, 2012). This emphasizes the important role stress plays in lexical access, supporting efficient and fast word recognition.

However, the aforementioned studies on the time-course of lexical stress perception (Jesse et al., 2017; Reinisch et al., 2010) focused solely on auditory perception, while speech perception often also involves the visual modality (Holler & Levinson, 2019; Perniss, 2018; Rosenblum, 2008). In the auditory modality, stress is usually expressed with modulations of acoustic cues such as fundamental frequency (F0), duration and intensity (Rietveld & Heuven, 2009; Severijnen et al., 2024). In the visual modality, stress can be visible on the face (Jesse & McQueen, 2014; Scarborough et al., 2009) and in the hands, reflected by the timing of gestures with stressed syllables (Krahmer & Swerts, 2007; Leonard & Cummins, 2011). That begs the question whether these visual cues can also affect on-line spoken word recognition immediately just like auditory cues. A recent study found no effect of visual articulatory stress cues on the face in *audiovisual* stress perception (Bujok et al., 2024): An audio with a video of a talker saying a word with initial stress (strong-weak; SW; e.g., *VOORnaam*) was perceived similarly to the same audio paired with a video of a talker saying a word with final stress (weak-strong; WS; e.g., *voorNAAM*). This is presumably because of the low salience and reliability of visual articulatory cues to stress. In contrast, manual beat gestures have been consistently found to affect stress perception (Bosker & Peeters, 2021; Bujok et al., 2024).

Beat gestures are simple, rhythmic up-and-down gestures of the hand, which are considered the most commonly used co-speech gestures (McNeill, 1992). Despite their redundancy (i.e., people can in principle communicate successfully without gesturing), they are believed to be inherently linked to speech production and planned together with speech (Kita & Özyürek, 2003). Temporal coupling of beat gestures and speech has even been found in young children (Florit-Pons et al., 2023). Specifically, their point of maximum extension, the so-called apex, is tightly linked to acoustically prominent parts of speech (Krahmer & Swerts, 2007), especially pitch accents (Leonard & Cummins, 2011). In languages with lexical stress, the apex of a beat gesture is most closely aligned to

the F0 peak of stressed syllables (Leonard & Cummins, 2011; Shattuck-Hufnagel & Ren, 2018; Yasinnika et al., 2004).

Beat gestures have been found to facilitate word recall (Kushch & Prieto, 2016) and integration of the words they were aligned to (e.g., Dimitrova et al., 2016). Moreover, their close temporal connection with stress can be used by listeners in their lexical stress perception. Recent studies have used two-alternative forced choice (2AFC) tasks to test the influence of beat gestures (Bosker & Peeters, 2021; Bujok et al., 2024). Participants were presented with Dutch lexical stress continua (e.g., /vo:r.na:m/; acoustically ranging from *VOORnaam* to *voorNAAM*) together with a beat gesture aligned to either the first or the second syllable. When asked to categorize the words as either having stress on the first (e.g., *VOORnaam*) or second syllable (e.g., *voorNAAM*), participants' categorization responses were affected by the alignment of the beat gesture they saw. That is, when they saw a beat gesture aligned to the first syllable, they were more likely to indicate perceiving stress on the first syllable, and when they saw a beat gesture aligned to the second syllable they were more likely to perceive stress on the second syllable. This effect of the talker's gestural timing affected perception of the entire stress continuum but was largest at the acoustically ambiguous steps. Yet, these studies cannot reveal when in time beat gestures affect perception. For example, these effects could reflect relatively late task-related decision effects. Alternatively, it is possible that beat gestures aid lexical access in an on-line manner in the same way auditory cues of lexical stress do (Jesse et al., 2017; Reinisch et al., 2010).

This study aimed to assess the time-course of lexical access in audiovisual speech with beat gestures. Specifically, our goal was to determine how and when beat gestures facilitate audiovisual lexical stress perception. To test this, we used eye-tracking with the visual world paradigm (VWP). The VWP is very well suited for testing on-line perception because the timing of the eye gaze is closely linked in time to relevant cues in the speech signal (for review see Huettig et al., 2011). Generally, auditory stimuli (e.g., speech) are presented while four response options (e.g., pictures or words) are presented in each corner of the screen. The dynamic changes in fixations to the response

options are closely linked to the processing of the stimulus and are thus taken as an on-line measure of stimulus processing.

Most studies have used the VWP with auditory stimuli (e.g., Jesse et al., 2017; Reinisch et al., 2010). We too adopted the VWP for the present study with a similar design as Reinisch et al. (2010), but this time using audiovisual stimuli. The audiovisual stimuli were presented in the center and four written response options in each corner of the screen. Participants were asked to click on the word they thought the speaker said as fast as possible, while we measured their eye movements. Using audiovisual stimuli could increase visual attention and fixations to the video stimuli, which could make it more difficult to find small differences in the fixations to target vs. competitor response options. However, one study testing spoken word recognition using audiovisual stimuli found that the visual world paradigm works best when cognitive load is low by means of a constant display of response options and increased attention to the speaker (by using a gesture) (Mitterer & Reinisch, 2017). Therefore, this finding makes us confident that the VWP with audiovisual stimuli can be applied to test our research question.

Considering the potential influence of the presence of video stimuli on screen on the proportion of looks to the response options, we first wanted to replicate earlier findings on the time-course of the processing of *auditory stress cues* on lexical stress perception (Jesse et al., 2017; Reinisch et al., 2010) with audiovisual stimuli (i.e., a talking person, but critically without any gestures). Because of the considerations about video stimuli, we decided to use lexical stress minimal pairs (e.g., *VOORnaam – voorNAAM;* /vo:r.na:m/), where lexical stress is critical for disambiguation (i.e., unlike *OCtopus* vs. *okTOber*; see Reinisch et al., 2010). This also facilitates comparison with previous studies that used the same pairs (Bosker & Peeters, 2021; Bujok et al., 2024). We expected to find more fixations to the target (e.g., *VOORnaam*) than competitor (e.g., *voorNAAM)* during ongoing auditory presentation of the word, and hardly any fixations to any segmentally distinct distractors (e.g., *CAnon* or *kaNON*). At the same time the audiovisual trials without gestural information serve as a control condition and can be directly compared to trials with beat gestures.

Finally, we combined both members of a stress pair with the talker's face articulating the word with initial stress and also with the talker articulating the word with final stress. However, given the lack of evidence for articulatory cues to influence audiovisual stress perception in prior work (Bujok et al., 2024), we predicted to find no evidence for effects of visual articulatory stress cues on participants' fixations.

The primary goal of this study was to assess the time-course of the uptake of beat gestures in audiovisual lexical stress perception. Therefore, in addition to the aforementioned trials without gestures, we also presented participants with videos of the talker producing unambiguously stressed words (e.g., original recordings of *VOORnaam* or *voorNAAM*), or ambiguously stressed words (e.g., /vo:r.na:m/ midway between clear *VOORnaam* and *voorNAAM*), paired with either a beat gesture on the first (Beat on 1st; Bo1) or on the second syllable (Beat on 2nd; Bo2).

Beat gestures align most consistently with the pitch accent in the speech signal, but considering the observation that a beat gesture's onset and preparation can precede the onset of a stressed syllable by as much as 300 ms (Leonard & Cummins, 2011), beat gestures could offer early visual information that could be used to predict the position of lexical stress and speed up word recognition. Hence, regarding the unambiguously stressed (henceforth called "original") trials, we predicted that congruently timed beat gestures facilitate word recognition and thus lead to earlier looks to the target word when compared to no-beat trials. That is, when presented with original words with congruent beat gestures (e.g., *VOORnaam* + Bo1 or *voorNAAM* + Bo2), participants may be faster to fixate the correctly-stressed target word (vs. stress-mismatching competitor) when a beat gesture is produced on the stressed syllable (compared to a no-beat condition). We also included trials with incongruently timed beat gestures (e.g., *VOORnaam* + Bo2 or *voorNAAM* + Bo1) for a balanced design. It is possible that incongruent beat gestures could delay and reduce the looks to the target as the incongruently timed beat gesture drives some looks to the competitor.

Regarding ambiguously stressed (henceforth called "ambiguous") trials, where the audio was perceptually midway between having word initial stress (strong-weak; SW) vs. word-final stress

(weak-strong, WS), we predicted to find no difference in looks to the SW (e.g., *VOORnaam*) vs. WS member (e.g., *voorNAAM*) when the beat gesture was absent because of the auditory ambiguity. In contrast, when a beat gesture was present and thus a critical disambiguating cue for lexical stress, we predicted more fixations to the word suggested by the temporal alignment of the beat gesture. That is, we predicted more looks to the word with word-initial stress on ambiguous trials with a beat gesture on the first syllable, but more looks to the word with word-final stress on ambiguous trials with a beat gesture on the second syllable. Moreover, we expected to find this effect of preferential looking before word offset and time-locked to the apex of the beat gesture, which has been found to be the most consistently aligned part of a beat gesture with stressed syllables (Leonard & Cummins, 2011). This would indicate that the beat gesture information is used to make fast and efficient decisions about lexical stress, playing an important role in early on-line lexical access, just like auditory cues (Reinisch et al., 2010). However, if we find late preferential looking behavior (i.e., after word offset) this would mean that listeners prioritize auditory processing and, unlike auditory information, likely use gestural information relatively late in task-related decision-making processes.

Methods

Participants

Participants gave informed consent to participate in this study, which was approved by the Ethics Committee of the Social Sciences department of Radboud University (project code: ECSW-2019-019). Criteria for eligibility for this study were normal or corrected-to-normal vision, no reported hearing or language deficit, and having Dutch as their native language. Participants were financially compensated for their participation. 32 participants (27 female, 5 male), were recruited from the Max Planck Institute for Psycholinguistics participant database. Median age was 21 (SD = 3.7, range = 18 - 36).

Materials

Stimuli

Materials used in the present study were adopted from a previous study (Bujok et al., 2024). Seven disyllabic segmentally identical minimal stress pairs of Dutch, which differed only in the placement of lexical stress, were chosen (see Table 1). We made high-definition video recordings of a male native speaker of Dutch (i.e., last author) producing all of these words naturally without any gesture. Additionally, we recorded the same speaker producing all of the words with a naturally aligned beat gesture on the stressed syllable. Video was recorded at 50 frames per second and audio was sampled at 48 kHz.

Table 1. Overview of the Dutch items used in this study [English translations]. Item pairs are segmentally identical (see IPA transcription) and only differ in the placement of lexical stress (indicated by capital letters). SW (strong-weak) = stress on the first syllable; WS (weak-strong) = stress on the second syllable).

SW (strong-weak)	WS (weak-strong)	IPA transcription
CAnon [canon]	*kaNON* [cannon]	/ka.nɔn/
CONtent [content]	*conTENT* [content]	/kɔn.tɛnt/
SERvisch [Serbian]	*serVIES* [tableware]	/sɛr.vis/
VOORnaam [first name]	*voorNAAM* [respectable]	/vo:r.na:m/
VOORruit [windshield]	*voorUIT* [forward]	/vo:r.œyt/
VOORspel [prelude]	*voorSPEL* [predict]	/vo:r.spɛl/
VOORtuin [front garden]	*forTUIN* [fortune]	/vo:r.tœyn/

We extracted the audio from the videos that did not have any gesture to create a stress-ambiguous token for each pair. We set the duration and intensity of individual syllables at fixed stress-ambiguous values (midway between stressed and unstressed) and then interpolated the F0 contours of the recordings linearly using PSOLA in Praat (Boersma & Weenink, 2024) (ranging from the original SW recording to the original WS recording, see Figure 1). Based on categorization results in a pretest we selected one ambiguous step for each pair, where participants were about equally likely to choose the SW and the WS response option (steps 4 from the stress continua in Bujok et al., 2024).

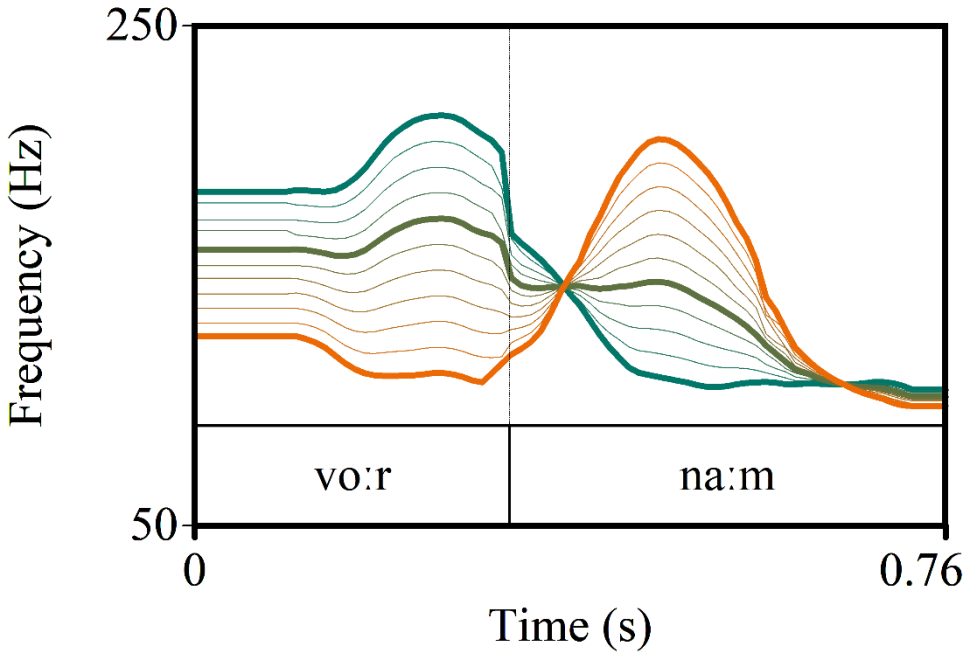
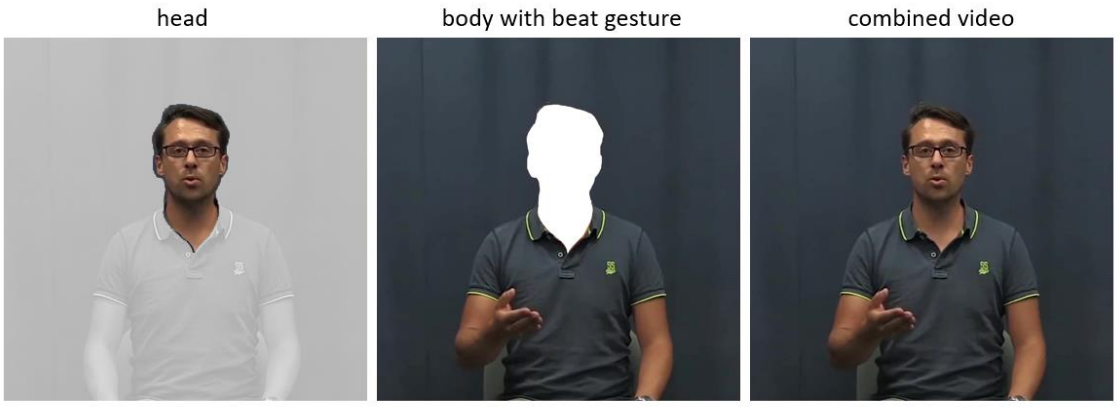


Figure 1. Visualization of the F0 stress manipulation for /vo:r.na:m/: F0 contours were interpolated in 11 steps to go from SW (green) to WS (orange). The most ambiguous step, determined by ca. 50% proportion SW responses in a pretest was selected as the ambiguous step in this study. Together with the original two recordings, it created our three Audio Conditions (SW, ambiguous, WS; highlighted in bold). Note: the original recordings (SW & WS) were unmanipulated and thus also contained clear duration and intensity cues to stress.

The original SW and WS video-recordings from the no-beat gesture videos were used as they were. The beat gesture stimuli were created by combining the face and audio from the no-beat videos with the gesturing body from the original beat recordings. That manipulation ensured that the only difference between no-beat and beat trials was the presence of a beat gesture (i.e., no difference in articulatory cues on the face, or audio). Using Adobe Premiere Pro CC 2018, we cut out the head and the neck from the videos without a beat gesture and superimposed it onto a video of the speaker producing either word of the same item (e.g., *VOORnaam* or *voorNAAM*) with a beat gesture. By moving the headless body forward/backward in time we temporally aligned the apex of

229 the beat gesture to vowel onset again. Lastly, a feathered mask seamlessly blended together both
230 videos (see Figure 2). Importantly, for this study we only used stimuli where the beat gesture
231 alignment was consistent with the visual articulatory cues on the face (e.g., Beat on 1st + lips
232 producing "*VOORnaam*"). This ensured that there was no conflicting information within the visual
233 modality, and that the only difference between beat and no-beat trials was the presence or absence
234 of a beat gesture (i.e., articulatory cues were identical in each comparison). These co-varying visual
235 stress cues (i.e., beat gesture and facial cues) could either be congruent (e.g., hearing *VOORnaam*) or
236 incongruent (e.g., hearing *voorNAAM*) with the auditory stress cues.

237



238

239 **Figure 2. Illustration of the Video Manipulation to create Beat Gesture Items:** Head from the non-gesture
240 videos (left image) was pasted onto a video of the speaker producing a beat gesture (middle image) resulting in
241 a seamlessly combined stimulus (right picture). Note: greyed- and whited-out areas only for illustration
242 purposes.

243

244 The manipulated ambiguous audio was combined with both the SW (i.e., Beat on 1st) and WS
245 video (i.e., Beat on 2nd), to create trials with ambiguous audio disambiguated by a beat gesture (and
246 co-varying articulatory cues) as either stresses on the first or on the second syllable. Since the
247 ambiguous audio was also manipulated in its duration, it created slight audiovisual asynchronies
248 (about 40 ms) between the audio and face. To minimize these asynchronies, we aligned audio and

face at the time point of the original second syllable onset. This also guaranteed that the beat gestures would always be aligned to the correct syllable and any misalignments would be contained and limited to within each syllable. This left us with 84 unique items (7 words x 3 auditory steps x 2 trial types (beat vs. no-beat) x 2 video conditions (SW-biasing vs. WS-biasing). Note that no-beat trials could be SW- or WS-biasing because of the visual articulatory cues on the face. Because of this a comparison between beat and no-beat trials allowed us to more easily draw conclusions about the beat gesture as it was the only difference between these trial types.

Design and Procedure

The experiment was run in Presentation® software (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA) and presented on a 24" full HD screen with a refresh rate of 144 Hz. Audiovisual (AV) stimuli appeared in the center of the screen as 720 x 720 pixel displays on a white background. Response options were presented in either corner of the screen (see Figure 3). Audio was presented through high quality headphones (Sennheiser HD 437) at a comfortable volume. Participants were seated at a distance of approximately 60 cm from the screen in a soundproof booth. We recorded eye movements with a desktop mounted EyeLink® 1000 Plus (EyeLink1000 Plus, SR Research, Ontario, Canada) at a sampling rate of 250 Hz. The eye-tracker was set to monitor the participant's dominant eye, which was determined before testing for each participant individually. A chin and forehead rest were used to stabilize the participant's head.

The experiment was designed as a four alternative forced choice task (4AFC). Participants were presented with all 84 unique items, but the ambiguous items (28), where we expected the greatest benefit of the beat gesture, were presented twice, resulting in 112 trials in total. All trial sequences were identical (see Figure 3). On each trial, participants were presented with a 3000 ms preview of all four response options in the corners of the screen to minimize scanning behavior of the response options in the critical window when the stimulus was playing. Then followed a fixation cross for 500 ms, followed by the video stimulus. Participants were instructed to select the word they

thought the talker produced, by clicking on one of the four response options as fast and accurately as possible. The selected word would then be highlighted in red for 1000 ms, followed by a blank screen for 500 ms before moving on to the next trial. If participants failed to respond within 4000 ms the trial would time out automatically and the next trial would start.

In order to facilitate fast eye-movements and minimize scanning behavior during stimulus presentation, we decided to take a few measures related to the response options. First, the horizontal position of the response options was counterbalanced between participants. That is, half of the participants saw all SW words on the right, and the other half of participants saw them on the left side of the screen. Second, the distractors were always segmentally different from the target and competitor in both syllables. For instance, on trials where the target and competitor started with /vo:r/ (e.g., *VOORnaam - voorNAAM*), the distractors did not start with /vo:r/ (e.g., *CAnon - kaNON*). The vertical position (i.e., whether the target pair was presented on the top half of the screen or the bottom half of the screen) was fully randomized within participants. Each word pair was used equally as often as a distractor pair (16 times).

Our labelling of the response options is dependent on the conditions. Generally, in the SW and WS original Audio Conditions (not in the ambiguous condition), the labels were assigned based on the auditory stress cues, regardless of the alignment of the beat gesture. This means that when a stimulus with unambiguous auditory stress (e.g., *VOORnaam*) was presented, the labels target (e.g., *VOORnaam*) and competitor (e.g., *voorNAAM*) were determined based on the auditory stress, even when participants were presented with incongruent beat gestures (e.g., *VOORnaam* + Beat on 2nd). However, for trials with ambiguous audio one cannot speak of a “correct” target and “incorrect” competitor. Instead, we refer to the stress pattern of the response options. That is, on a trial with ambiguous audio (e.g., /vo:r.na:m/) there would be an SW target (e.g., *VOORnaam*) and WS target (e.g., *voorNAAM*), referring to the *segmental,* orthographic target, regardless of the alignment of the beat gesture.

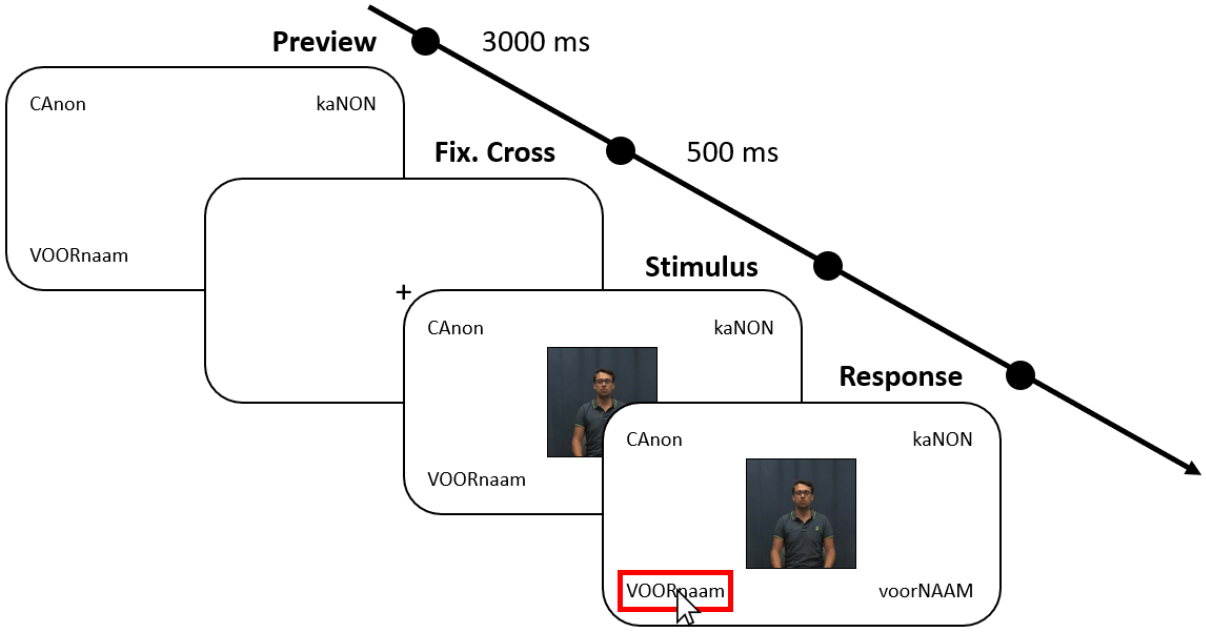


Figure 3. Trial Sequence Example: Each trial progressed in the same sequence. First participants saw all four response options for a preview of 3000 ms, followed by a fixation cross for 500 ms. Then followed the audiovisual stimulus. Participants then had to click on the word they thought the speaker said as fast as possible. They could respond while the video was still playing. Eye-movements were recorded throughout the whole trial.

Analysis of Fixations

Data were exported with EyeLink Data Viewer software package (SR Research Ltd., Version 4.3.1) and further analyzed in R (R Core Team, 2021). For the analysis of the fixation data, we defined five areas of interest (see Figure 5): The video as a 720x720 pixel square in the center of the screen, and each corner of the screen, extending from the edge of the video to the border of the screen horizontally, and from the center of the screen to the border of the screen vertically. Fixations were counted if they fell within the predefined area of interest. We assumed an average duration of 200 ms for initiating a saccade (see Matin et al., 1993) and hence all relevant time windows, related to timings in the stimuli, were shifted by +200 ms. We calculated the dependent variable target preference as the difference of logit transformed proportions of fixations to target vs. competitor. Hence, a positive difference indicates a preference to fixate the target (e.g., *VOORnaam*) over the

stress competitor (e.g., *voorNAAM*). For analysis of ambiguous trials with no clear target or competitor, we instead used SW preference as the dependent variable, calculated as the difference of logit transformed proportions of fixations to the SW target response option vs. WS target response option. Hence, a positive difference indicates a preference to fixate the SW target over the WS target.

Data were first analyzed with a series of linear mixed effects models in predefined broad time windows corresponding to syllable 1 and syllable 2. These first analyses were run to assess the presence of an effect in these relevant time windows (e.g., more fixations to target *VOORnaam* over competitor *voorNAAM*). Only if an effect was observed did we determine *at what time point* the effect could be reliably detected. For this second analysis we used divergence point analyses (DPA), which are able to determine the time point at which the emergence of an effect becomes stable and reliably detectable (Ito & Knoeferle, 2023; Stone et al., 2021). Moreover, unlike other methods (e.g., growth curve analysis, cluster-based permutation analysis, bootstrapped differences of timeseries) this method allows one to compare the onset of an effect across relevant conditions (for review see Ito & Knoeferle, 2023). The DPA determined a divergence point when participants showed a sustained preferential fixation bias for at least 200 ms.

Results

Categorization Data

We analyzed participants' behavioral mouse-click responses on all beat trials to test the effect of beat gesture alignment on lexical stress perception. We used Generalized Linear Mixed Models, with the categorization responses as the dependent variable (SW coded as 1, e.g., *VOORnaam*; WS coded as 0, e.g., *voorNAAM*). Audio Condition (categorical; SW, WS, ambiguous (on the intercept)), Beat Alignment (categorical, deviance coded: Beat on 1st as 0.5 and Beat on 2nd as -0.5) and their interaction were included as predictors. The model also included random intercepts for Participants and Items and by-participant random slopes for all predictors.

When compared to the ambiguous Audio Condition, SW audio was generally rated as more SW-like (β = 2.823, *SE* = 0.369, *z* = 7.644, *p* < 0.001), and WS audio was rated as less SW-like (β = -3.64, *SE* = 0.494, *z* = -7.372, *p* < 0.001), confirming the different stress patterns in the three levels of Audio Condition. Importantly, the effect of Beat Alignment was also significant (β = 1.405, *SE* = 0.201, *z* = 6.979, *p* < 0.001), and no interaction with either SW audio (β = 0.43, *SE* = 0.465, *z* = 0.924, *p* = 0.355) or WS audio (β = -0.316, *SE* = 0.461, *z* = -0.685, *p* = 0.494) was found. That means that participants generally gave more SW responses when they saw a beat gesture on the first syllable than when they saw a beat gesture on the second syllable (see Figure 4). Relevelled models confirmed that the Beat Alignment effect was even significant for original SW (β = 1.835, *SE* = 0.457, *z* = 4.013, *p* < 0.001) and WS audio (β = 1.09, *SE* = 0.453, *z* = 2.406, *p* = 0.016). This demonstrates a consistent influence of beat gestures on lexical stress perception, even when auditory stress cues are clear and unambiguous.

A similar analysis was run on no-beat trials, where the predictor Beat Alignment was replaced by Face (i.e., articulatory cues). This analysis confirmed that different articulatory stress cues on the face did not affect categorization responses (β = 0.212, *SE* = 0.145, *z* = 1.459, *p* = 0.145) and did not interact with either SW Audio Condition (β = -0.212, *SE* = 0.451, *z* = -0.47, *p* = 0.639) or WS Audio Condition (β = -0.068, *SE* = 0.407, *z* = -0.167, *p* = 0.868) (see Figure 4). This demonstrates that the effect of Beat Alignment on beat trials was driven primarily by beat gestures, and not the co-varying articulatory cues.

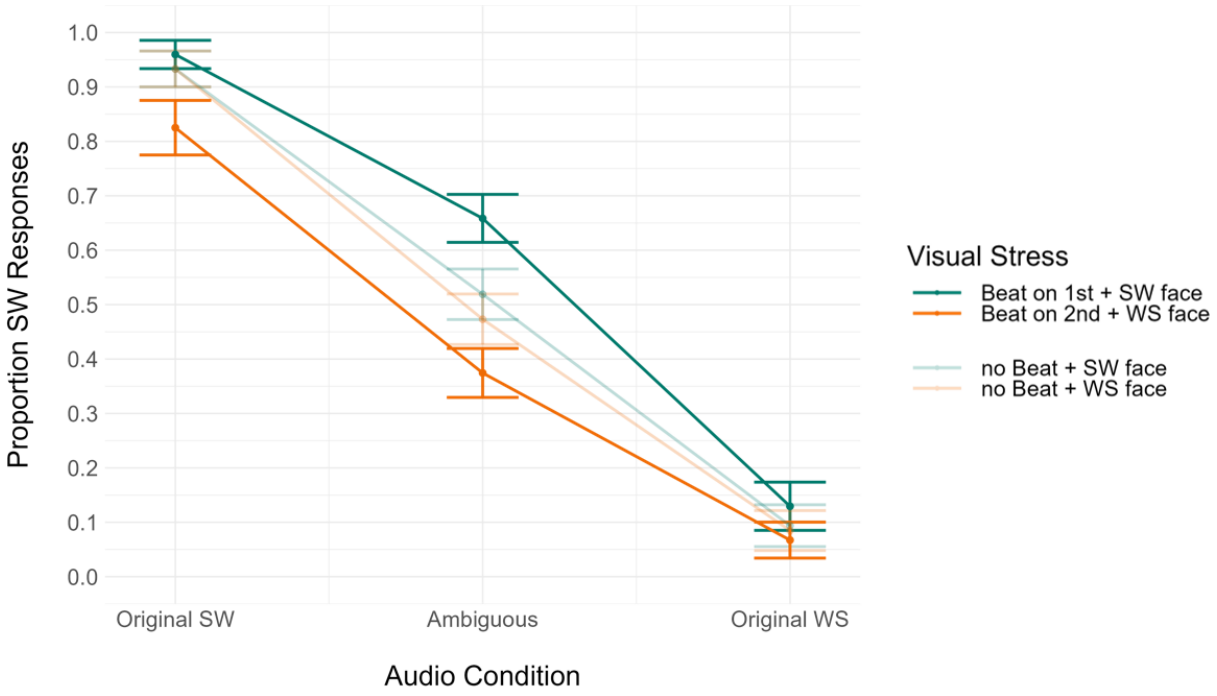


Figure 4. Categorization responses. Participants gave more SW responses when presented with a beat gesture on the first syllable when compared to seeing a beat gesture on the second syllable. This effect was present across all audio conditions. When participants were presented with videos of a talker with articulatory stress cues on the face, but without a beat gesture, categorization responses were not influenced by the visual articulatory cues (transparent lines). SW = strong-weak, stress on first syllable; WS = weak-strong, stress on second syllable.

Fixation Data

The general fixation pattern across all trials, showing the number of fixations on every part of the screen suggests that the response options and the face of the talker were fixated the most (see Figure 5). In contrast, it appears that only relatively few fixations fell on the position of the beat gesture (i.e., bottom left corner of the video).

We analyzed the fixation data separately for the following three trial types: Original Audio, Ambiguous Audio, Original Audio - Incongruent Visual Cues. All data and scripts used for these analyses are publicly available on OSF (https://osf.io/57dvh/). The analyses on the Original Audio assessed whether congruent beat gestures on the stressed syllable would facilitate word recognition

compared to the no-gesture condition as measured by preferential looking behavior towards the target (e.g., *VOORnaam*) vs. stress competitor (*voorNAAM*). We assessed this separately for the original SW and original WS words because of their intrinsic differences in when in time the multisensory stress cues arrived.

Then we analyzed the trials with Ambiguous Audio. In these analyses we tested whether the alignment of the beat gesture would affect fixation patterns in an on-line fashion. That is, we tested whether a beat gesture on the first syllable would lead to more fixations to the response option with an SW stress pattern (i.e., *VOORnaam*) than to the response option with a WS stress pattern (i.e., *voorNAAM*). For beat gestures aligned to the second syllable we would expect the opposite fixation bias (i.e., more fixations to *voorNAAM*). Because the audio was identical for trials with either beat gesture alignment, all ambiguous trials were analyzed together. Finally, we analyzed the trials with Original Audio and incongruent beat gestures, to test if incongruent beat gestures lead to delayed preferential looking to the auditory target over auditory competitor because of conflicting visual information. These trials were, like the analyses of trials with congruent beat gestures, analyzed in separate subsets for SW and WS audio.

The procedure and models for the analyses were similar for all trial types. For Original Audio (with congruent or incongruent beat), we calculated the proportion of looks based on the five interest areas (see Figure 5) within two time windows: the time window of the first and second syllable. We then logit transformed the proportion of looks (proportions 0 and 1 remapped to 0.025, 0.975 respectively) to the crucial interest areas of (auditory) target and competitor and subtracted them from one another to arrive at difference of looks to target over competitor as the dependent variable. As the predictors, we included Beat Presence (categorical predictor, dummy-coded: no-beat as 0 and beat as 1), and Time Window (categorical predictor, deviation-coded: 1st syllable as -0.5 and 2nd syllable as 0.5), and their interaction as predictors. The random effects structure contained Item and Participant random intercepts and by-Participant slopes for Beat Presence. Models with more complex random effects structures failed to converge.

For Ambiguous Audio trials, the dependent variable was the logit transformed difference of fixations to the SW target (e.g., written *VOORnaam*) over the WS target (e.g., written *voorNAAM*). In addition to the predictors Beat Presence and Time Window (see above), we also included the predictor Visual Stress (categorical predictor, deviation-coded: Visual Stress on 1st syllable as -0.5 and Visual Stress on 2nd syllable as 0.5). Note that Visual Stress refers to the articulatory cues of the face on trials without beat gesture. On trials with beat gesture, Visual Stress refers to the co-varying articulatory cues *and* beat gesture (i.e., aligned to either the first or second syllable). Moreover, we included all interactions of the predictors. This allowed us to compare the influence of beat gestures (on beat trials) with the influence of articulatory cues on the face (on no-beat trials). The random effects structure contained Item and Participant random intercepts and by-Participant slopes for all predictors.

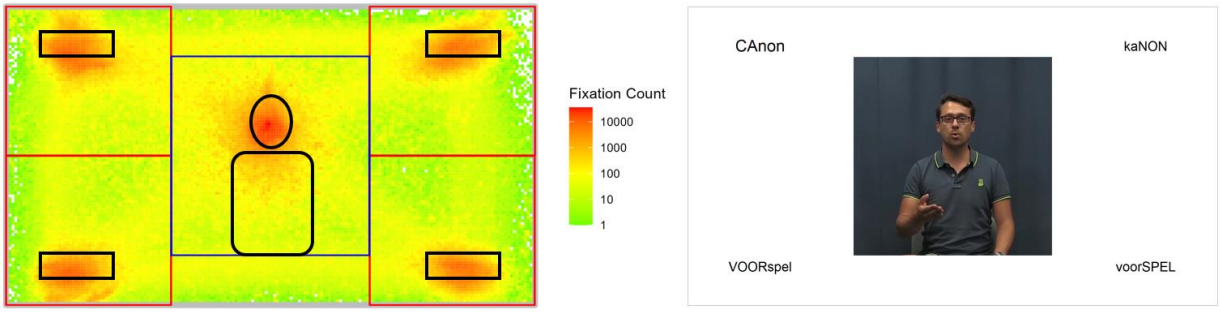


Figure 5. Heatmap of the fixation data across all participants and trials (left) and screenshot from a trial for comparison (right). Relevant areas of interest drawn in with red rectangles (around the words) and blue rectangle (around the video in the center). Position of the response options and the talker are drawn with black outlines (left panel). Most fixations within each interest area were found at the position of the words and the talkers face, but hardly any at the position of the gestures (i.e., arms, hands, chest). Note: the color scale of the heatmap is logarithmic.

Original Audio – congruent Beat vs. no Beat

In the first set of analyses, we created two subsets (one for original SW words and one for original WS words) for all trials with congruent auditory information and visual information (see

Figure 6). These subsets included no-beat trials with congruent visual stress information of the face as well as beat trials with congruent visual information of the face *and* a congruent beat gesture aligned to the stressed syllable. The only difference between the no-beat and beat trials was thus the presence or absence of a beat gesture. Generally, for both SW audio and WS audio, fixations to the video stimulus in the center decreased over time already starting from word onset (see Figure 6, light blue lines) as looks to target (green lines) and competitor (orange lines) increased. During the second syllable, the proportion of fixations to the target seems higher than those to the competitor. For trials with WS audio the fixation pattern for trials with congruent beat gestures (solid lines) and trials without beat gestures (dashed lines) appears similar. The fixation data on trials with SW audio hints at fewer fixations of the video, and a bit more fixations of the target when a beat trial rather than a no-beat trial was presented.

SW Audio. The model analyzing fixations on trials with original SW audio showed an effect of the intercept (β = 1.076, *SE* = 0.306, *t* = 3.512, *p* = 0.002), which can be interpreted as a general target preference on no-beat trials when considering the entire target word duration. An effect of Time Window showed that this target preference was larger in the Time Window of the second syllable than the first syllable (β = 1.924, *SE* = 0.318, *t* = 6.053, *p* < 0.001). Relevelled models revealed that the target preference was only present for the Time Window of the second syllable (β = 2.038, *SE* = 0.345, *t* = 5.905, *p* < 0.001) but not first syllable (β = 0.114, *SE* = 0.345, *t* = 0.33, *p* = 0.743). However, the target preference was not affected by Beat Presence (β = 0.178, *SE* = 0.296, *t* = 0.602, *p* = 0.551). Hence, participants' proportions of looks to target vs. competitor was similar on Beat and no-beat trials when hearing clear acoustic cues to an SW stress pattern (see Figure 6, left panel). Subsequent Divergence Point Analyses determining the time point of emergence of the target preference reached significance around 610 ms (95% CI = [560, 700]) for trials with a beat gesture, compared to 675 ms (95% CI = [660, 700]) on no-beat trials. Numerically, participants' divergence point occurred 65 ms earlier on trials with a beat gesture, but this difference was not significant (*p* = 0.917). In summary, we find that participants used the acoustic cues to stress to successfully fixate

456 the SW target word over the WS competitor early in time. However, we did not find reliable
457 statistical evidence that a congruent beat gesture on the first syllable sped up this preferential
458 looking behavior.

459 **WS Audio.** The results from the WS model showed a similar pattern. There was a general
460 target preference when considering the entire word duration (β = 1.073, *SE* = 0.253, *t* = 4.246, *p* <
461 0.001), but no effect of Beat Presence (β =-0.109, *SE* = 0.235, *t* = -0.466, *p* = 0.644), suggesting similar
462 target preference effects on beat and no-beat trials (see Figure 6, right panel). The effect was larger
463 in the Time Window of the second syllable than the first syllable (β = 1.977, *SE* = 0.319, *t* = 6.191, *p* <
464 0.001). In fact, relevelled models showed that the effect was only significant during the second
465 syllable (β = 2.062, *SE* = 0.3, *t* = 6.897, *p* < 0.001) and not the first syllable (β = 0.085, *SE* = 0.3, *t* =
466 0.283, *p* = 0.778). The divergence point of the target fixation preference was at 603ms (95% CI =
467 [520, 660]) for beat trials and 594ms (95% CI = [500, 700]) for no-beat trials, showing no significant
468 difference (p = 0.669). Taken together, this shows that participants' fixations were guided by acoustic
469 cues to stress on trials with original WS speech in an on-line manner as the target preference
470 emerged while the word was still being uttered. We found no reliable influence of additional
471 congruent beat gestures on fixation patterns.
472

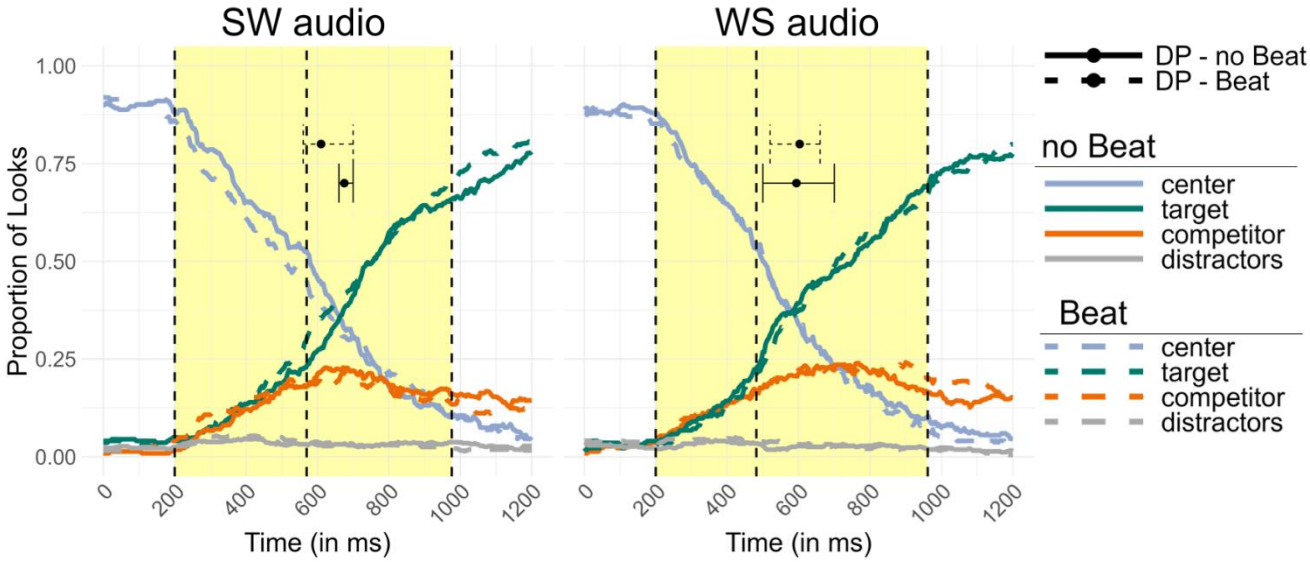


Figure 6. Comparison of fixations across time between trials with no beat and congruent beat with original audio. Fixation patterns between both conditions (no-beat vs. beat) were similar. Participants were able to fixate the target over the competitor well before word offset for SW and WS words. This effect was significant in the time window of the second syllable. Divergence points (i.e., reliably detectable preference for targets over competitors) also fell within the second syllable (SW audio: no-beat = 675 ms, Beat = 610 ms; WS audio: no-beat = 594 ms, beat = 603 ms). Note: shaded areas indicate relevant time windows of analysis. Vertical dashed lines correspond to word onset, second syllable onset and word offset. All timings were shifted by 200 ms. Both distractors were averaged and plotted together. DP = Divergence Point. Error bars show 95% confidence intervals.

The analyses on original audio demonstrated participants' ability to fixate the target over the competitor well before word offset, replicating a previous audio-only study (Reinisch et al., 2010). However, on these original trials, there were clear auditory cues to stress as well as clear visual articulatory cues to stress, limiting the contribution of the arguably redundant beat gestures. Hence in the following analysis we focused on trials with ambiguous audio to isolate the influence of the beat gesture, when audio was not informative regarding lexical stress (see Figure 7).

Moreover, in the analysis of the ambiguous trials we could also isolate the effect of beat gestures from any effects of articulatory cues on the face. That is, on ambiguous trials without beat gesture, the video stimuli still differed in terms of Visual Stress (e.g., lips producing SW word or WS

492 word). On ambiguous trials with beat gesture, Visual Stress was indicated by beat gesture alignment
493 (i.e., Beat on 1st and Beat on 2nd) *and* (congruent) articulatory cues. Hence the only difference
494 between trials with beat gesture and no beat gesture, was the presence or absence of a Beat
495 gesture.
496
497 ***Ambiguous audio – Beat gesture alignment***
498 The analysis of ambiguous audio trials revealed no effect of Visual Stress on no-beat trials (β
499 = 0.008, *SE* = 0.223, *t* = 0.035, *p* = 0.973), meaning that articulatory cues alone did not affect fixation
500 patterns. Hence for clarity the top left panel Figure 7 shows combined SW and WS articulatory cues
501 (see Figure S1 in Supplementary Materials for extensive plot: https://osf.io/57dvh/). However, there
502 was an interaction of Visual Stress with Beat Presence (β = -0.718, *SE* = 0.226, *t* = -3.177, *p* = 0.002),
503 and a three-way interaction of Visual Stress, Beat Presence and Time Window (β = -0.97, *SE* = 0.452, *t*
504 = -2.147, *p* = 0.032). A relevelled model with the beat trials on the intercept confirmed that when a
505 beat gesture was present, Visual Stress (articulatory cues *and* Beat Alignment) affected fixations
506 significantly (β = -0.654, *SE* = 0.162, *t* = -4.043, *p* < 0.001). This means, that the alignment of the beat
507 gesture affected the SW preference (see Figure 7).
508 In this relevelled model, the effect of Visual Stress interacted with Time Window (β = -1.242,
509 *SE* = 0.32, *t* = -3.887, *p* < 0.001). That is, beat gestures did not affect fixations in the Time Window of
510 the first syllable (β = -0.033, *SE* = 0.229, *t* = -0.143, *p* = 0.887), but only during the second syllable (β =
511 -1.274, *SE* = 0.229, *t* = -5.588, *p* < 0.001). That is, in this second time window the preference to look at
512 the SW target was smaller when the beat gesture was aligned to the second syllable than when it
513 was aligned to the first syllable. In fact, relevelled models showed that when beat was aligned to the
514 first syllable, participants showed a significant fixation preference of SW target over WS target in the
515 Time Window of the second syllable (β = 0.921, *SE* = 0.249, *t* = 3.695, *p* < 0.001) (see bottom left
516 panel in Figure 7). When beat gesture was aligned to the second syllable participants showed a
517 fixation preference of WS target over SW (i.e., reversed pattern). Note that because of the later

timing of the beat gesture, this last effect only reached significance in an additional post-hoc model which tested the effect in the Time Window of the second syllable + 200 ms (β = -0.532, *SE* = 0.237, *t* = -2.242, *p* = 0.028) (see bottom right panel in Figure 7).

According to a DPA, the divergence point of the fixation preference on trials with Beat on 1st became significant at 727 ms (95% CI = [680, 780]), which is 402 ms after the beat apex (325 ms). The DPA on trials with Beat on 2nd calculated the divergence point at 988 ms (95% CI = [960, 1000]), which is only 14 ms after average word offset at 974ms, and 380 ms after the apex of the beat gesture at 608 ms. In sum, beat gesture alignment (but not articulatory stress cues) affected fixation preference of the response options. Specifically, when beat was aligned with the first syllable, participants fixated the SW target more than the WS target while the word was still being uttered. When Beat was aligned with the second syllable, participants looked at the WS target more than the SW target and the preference became reliably detectable around word offset.

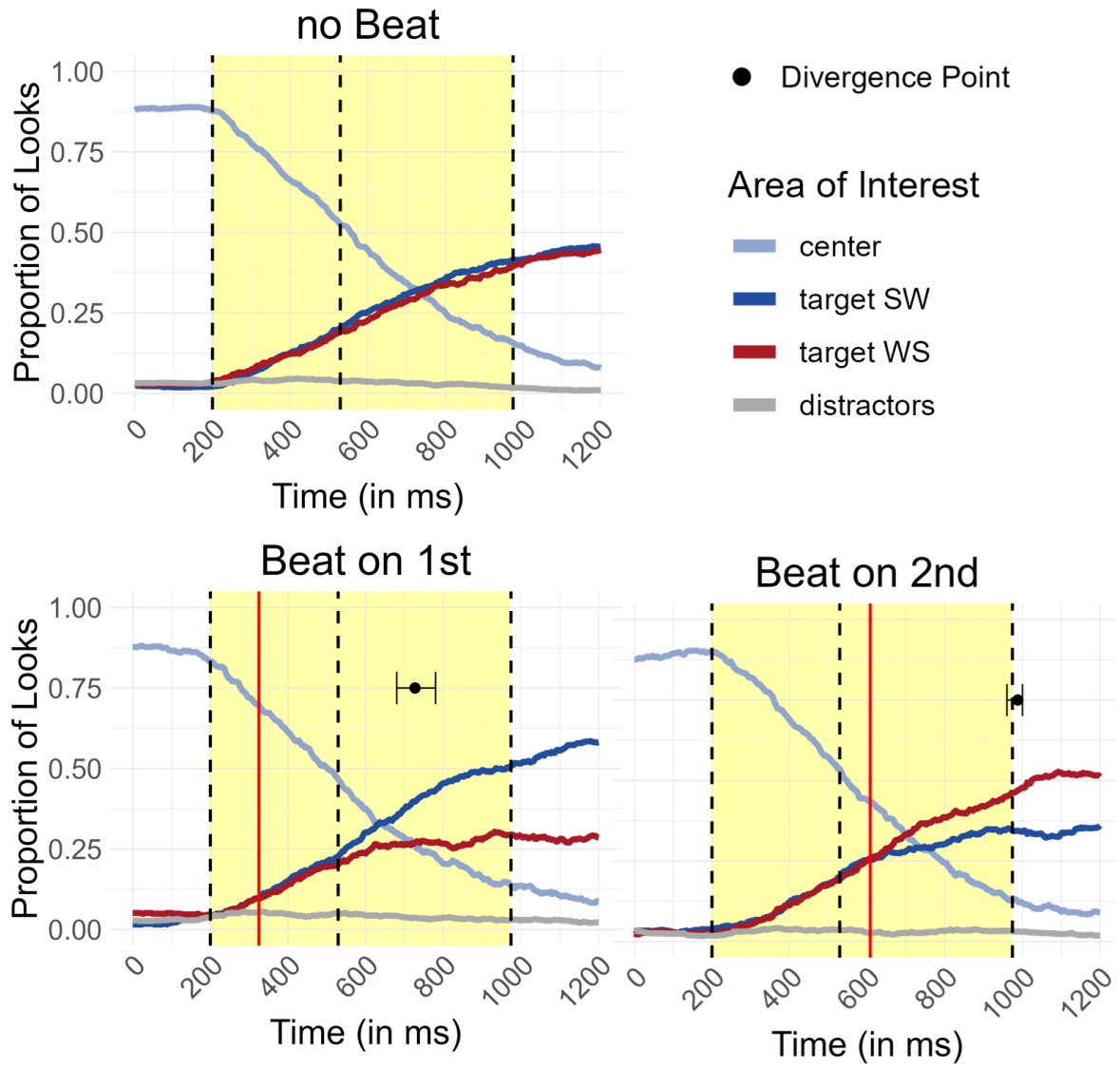


Figure 7. Fixations across time on trials with ambiguous audio. When presented with ambiguous audio without a beat gesture, participants did not preferentially look to either the SW or WS target word (Note: top left panel shows a combined plot with SW and WS articulatory cues for simplicity). When an early beat gesture was present (Beat on 1st syllable), participants looked significantly more to the SW target than the WS target in the time window of the second syllable (divergence point at 727 ms). Late beat gestures (Beat on 2nd syllable) made participants look more to the WS target than the SW target. However, the effect was only reliably detectable 14 ms after word offset (divergence point at 988 ms). Note: shaded areas indicate relevant time windows of analysis. Vertical dashed lines correspond to word onset, second syllable onset and word offset. Vertical red line indicates average point of maximum extension of the beat gesture. All timings were shifted by 200 ms. Both distractors were averaged and plotted together.

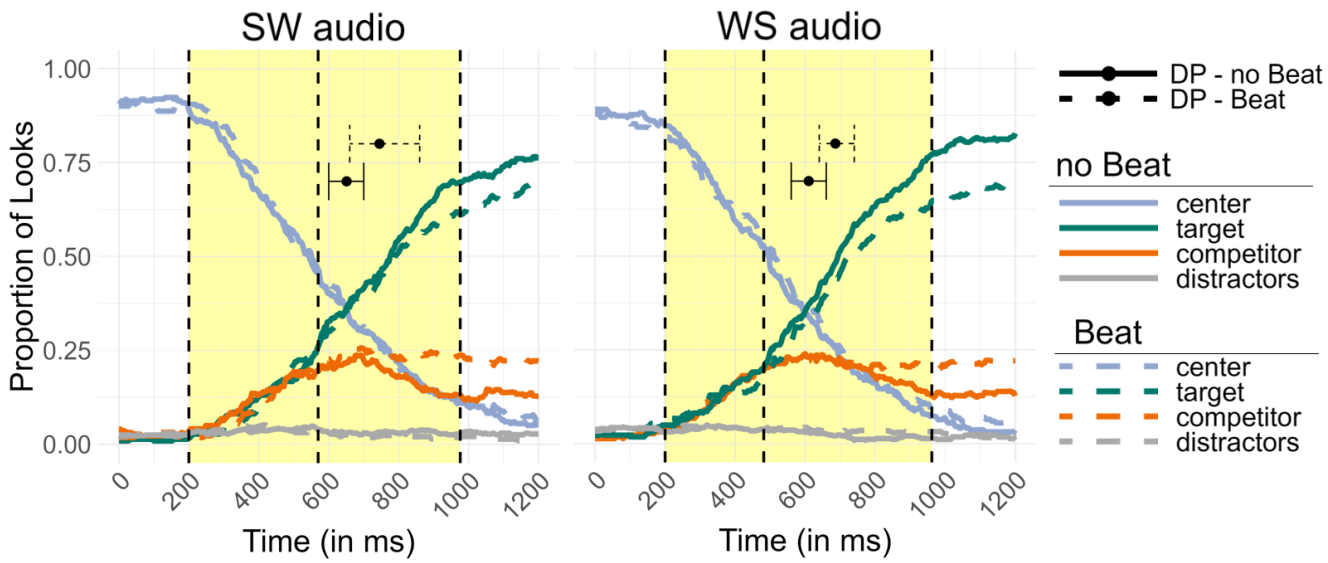
Incongruent Stress Cues

Our dataset also included trials with incongruently timed beat gestures and thus conflicting stress cues. We created two subsets based on the auditory stress cues (SW and WS) and ran the models on them separately. The models were identical to the ones used for analyzing and comparing no-beat and Congruent Beat trials with original audio, as described above. Thus, we tested whether incongruent beat gestures affected the time-course of lexical stress perception. Note that this time all trials were incongruent. Trials without beat gesture still had conflicting articulatory cues (e.g., audio: *VOORnaam*; visual articulatory cues: *voorNAAM*). However, considering that we did not find any effect of the articulatory facial cues to stress on behavioral categorization data nor on participants' fixation behavior, we did not expect any effects of incongruent articulatory cues here. Moreover, beat trials had incongruent articulatory cues and additionally incongruent beat gesture alignment. The only difference between beat trials and no-beat trials was thus the presence of a beat gesture and hence we can interpret any differences between beat trials and no-beat trials as effects of Beat Presence.

SW Audio & WS Visual Stress. The first analysis on the subset with auditory SW stress and thus incongruent visual stress cues on the second syllable found similar effects to the congruent trials (see Figure 8). Generally, participants showed preferential fixations to the auditory target over the auditory competitor when considering the entire word duration (β = 1.093, *SE* = 0.225, *t* = 4.868, *p* < 0.001). Moreover, the effect was bigger in the Time Window of the second syllable than in the Time Window of the first syllable (β = 2.168, *SE* = 0.329, *t* = 6.598, *p* < 0.001). When we relevelled the model so that the second syllable was on the intercept, we found that Beat Presence was close to the threshold of significance (β = -0.669, *SE* = 0.365, *t* = -1.835, *p* = 0.07). Visual inspection of the data also confirmed that the target preference appeared smaller for beat vs. no-beat trials. Therefore, we ran a post-hoc model with an extended Time Window with the duration of the second syllable + 200 post word offset. The Beat Presence effect was significant (β = -0.716, *SE* = 0.33, *t* = -2.174, *p* = 0.032), suggesting a smaller auditory target preference when an incongruent beat gesture was

presented. A DPA estimated the divergence point for beat trials at 745ms (95% CI = [660, 860]), and for no-beat trials at 651ms (95% CI = [600, 700]). However, the 94 ms difference was not statistically significant (p = 0.8). Therefore, even though it didn't delay gaze behavior, post-hoc analyses suggest that an incongruently timed beat gesture decreased the target-over-competitor preference, demonstrating greater uncertainty in word recognition.

WS Audio & SW Visual Stress. The second subset included the trials with auditory WS stress and incongruent visual stress cues on the first syllable. The model revealed the expected auditory target preference over the entire word duration (β = 1.08, *SE* = 0.243, *t* = 4.446, *p* < 0.001), as well as an effect of Time Window (β = 2.216, *SE* = 0.311, *t* = 7.115, *p* < 0.001) (see Figure 8). Similar to the prior analysis, the effect of Beat Presence was not significant in the Time Window of the second syllable (β = -0.535, *SE* = 0.356, *t* = -1.501, *p* = 0.137). But in a post-hoc analysis with the extended Time Window (i.e., second syllable + 200 ms), Beat Presence affected preferential fixations (β = -0.744, *SE* = 0.342, *t* = -2.176, *p* = 0.033). This indicates that the auditory target preference was smaller for beat trials (with incongruent beat gesture) than no-beat trials. The divergence points for beat trials was found at 686 ms (95% CI = [640, 740]), and for no-beat trials at 610 ms (95% CI = [560, 660]). The difference, however, was not significant (*p* = 0.986). So once again, despite the fact that the time point of the effect was not significantly delayed, we found a decreased target-over-competitor preference in proportion of looks in an extended time window, suggesting greater uncertainty.



588

589 Figure 8. Comparison of fixations across time between no-beat trials and trials with *incongruent* beat.
590 Participants looked at the target more than the competitor in the time window of the second syllable.
591 However, when an incongruent Beat was presented, the effect was smaller when compared to no-beat trials.
592 divergence points on trials with an incongruent beat gestures were numerically later but the effect was not
593 significant (SW audio: no Beat = 651 ms, Beat on 2nd = 745 ms; WS audio: no Beat = 610 ms, Beat on 1st = 686
594 ms). Note: shaded areas indicate relevant time windows of analysis. Vertical dashed lines correspond to word
595 onset, second syllable onset and word offset. All timings were shifted by 200 ms. Both distractors were
596 averaged and plotted together. DP = Divergence Point.

597

598

599 **General Discussion**

600 The current study investigated the use of auditory and visual beat gesture cues in the
601 perception of lexical stress, and the time-course of their effects. Our categorization data confirmed
602 that the alignment of beat gestures to the speech input affected lexical stress perception. When
603 presented with a beat gesture aligned to the first syllable, participants were more likely to perceive
604 stress on the first syllable and when hearing the exact same speech but with a beat gesture aligned
605 to the second syllable, participants were more likely to perceive stress on the second syllable. In
606 contrast, no evidence was found that visual articulatory cues on the face affected categorization

responses. Whether participants saw a video of a talker visually producing a SW word or WS word, they perceived it similarly, even when the audio was ambiguous. This is in line with a previous study which found no effects of visual articulatory cues on audiovisual lexical stress perception (Bujok et al., 2024).

In contrast, the effect of beat gesture alignment on lexical stress perception – as observed here – has been consistently demonstrated by several earlier studies (Bosker & Peeters, 2021; Bujok et al., 2024; Rohrer et al., 2024). Note that the effect was present in all audio conditions, so even when the audio was unambiguous and the beat gestures hence redundant, but the effect was numerically largest when speech was most ambiguous. Other studies similarly report the largest beat effects when the speech is most ambiguous (Bujok et al., 2024; Rohrer et al., 2024), although one does not (Bosker & Peeters, 2021). It is thus likely that the effect of the beat gestures on lexical stress perception is moderated by the auditory properties of the stimuli.

Turning to participants' fixations within the video stimuli, as shown in Figure 5, most looks within this area were focused on the talker's face, corroborating earlier work using gesture videos (Gullberg & Holmqvist, 1999, 2006) and only relatively few looks fell on the gestural space (i.e., where the beat gesture was produced). Yet, we still found consistent biasing effects of the beat gesture alignment on lexical stress perception. Beat gestures appear to affect speech perception even when only viewed in the visual periphery. This is in line with previous research finding no reduced uptake of gesture information when gestures were not fixated but seen peripherally (Gullberg & Kita, 2009).

Our analyses of fixations on the various response options first assessed the uptake of auditory stress cues. Thus, we aimed to replicate earlier studies (Jesse et al., 2017; Reinisch et al., 2010), but this time using audiovisual videos as stimuli. Results revealed that participants began to show preferential fixations to target over competitor before word offset. This means that auditory lexical stress cues were tracked continuously and affected lexical access in an on-line fashion. Previous studies using audio-only stimuli have reported similar findings (Jesse et al., 2017; Reinisch et

al., 2010) but there are also some differences. These studies used stimuli which overlapped only in the first two syllables (e.g., *OCtopus* vs. *okTOber*), and found that participants used the lexical stress information on the first two segmentally identical syllables for on-line word recognition. Specifically, they found that the target preference emerged earliest for words with word-initial stress (e.g., *OCtopus*; compared to medial stress position, e.g., *okTOber*). The authors concluded that the presence of stress cues on the first syllable (e.g., OC-) is likely more informative than the absence of stress (e.g., oc-).

In contrast, our results did not show earlier divergence points for SW original audio than WS original audio. Judging from the divergence points, we see that this lack of difference is driven in part by an early target preference in trials with the WS original audio. Perhaps participants realized the critical role of lexical stress for our minimal pair stimuli (in contrast to the stimuli used by Reinisch et al., 2010) and hence used the absence of stress cues (i.e., on the first syllable in WS trials) to guide their word recognition as early as the presence of stress cues (i.e., on the first syllable in SW trials). However, we also see that the target preference in trials with SW original audio is relatively late compared to earlier studies. This may be attributed to the use of *audiovisual* stimuli in our study. The attention-grabbing videos may have delayed looks to the response options in general (compared to audio-only stimuli), making it difficult to find a statistically reliable target preference early in the word in original SW trials. Still, even with audiovisual stimuli, we arrive at the same main conclusion as previous studies (Jesse et al., 2017; Reinisch et al., 2010), namely that auditory lexical stress affects lexical access on-line.

Crucially, the present study found that simple up-and-down beat gestures can guide on-line word recognition as well. When exposed to ambiguously stressed audio, paired with a disambiguating beat gesture (e.g., Beat on 1st syllable), participants began to preferentially fixate the word with the stress pattern indicated by the beat gesture (e.g., SW word) even before the talker finished speaking. This is in line with the categorization data, where participants showed the same gesture-driven bias. Importantly, the fixation data showed that the effect of early beat gestures (i.e.,

on the first syllable) could be detected well before word offset, highlighting their on-line contribution to lexical access. When participants saw a late beat gesture (i.e. on the second syllable), we detected a statistically reliable preference for the WS over SW words around word offset, suggesting that the fixations to WS and SW words already started to diverge before word offset. Moreover, the time interval between the beat gesture apex and the divergence point was similar for beat gestures on the first (~402 ms) and second syllable (~380 ms) at around 400 ms. Together, our data support the view that beat gestures can be used in an on-line manner, guiding lexical access as a spoken word unfolds over time. These observations argue against possible explanations that beat gestures are only used in relatively late task-related decision making.

Given the fact that the onset of beat gestures can precede the onset of the stressed syllable by approximately 300 ms (Leonard & Cummins, 2011), one might have expected even earlier biasing effects of the beat gestures. However, we posit that participants focused mostly on the timing of the apex of the beat gesture, as it is the least variable and most informative part of the gesture (Leonard & Cummins, 2011). Moreover, the divergence point is only a measure of when the divergence is *statistically detectable*. Importantly, the DPA we used required a sustained preference of fixations of at least 200 ms to establish a divergence point, which might have been relatively conservative to find small differences in timing. In fact, visual inspection of the results of trials with ambiguous audio suggests a possibly earlier effect. Furthermore, the ambiguous audio might have added uncertainty about the "correct" response, which is reflected in relatively small differences in the proportions of looks to SW target and WS target. This uncertainty could also have delayed the point at which the divergence was reliably detectable. Still, the present study can conclude that beat gestures affect on-line lexical access, highlighting how both auditory *and* visual stress cues can be used in incremental word recognition.

Furthermore, additional analyses on trials with original audio (i.e., unambiguous stress cues) and incongruent (i.e., misaligned) beat gestures revealed reduced looks to the auditory target when compared to trials without beat gestures. This supports the behavioral data that showed that even

when audio had a clear stress pattern, incongruent beat gestures significantly affected responses in the other direction (i.e., opposite to the auditory stress pattern). That means that even on trials where categorization based on audio was sufficient, incongruent visual cues still affected on-line lexical activation. This finding suggests that stress information conveyed by the temporal alignment of beat gestures is difficult to ignore and can even affect the processing of clear, unambiguous speech. Still, when participants were presented with original audio and *congruent* beat gestures, beat gestures did not change fixation behavior when compared to no-beat trials. That is, we found no facilitatory, let alone a predictive effect of beat gestures on word recognition when speech was clear. However, some electrophysiological studies (Biau et al., 2015; Biau & Soto-Faraco, 2013, 2015) have found that beat gestures induce early phase-resetting of brain oscillations associated with the prediction of the timing of words (Arnal & Giraud, 2012). Therefore, while beat gestures congruently timed with clear acoustic stress cues might not further facilitate lexical stress perception, they still have an important function in word recognition in a larger context, predicting the onset of upcoming words.

Another consideration is that participants' ability to perceive lexical stress in clear, unambiguous audio might have already been at ceiling making it difficult to detect a small facilitatory effect of congruently timed beat gestures. Yet most speech isn't as clear as in a lab-based setting. In many conversational settings, the auditory signal is less accessible, for instance through masking by background noise. In such scenarios, speech perception is not at ceiling and visual processing is enhanced (e.g., Drijvers & Özyürek, 2017). For example, degrading speech with noise could make participants more likely to weight the salient beat gesture more heavily, which could enhance prediction of the stressed syllable based on said beat gesture (Biau et al., 2015; Biau & Soto-Faraco, 2013, 2015) and reveal facilitatory effects of congruently timed beat gestures. Indeed, the emergence of the target preference effect in lexical stress perception was numerically earlier on trials with original SW words and congruent beat gestures (i.e., Beat on 1st) when compared to no-

beat trials, and numerically later on trials with original audio and incongruent beat gestures. Thus, testing this effect in degraded speech could be very informative.

Although the present study was not designed to test any specific models of speech perception, our findings have implications for such models. Current models of audiovisual speech perception (e.g., the Fuzzy Logic Model of Perception (see Massaro, 1998); the Supramodal Brain (Rosenblum, 2019)) were primarily designed to explain segmental speech perception from "talking faces" (Massaro, 1998), and thus do not make predictions about prosody and gestures. The outcomes of our study therefore warrant extension of such models to include suprasegmental speech cues as well as non-articulatory visual cues, such as gestures. Still, they do make predictions about the timing of audiovisual integration. Some models, for example the Fuzzy Logic Model of Perception (FLMP; see Massaro, 1998), assume that visual and auditory information are identified separately and matched to learned prototypes in parallel (e.g., how much a phoneme matches an abstract prototype representation) before they are integrated into one multisensory percept. In contrast, the Supramodal Brain (Rosenblum et al., 2017) is a model that assumes that the multisensory speech cues that we perceive are not tied to a specific modality and are thus processed supramodally. It would suggest that a beat gesture is not processed independently as a visual cue alone but rather together with the auditory (and other multimodal) cues. Hence the main difference between these two models is the timing of the integration of information from different modalities. The FLMP proposes a rather late integration stage, whereas the Supramodal Brain assumes early integration. From the findings of the current study, we cannot make conclusive statements about early vs. late integration. However, we can conclude that integration must be fast and continuous, since we clearly find effects while the word is still being uttered. In other words, audiovisual integration must occur on a smaller temporal level than the word-level.

In conclusion, listeners can use relevant auditory and visual cues to make quick and accurate decisions about word prosody and guide word recognition. Crucially, the alignment of simple hand gestures with speech can affect word recognition immediately, even before a word has been fully

uttered. This supports the idea that beat gestures do not only affect later task-related decision making, but are processed quickly and continuously during audiovisual speech perception. Together these findings highlight the inherent multimodality of speech perception in face-to-face communication.

Data Availability

All experimental data, including scripts and stimuli are publicly available on OSF (https://osf.io/57dvh/) under a CC-By Attribution 4.0 International license.

Acknowledgements

Funded by an ERC Starting Grant (HearingHands, 101040276) from the European Union awarded to Hans Rutger Bosker. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. The authors would also like to express gratitude to Maarten van den Heuvel for the technical support and Jennifer Sander for advice on eye-tracker usage and data analysis. Finally, we would like to thank the SPEAC research group, as well as the Sound Learning research group, and the Psychology of Language Department for their valuable feedback and comments on the study.

References

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398. https://doi.org/10.1016/j.tics.2012.05.003

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, *124*(2), 143–152. https://doi.org/10.1016/j.bandl.2012.10.008

Biau, E., & Soto-Faraco, S. (2015). Synchronization by the hand: The sight of gestures modulates low-frequency activity in brain responses to continuous speech. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00527

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. https://doi.org/10.1016/j.cortex.2014.11.018

Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer [Computer program]* [Computer software]. http://www.praat.org/

Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1943), 20202419. https://doi.org/10.1098/rspb.2020.2419

Bujok, R., Meyer, A. S., & Bosker, H. R. (2024). Audiovisual Perception of Lexical Stress: Beat Gestures and Articulatory Cues. *Language and Speech*, 00238309241258162. https://doi.org/10.1177/00238309241258162

Connell, K., Hüls, S., Martínez-García, M. T., Qin, Z., Shin, S., Yan, H., & Tremblay, A. (2018). English Learners' Use of Segmental and Suprasegmental Cues to Stress in Lexical Access: An Eye-Tracking Study. *Language Learning*, *68*(3), 635–668. https://doi.org/10.1111/lang.12288

Cutler, A., & Jesse, A. (2021). Word Stress in Speech Perception. In *The Handbook of Speech Perception* (pp. 239–265). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119184096.ch9

780 Dimitrova, D., Chu, M., Wang, L., Özyürek, A., & Hagoort, P. (2016). Beat that Word: How Listeners
781 Integrate Beat Gesture and Focus in Multimodal Speech Discourse. *Journal of Cognitive*
782 *Neuroscience*, *28*(9), 1255–1269. https://doi.org/10.1162/jocn_a_00963
783 Drijvers, L., & Özyürek, A. (2017). Visual Context Enhanced: The Joint Contribution of Iconic Gestures
784 and Visible Speech to Degraded Speech Comprehension. *Journal of Speech, Language, and*
785 *Hearing Research*, *60*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101
786 Florit-Pons, J., Vilà-Giménez, I., Rohrer, P. L., & Prieto, P. (2023). Multimodal Development in
787 Children's Narrative Speech: Evidence for Tight Gesture–Speech Temporal Alignment
788 Patterns as Early as 5 Years Old. *Journal of Speech, Language, and Hearing Research*, *66*(3),
789 888–900. https://doi.org/10.1044/2022_JSLHR-22-00451
790 Gullberg, M., & Holmqvist, K. (1999). Keeping an eye on gestures: Visual perception of gestures in
791 face-to-face communication. *Pragmatics & Cognition*, *7*(1), 35–63.
792 https://doi.org/10.1075/pc.7.1.04gul
793 Gullberg, M., & Holmqvist, K. (2006). What speakers do and what addressees look at: Visual attention
794 to gestures in human interaction live and on video. *Pragmatics & Cognition*, *14*(1), 53–82.
795 https://doi.org/10.1075/pc.14.1.05gul
796 Gullberg, M., & Kita, S. (2009). Attention to Speech-Accompanying Gestures: Eye Movements and
797 Information Uptake. *Journal of Nonverbal Behavior*, *33*(4), 251–277.
798 https://doi.org/10.1007/s10919-009-0073-2
799 Holler, J., & Levinson, S. C. (2019). Multimodal Language Processing in Human Communication.
800 *Trends in Cognitive Sciences*, *23*(8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006
801 Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language
802 processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171.
803 https://doi.org/10.1016/j.actpsy.2010.11.003

Ito, A., & Knoeferle, P. (2023). Analysing data from the psycholinguistic visual-world paradigm: Comparison of different analysis methods. *Behavior Research Methods*, *55*(7), 3461–3493. https://doi.org/10.3758/s13428-022-01969-3

Jesse, A., & McQueen, J. M. (2014). Suprasegmental Lexical Stress Cues in Visual Speech can Guide Spoken-Word Recognition. *Quarterly Journal of Experimental Psychology*, *67*(4), 793–808. https://doi.org/10.1080/17470218.2013.834371

Jesse, A., Poellmann, K., & Kong, Y.-Y. (2017). English Listeners Use Suprasegmental Cues to Lexical Stress Early During Spoken-Word Recognition. *Journal of Speech, Language, and Hearing Research*, *60*(1), 190–198. https://doi.org/10.1044/2016_JSLHR-H-15-0340

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32. https://doi.org/10.1016/S0749-596X(02)00505-3

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*(3), 396–414. https://doi.org/10.1016/j.jml.2007.06.005

Kushch, O., & Prieto, P. (2016). *The Effects of pitch accentuation and beat gestures on information recall in contrastive discourse*. https://doi.org/10.21437/SpeechProsody.2016-189

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, *26*(10), 1457–1471. https://doi.org/10.1080/01690965.2010.500218

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. MIT Press.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, *53*(4), 372–380. https://doi.org/10.3758/BF03206780

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

Mitterer, H., & Reinisch, E. (2017). Visual speech influences speech perception immediately but not automatically. *Attention, Perception, & Psychophysics*, *79*(2), 660–678. https://doi.org/10.3758/s13414-016-1249-6

Perniss, P. (2018). Why We Should Study Multimodal Language. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01109

Reinisch, E., Jesse, A., & McQueen, J. M. (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. *Quarterly Journal of Experimental Psychology*, *63*(4), 772–783. https://doi.org/10.1080/17470210903104412

Rietveld, T., & Heuven, V. J. van. (2009). *Algemene Fonetiek (3e geheel herziene druk)*. Bussum : Coutinho. https://repository.ubn.ru.nl/handle/2066/79395

Rohrer, P. L., Bujok, R., van Maastricht, L., & Bosker, H. R. (2024). The timing of beat gestures affects lexical stress perception in Spanish. *Speech Prosody 2024*. Speech Prosody. https://pure.mpg.de/rest/items/item_3582989/component/file_3582990/content

Rosenblum, L. D. (2008). Speech Perception as a Multimodal Phenomenon. *Current Directions in Psychological Science*, *17*(6), 405–409. https://doi.org/10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D. (2019). Audiovisual Speech Perception and the McGurk Effect. In L. D. Rosenblum, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.420

Rosenblum, L. D., Dias, J. W., & Dorsi, J. (2017). The supramodal brain: Implications for auditory perception. *Journal of Cognitive Psychology*, *29*(1), 65–87. https://doi.org/10.1080/20445911.2016.1181691

Scarborough, R., Keating, P., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English. *Language and Speech*, *52*(2–3), 135–175. https://doi.org/10.1177/0023830909103165

Severijnen, G. G. A., Bosker, H. R., & McQueen, J. M. (2024). Your “VOORnaam” is not my “VOORnaam”: An acoustic analysis of individual talker differences in word stress in Dutch. *Journal of Phonetics*, *103*, 101296. https://doi.org/10.1016/j.wocn.2024.101296

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*. https://doi.org/10.3389/fpsyg.2018.01514

Stone, K., Lago, S., & Schad, D. J. (2021). Divergence point analyses of visual world data: Applications to bilingual research. *Bilingualism: Language and Cognition*, *24*(5), 833–841. https://doi.org/10.1017/S1366728920000607

Sulpizio, S., & McQueen, J. M. (2012). Italians use abstract knowledge about lexical stress during spoken-word recognition. *Journal of Memory and Language*, *66*(1), 177–193. https://doi.org/10.1016/j.jml.2011.08.001

Xie, X., Buxó-Lugo, A., & Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition*, *211*, 104619. https://doi.org/10.1016/j.cognition.2021.104619

Yasinnika, Y., Renwick, M., & Shattuck-Hufnagel, S. (2004). The Timing of Speech-Accompanying Gestures with Respect to Prosody. *Journal of the Acoustical Society of America*, *5*. https://doi.org/10.1121/1.4780717