Introduction

Understanding spoken language is a cognitively challenging task. The primary medium of communication in spoken languages, namely auditory speech, is highly variable and noisy. That is, the same word can sound differently depending on the talker, the adjacent phonemes, and even the room acoustics. In face-to-face communicative settings, listeners can make use of at least two sources of information to disambiguate the variable speech signal: surrounding acoustic context (e.g., interpreting ambiguous sounds relative to the preceding speech rate and/or average formant frequencies) and visual articulatory cues (e.g., lip and mouth movements). However, how these two sources of information interact in audiovisual spoken language comprehension is unclear. For instance, do visual cues to prosodic context, such as a fast moving mouth cueing high speech rate, influence following vowel length perception? The present two experiments demonstrate that speech comprehension is indeed influenced by visual articulatory cues to fast and slow speech rates, but only at the stage of audiovisual integration.

Words are typically encountered in rich acoustic contexts, including, for instance, the preceding words in a sentence. Speech researchers have long recognised that the prosodic characteristics of the surrounding acoustic context can influence the perception of subsequent words.
Penel, 2004), but this auditory advantage in rhythmic synchronisation has been attributed to the unrealistic nature of the visual stimuli used (Iversen et al., 2015). Note however that these studies all concerned non-speech stimuli.

Electrophysiological evidence from audiovisual speech processing suggests that watching a talker speak enhances the cortical capacity to track the temporal speech envelope (relative to audio-only stimuli), especially in multitalker settings (Crosse et al., 2015; Golumbic et al., 2013; Schroeder et al., 2008). This presumably involves neural oscillations in visual cortex aligning to a talker’s lip movements during continuous speech processing, aiding intelligibility (Park et al., 2016). However, how these two oscillatory functions (sustained entrainment based on acoustic cues influencing temporal sampling vs. visually induced entrainment) interact is unknown. For instance, could only watching a talker produce fast and slow articulatory lip movements induce oscillatory speech-tracking at fast vs. slow frequencies, with consequences for the temporal sampling of following auditory words?

There is some evidence in the literature that visual prosodic cues to speech rate may influence speech perception. Listeners can estimate a talker’s speech rate as accurately from visual-only as from audio-only stimuli (Green, 1987). Moreover, Green and Miller (1985) demonstrated that these visual cues to speech rate can influence audiovisual integration. They presented participants with ambiguous auditory /bi-pi/ continua, varying VOT from short values (most /bi/-like) to long values (most /pi/-like). These auditory speech sounds were combined with videos of a talker saying /bi/ and /pi/ at fast and slow rates. Results showed that participants were more likely to report hearing /pi/ (i.e., long VOT) when the ambiguous target sounds were combined with visual cues to a fast speech rate (and vice versa). The same effect has been observed using fast and slow videos that differ from the auditory target sounds in place of articulation: hearing ambiguous /bi-pi/ tokens combined with videos of a talker saying /ti/ at a fast speech rate (vs. slow speech rate) also biases perception towards /pi/ (Brancazio & Miller, 2005). This suggests that this perceptual bias is indeed driven by visual cues to speech rate and not by visual articulatory cues about the consonant itself. However, these two studies only assessed effects of visual speech rate on the target word itself (i.e., concurrent with target word presentation). As such, it remains unclear whether contextual speech rate (i.e., the speech rate in a lead-in sentence) can influence following target word perception.

To our knowledge, there is only one piece of evidence suggesting that visual cues to contextual speech rate induce rate-dependent perception of following target words. At the Annual Meeting of the Psychonomic Society in 2013, Jesse and Newman (2013) reported an experiment in which participants were shown visual-only stimuli (i.e., a muted talking face) of a talker producing the context
sentence “Sarah brought a bag so Paul could get the . . .” This context sentence was followed by audiovisual targets (face + voice) that were ambiguous between /bɪn/ (with short VOT) and /pɪn/ (with long VOT). The authors observed a visual rate effect on audiovisual target perception: fast (i.e., linearly compressed) versions of the context video induced more /pɪn/ responses than slow versions did. However, because only muted videos were used as contexts, comparison to audio-only and audiovisual rate-manipulated contexts could not be made. Moreover, the target words in the aforementioned study were presented audiovisually. As such, it remains unknown whether the visually cued contextual speech rate influenced the perception of the auditory target cues or the perception of the visual target cues. Specifically, we here ask whether speech rate information is represented in a modality-specific or a modality-independent manner. That is, would we also be able to observe rate-dependent perception of auditory target words in the context of a complete switch in modality (e.g., from visual-only to audio-only)?

It could be that audiovisual cues to contextual speech rate are encoded in a modality-independent manner. That is, the speech rate of the context sentence may be represented in a manner that is not specific to the modality that cued the speech rate. This would predict that visual cues to contextual speech rate can influence the perception of subsequent auditory target cues, guiding participants’ target categorisation responses. We will refer to this account as the “cross-modal transfer” account (cf. Figure 1). This proposal would be grounded in the notion of a “supramodal” architecture of multisensory speech comprehension (as advocated by Rosenblum, 2019; Rosenblum et al., 2017), proposing that the speech processing system acts to extract supramodal informational patterns that are common in form across sensory streams. Support for such a supramodal architecture comes from observations that viewing a silent video of an articulating face can induce activity in auditory brain areas in novice lipreaders (Calvert et al., 1997); experience with silently lipreading a talker allows individuals to subsequently better comprehend that talker’s audio-only speech (Rosenblum et al., 2007); imagined visual gender information influences vowel category boundaries (Johnson et al., 1999); and non-articulatory visual information can change vowel perception (Hay & Drager, 2010).

Alternatively, the speech rate of a context sentence, presented only in the visual modality, could also be encoded in a modality-specific manner. For instance, consider the situation where participants are presented with visual-only context sentences, followed by audio-only target words ambiguous between containing /ɑ/ or /a:/ . This modality-specific proposal would predict that a fast visually cued contextual speech rate would not bias the perception of the following audio-only target words towards long /a:/ . However, it also predicts that, if the target were presented audiovisually (i.e., concurrently), the visual speech rate in a context sentence might influence the perception of visual target cues. That is, perhaps seeing fast mouth movements in the context time window (i.e., in the visual modality) will influence the perception of the duration of the opening of the mouth in the target window (i.e., also in the visual modality). Then, at the audiovisual integration stage, these visual target cues are integrated with the auditory target cues, resulting in a higher proportion of long /a:/ responses. We will refer to this account as the “cross-modal integration” account (cf. Figure 1). Relevant to this modality-specific account of speech rate encoding are findings reported by Bosker et al. (2017). In their study, participants
were presented with auditory rate-manipulated context sentences followed by ambiguous target words. In addition, participants had to simultaneously perform a demanding concurrent task in the visual domain (easy vs. difficult visual search). Outcomes demonstrated that rate-dependent perception effects were as strong with a difficult vs. easy visual concurrent task (Bosker et al., 2017), possibly suggesting that speech rate encoding is relatively “immune” to modulations in the visual modality.

This study addressed the question whether speech rate is encoded in a modality-specific or in a modality-independent manner by assessing whether and how visual cues to speech rate induce rate-dependent perception of following target words. Participants in the two current experiments were told they would take part in a “Go Fish”–like guessing game: they were presented with a talker telling them which five objects were on her cards (e.g., “This time I have the bone, the pipe, the fruit, the caterpillar, and the . . . ”). Critically, the last object was always ambiguous between containing the short vowel /ɑ/ vs. the long vowel /a:/ in Dutch, for instance, bal /baːl/ “ball” vs. bal /baːl/ “bale.” Participants’ task was to select from two options presented on screen the card they thought was in the talker’s hand (e.g., card with a ball vs. card with a bale of hay). Context sentences were presented at fast vs. slow speech rates allowing assessment of how target perception would change as a function of the preceding contextual speech rate.

Crucially, each experiment manipulated the modality of the context sentences (see Figure 2): audio-only (auditory speech with a static fixation cross on screen; A-only), visual-only (dynamic but mute video of the talker’s face producing the context sentence; V-only), and audiovisual presentation (speech + dynamic face; AV). By contrast, the modality of the targets was always fixed within an experiment. In Experiment 1, the targets were presented auditorily (speech with static fixation cross on screen; A-only), and the targets were presented visually (dynamic videos). In Experiment 1, context sentences were presented auditorily with a fixation cross on screen (i.e., without visual articulatory cues). In the V-only condition, context sentences were only presented visually without any sound (dynamic videos of female talker). In the AV condition, audiovisual context sentences were presented. In Experiment 1, context sentences were followed by auditory target phrases with a fixation cross on screen (i.e., without visual articulatory cues). In Experiment 2, the same context sentences were followed by audiovisual target phrases (dynamic videos).
perception, as there are additional visual articulatory cues (beyond the auditory cues in the speech) to the talker’s rate of speaking. Critically, the V-only condition assessed whether visually cued contextual speech rate is encoded in a modality-independent or in a modality-specific manner. That is, if fast mute videos bias perception of audio-only target words towards long /a:/ while slow mute videos bias perception towards short /a/, this would be evidence for a modality-independent influence of visual rate cues on auditory perception. This influence would then not require the two modalities to be concurrently present at any point, as no visual target cues were provided in Experiment 1.

In contrast, Experiment 2 was identical to Experiment 1 except that targets were presented audiovisually (speech + dynamic face). If Experiment 1 would fail to find evidence for modality-independent rate effects, perhaps visual contextual speech rate does influence the perception of visual target cues. In turn, these visual cues, combined with the auditory target cues, may bias participants’ categorisation responses at an audiovisual integration stage. Hence, if Experiment 2 would find rate-dependent perception effects in the V-only condition, this would be evidence for an initial modality-specific encoding of speech rate, resulting in cross-modal integration at the audiovisual integration stage (if and only if the two modalities are concurrently presented during the target word). Finally, comparison of the rate effects between the various conditions may reveal potential variation in auditory- vs. visually-induced effects.

**Experiment 1**

Experiment 1 involved presenting participants with fast and slow context sentences, followed by target words ambiguous between containing the short vowel /a/ vs. the long vowel /aː/ (e.g., bal /ba:l/ “ball” vs. baal /baal/ “bale”). Context sentences were presented in three conditions: audio-only (A-only; audio with a static fixation cross on screen), visual-only (V-only; mute video of talker’s face), or audiovisual (AV; same video with audio). Target words were always presented audio-only (no video; only fixation cross).

**Methods**

**Participants.** Native Dutch participants (N = 38; 30 females, 8 males; mean age = 22, range = 18–28) were recruited from the Max Planck Institute’s participant pool. Participants had normal hearing, had no speech or language disorders, and took part in only one of our experiments. Participants in all experiments reported in this study 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. We decided a priori to exclude participants with a proportion of “long vowel” responses below 0.1 or above 0.9, as for these participants the presented stimuli would be insufficiently ambiguous to establish reliable effects of speech rate. Based on this criterion, data from six participants were excluded because they reported hearing the long vowel target words in over 90% of the trials (ca. 15%; comparable to other studies testing rate-dependent perception of the /aː-a:/ contrast in Dutch; cf. ca. 15% in Maslowski et al., 2019a; ca. 10% in Maslowski et al., 2019b). Data from a further two participants were excluded because of technical errors. The data from the remaining 30 participants (24 females, 6 males; mean age = 22, range = 18–28) were entered into the analyses described below.

**Stimuli.** Ten Dutch sentences were constructed containing lists of five monosyllabic items (e.g., Dit keer heb ik het bot, de pijp, het fruit, de rups, en de bal/baal; “This time I have the bone, the pipe, the fruit, the caterpillar, and the ball/bale”; see Table S1 in the Supplementary Materials). Sentences always ended in a monosyllabic minimal word pair (target words). These target word pairs differed only in their vowel, containing either short /a/ or long /aː/ (e.g., bal /ba:l/ “ball” vs. baal /baal/ “bale”; see Table S1 in the Supplementary Materials). None of the other items in the lists contained /a/ or /aː/. Coloured line drawings of all target words were retrieved from various online resources (all permitting the noncommercial reuse of the materials).

A female Caucasian native speaker of Dutch was video-recorded using a Canon XF205 camera (50 frames per second; resolution: 1280 by 720 pixels) with an external Sennheiser ME64 directional microphone (audio sampling frequency: 48 kHz). Recordings were made from the shoulders up (entire face visible; neutral background), with the talker speaking all sentences ending in either member of the target minimal pair. Both members of the target minimal pairs were required to create the ambiguous target tokens used in the experiment (see below). She was instructed to speak at a comfortable speech rate. Recorded sentences were divided into context sentences (all speech up to and including the fourth item in the list) and target phrases (including en de “and the,” and the sentence-final target word). One context sentence recording was selected for each target minimal pair. The video content of the context sentences was rate-manipulated using the ffmpeg tool for batch video processing (version 4.0; available from http://ffmpeg.org/). Two ratios were used, resulting in fast context sentences (ratio = 0.66) and slow context sentences (ratio = 1/0.66 = 1.5). The audio content of the context sentences was separated from the video content using ffmpeg, manipulated in rate to match the speed of the videos using PSOLA in Praat (Boersma & Weenink, 2016) which maintains the spectral cues in the audio (formants, F0, etc.), and then combined with the rate-manipulated videos using ffmpeg.

Video-recordings of the target phrases (e.g., en de bal) were converted into an audio-only format, removing the
video from the target phrases. For each of the 10 target word pairs, an individual duration continuum was created from short /a:/ to long /a:/.

Procedure. Participants were tested individually in a sound-conditioned booth. Before the experiment, they were informed that they would take part in a “Go Fish”-like guessing game: they would be presented with a talker telling them which five objects were on her cards (e.g., “This time I have the bone, the pipe, the fruit, the caterpillar, and the [target]”). Their task was to indicate what they thought was on the last card by selecting one out of two response options (i.e., two cards showing two target images of a minimal target pair), one on the left and one on the right (position counterbalanced across participants). Participants entered their response as to which of the two response options they had heard (bal or baal) by pressing the “Z” button on a regular computer keyboard for the image on the left, or “M” for the image on the right. After their response (or time-out after 4 s), the screen was replaced by an empty screen for 500 ms, after which the next trial was initiated automatically.

The V-only block was identical to the AV block (but with a unique random order of trials), except that no audio was played during the context window. That is, participants saw silent videos of fast and slow context sentences, which at context offset were followed by a sudden transition to a fixation cross together with audio-only target phrases. The A-only block was identical to the AV block, except that the visual video stimulus was replaced by a fixation cross. That is, participants continuously saw a fixation cross on screen while listening to audio-only contexts and target phrases.

Block order was counterbalanced across participants (six different lists with five participants each). Six practice trials were presented to participants (two in each condition) to familiarise them with the materials and the task. Participants were given opportunity to take a short break after each block.

Results

Trials with missing data (n=7; <0.1%) were excluded from analyses. Note that, because all target words contained a vowel ambiguous between short /a/ and long /a:/,
our task does not measure “accuracy,” as neither of the two response options can be labelled as “accurate” or “inaccurate.” Instead, we analysed the categorisation data, calculated as the proportion of long /a:/ responses, presented in Figure 3. The average proportion of long /a:/ responses across all data from Experiment 1 was 0.55. As expected, higher steps on the duration continua led listeners to report more /a:/ responses (lines have a positive slope). Differences between the blue/darkgray (slow contexts) and orange/lightgray lines (fast contexts) are indicative of an influence of the preceding context. Fast contexts seem to induce more long /a:/ responses than slow contexts in the A-only block. This rate effect seems to be reduced in the AV block and appears to be absent in the V-only block.

Data were statistically analysed using a generalised linear mixed model (GLMM; Quené & Van den Bergh, 2008) with a logistic linking function as implemented in the lme4 library (version 1.0.5; Bates et al., 2015) in R (R Development Core Team, 2012). The binomial dependent variable was participants’ categorisation of the target as either containing /a:/ (e.g., baal; coded as 1) or containing /ɑ/ (e.g., bal; coded 0). Fixed effects were Continuum Step (continuous predictor; centred around the mean), Context Rate (categorical predictor; deviation coding, with slow coded as –0.5 and fast as +0.5), Condition (categorical predictor with the AV condition mapped onto the intercept), and all interactions. The use of deviation coding of two-level categorical factors (i.e., coded with –0.5 and +0.5) allows us to test main effects of these predictors, because with this coding the grand mean is mapped onto the intercept. All models reported in this study included Participant and Target Item as random factors, with by-participant and by-item random slopes for Context Rate (more complex models failed to converge). Note that simple effects should be interpreted with respect to the AV condition only, because the AV condition was mapped onto the intercept. Interactions with Condition would reveal differential effects in the two other conditions.

The model showed a significant effect of Continuum Step (β = 0.751, SE = 0.037, z = 20.194, p < .001), indicating that, in the AV condition, higher continuum steps led to more long /a:/ responses. It also showed an effect of Context Rate (β = 0.364, SE = 0.105, z = 3.460, p < .001), indicating that, in the AV condition, fast contexts biased listeners towards /a:/.

An interaction between Context Rate and the A-only condition (β = 0.455, SE = 0.141, z = 3.219, p < .001) showed that the effect of Context Rate was greater in the A-only condition. Conversely, an interaction between Context Rate and the V-only condition (β = –0.392, SE = 0.139, z = –2.814, p < .005) showed that the effect of Context Rate was greatly reduced in the V-only condition relative to the AV condition. In fact, a mathematically equivalent model, this time mapping the V-only condition onto the intercept, showed no evidence for an effect of Context Rate in the V-only condition (p = .781).

It could be that the absence of an effect of Context Rate in the V-only condition was due to participants not looking...
on screen during the presentation of the mute videos. Therefore, we had recorded eye fixations on screen so as to assess whether participants followed the instructions to look at the screen during the trials of each block. Unfortunately, however, unexpected technical limitations of the equipment led to unreliable data for a considerable number of trials (e.g., no fixations registered while the experimenter clearly saw the participant looking on screen). Nevertheless, even when we analysed these gaze data, we found few differences between the three conditions (percentage time looking on screen: A-only: 70%; AV: 73%; V-only: 71%) or the two rates (fast: 71%; slow: 72%). When we excluded trials in which participants supposedly looked off screen for more than 25% of the context time window (31% data loss), statistical analyses on this subset of trials did not lead to qualitatively different interpretations.

Finally, one might expect that participants who were presented with the V-only block first (in two out of the six lists; n=10 participants) would have less access to the speech of the speaker than participants who had already heard the speaker talk (auditorily) before being presented with the V-only block. However, extending the model reported above with the predictor V-First (categorical predictor, with the lists that received the V-only block first mapped onto the intercept), interacting with all other predictors, did not reveal a significant interaction between Context Rate and V-First (p=.539). This suggests that being presented the V-only condition first did not change the effect of Context Rate.

**Interim discussion**

The results of Experiment 1 showed that our target duration continua appropriately sampled the perceptual continuum from /a/ (e.g., bal) to /a:/ (e.g., baal). They also demonstrated that audio-only contexts with fast speech rates biased target perception to more /a:/ responses in the A-only block (relative to slow contexts), replicating earlier audio-only studies on rate-dependent perception (Bosker, 2017a, 2017b; Reinisch & Sjerps, 2013). No evidence was found for effects of visual-only contextual rates on target speech perception (no effect of context rate in V-only block), which may indicate that visually cued speech rate is encoded in a modality-specific manner: visual cues to speech rate (in the context window) do not influence the perception of auditory cues to vowel duration (in the target window). However, it does not exclude the possibility that visual rate cues in the context window (e.g., fast or slow moving lips) could potentially influence the perception of following visual cues to vowel duration (e.g., a shorter vs. longer duration of mouth opening), which in turn—at an audiovisual integration stage—may influence the perception of speech sounds. Therefore, Experiment 2 was identical to Experiment 1 except that the target phrases were presented audiovisually (with both audio and accompanying video; see Figure 2).

Note also that the rate-dependent effect in the AV condition was reduced relative to the A-only condition, even though the AV condition presented listeners with more sensory cues to speech rate (audio and video) compared with the A-only condition. As the target phrases in Experiment 1 only involved auditory cues (audio with fixation cross), this meant that the AV and V-only conditions shared a sudden visual change at context offset. That is, in both conditions, the video of the talker was suddenly replaced by a fixation cross at context offset. In contrast, in the A-only condition, participants only ever saw a fixation cross during stimulus presentation. The sudden transition between contexts and targets in the AV and V-only conditions may have had had a detrimental effect on the perceptual binding of contexts and targets, hence potentially reducing context effects in these two conditions (relative to the A-only condition). Using audiovisual targets in Experiment 2 removed the sudden visual changes from the AV and V-only condition, while introducing them in the A-only condition (see Figure 2). Thus, this experimental design additionally allowed us to investigate whether sudden visual transitions in between contexts and targets modulate the contextual rate effect.

**Experiment 2**

Experiment 2 was identical to Experiment 1, except that this time target phrases were presented audiovisually (see Figure 2). As a consequence, the AV and V-only conditions did not have sudden changes from videos to a fixation cross. Instead, in Experiment 2, it was the A-only condition that this time involved a sudden visual change: auditory context sentences were presented with a fixation cross, followed by audiovisual target phrases.

**Methods**

**Participants.** Native Dutch participants (N=39; 29 females, 10 males; mean age = 23, range = 18–28) were recruited from the Max Planck Institute’s participant pool. Participants had normal hearing, had no speech or language disorders, and had not taken part in Experiment 1 nor in the pretest. Based on the exclusion criterion introduced in Experiment 1, data from seven participants were excluded because they reported hearing the long vowel target words in over 90% of the trials. Data from a further two participants were excluded because of technical errors. The data from the remaining 30 participants (23 females, 7 males; mean age = 23, range = 18–28) were entered into the analyses described below.

**Stimuli.** The stimuli were identical to those used in Experiment 1, except that this time audiovisual versions of the
target phrases were presented. Also, in Experiment 1, the context stimuli and the target stimuli involved separate files, which were presented in sequence. Adopting a similar procedure for Experiment 2 could, however, result in slight transition delays (depending on the loading time of the videos), with the onset of the target video being temporally separated from the offset of the context video. Therefore, we decided to use the 10 original sentence-long video-recordings containing each long /a:/ token and manipulate the rate of the context, the duration of the auditory vowel, and the modality of the context (A-only, V-only, AV) manually for each item. This ensured that the visual stimulus was always continuous. This also meant that we had to use a different software package than ffmpeg and Praat, because they do not allow manipulating separate parts of a given video stimulus.

Stimulus manipulations were performed in Adobe Premiere Pro CC 2015. Once more, original video-recordings were divided into context sentences (all speech up to and including the fourth item in the list) and target phrases (including en de, and the sentence-final target word). The context rate was compressed/expanded by the same factors as in Experiment 1 (0.66 and 1.5) using the “Speed/Duration…” function in Adobe Premiere, which maintains pitch and formant frequencies. The duration of the vowel was manipulated by removing the original audio stream in the entire target phrase and replacing it with the manipulated target phrase audio materials from Experiment 1 without any noticeable synchronisation error. Finally, the modality of the context sentence was manipulated by either removing the audio stream (V-only) or by replacing the video stream of the context sentence by a single frame, showing a black fixation cross on a white background (A-only). Note that this resulted in unique video files for each item in all conditions, with only a sudden transition between contexts and targets in the A-only condition (from a fixation cross in the context window to a video of the talker pronouncing the target phrase).

Procedure. The procedure of Experiment 2 was identical to the one in Experiment 1. That is, participants took part in a “Go Fish”–like guessing game, while their eye fixations were recorded, but this time using a tower-mounted EyeLink 1000 eye-tracking system (SR Research) with a chin and forehead rest. Thus, we hoped to collect more reliable eye gaze data.

Results
Trials with missing data (n=9; <0.1%) were excluded from analyses. Categorisation data, calculated as the proportion of long /a:/ responses, are presented in Figure 4. The average proportion of long /a:/ responses across all data from Experiment 2 was 0.71. It would seem that, on the whole, there was a higher proportion of long /a:/ responses in Experiment 2 than in Experiment 1 (0.55), possibly due to the addition of target video stimuli of the talker pronouncing the long vowel /a:/.

More interestingly, however, there seems to be a rate effect in all three conditions: fast contexts in the A-only, AV, and even the V-only block induced a higher proportion of long /a:/ responses than slow contexts.

Data were statistically analysed using another GLMM with the same structure as specified in Experiment 1.
condition was mapped onto the intercept. Interactions with Condition would reveal differential effects in the two other conditions.

This GLMM showed a significant effect of Continuum Step ($\beta = 0.749, SE = 0.039, z = 19.269, p < .001$), indicating that, in the AV condition, higher continuum steps led to more long /a:/ responses. It also showed an effect of Context Rate ($\beta = 0.424, SE = 0.115, z = 3.691, p < .001$), indicating that, in the AV condition, fast contexts biased listeners towards /a:/. There was also an overall difference in the proportion of long /a:/ responses between the AV and V-only condition ($\beta = -0.393, SE = 0.073, z = -5.393, p < .001$) and between the AV and A-only condition ($\beta = -0.227, SE = 0.073, z = -3.107, p = .002$), indicating an overall higher proportion of long /a:/ responses in the AV condition relative to the V-only and A-only conditions.

Although no statistically significant interactions were found, we note that, numerically, the effect of Context Rate was largest in the AV condition ($\beta = 0.424$). Mathematically equivalent models, rotating which Condition was mapped onto the intercept, also showed Context Rate effects for the A-only and V-only conditions (A-only: $\beta = 0.255, SE = 0.108, z = 2.364, p = .018$; V-only: $\beta = 0.257, SE = 0.107, z = 2.404, p = .016$).

We also assessed whether participants who had been presented with the V-only block first showed a smaller Context Rate effect than other participants. Extending the GLMM reported above with the predictor V-First (categorical predictor, with the lists that received the V-only block first mapped onto the intercept), interacting with all other predictors, did not reveal a significant interaction between Context Rate and V-First ($p > .6$). As in Experiment 1, this suggests that being presented the V-only condition first did not change the effect of Context Rate.

The results from Experiment 2 would seem to differ from Experiment 1 in two ways. First, the overall proportion of long /a:/ responses seems to be higher in Experiment 2 than in Experiment 1. Second, the effect of Context Rate was observed for all three conditions in Experiment 2, while it was absent in the V-only condition in Experiment 1. These observations were statistically verified by running an omnibus GLMM on the combined data from Experiment 1 and 2. The structure of this GLMM was identical to the previous GLMMs, except that the additional predictor Experiment (categorical predictor with Experiment 1 mapped onto the intercept) was included, interacting with all other predictors. The simple effects of this omnibus GLMM demonstrated all the results reported in Experiment 1 (e.g., effect of Step, Context Rate, interactions between Context Rate and Conditions), because Experiment 1 was mapped onto the intercept. In addition, it showed an effect of Experiment, confirming that there was indeed a higher proportion of long /a:/ responses in Experiment 2 relative to Experiment 1 ($\beta = 1.321, SE = 0.283, z = 4.661, p < .001$).

With the AV condition mapped onto the intercept, no interaction between Context Rate and Experiment was observed ($\beta = 0.126, SE = 0.152, z = 0.831, p = .406$), suggesting that the effect of Context Rate in the AV condition was comparable in both experiments. Mapping the A-only condition onto the intercept did reveal a Context Rate × Experiment interaction ($\beta = -0.462, SE = 0.148, z = -3.119, p = .002$), indicating that the effect of Context Rate was significantly smaller in Experiment 2 than in Experiment 1. Finally, mapping the V-only condition onto the intercept also revealed a Context Rate × Experiment interaction, but in the opposite direction ($\beta = 0.307, SE = 0.147, z = 2.095, p = .036$), indicating that the effect of Context Rate was significantly larger in Experiment 2 than in Experiment 1.

Similar to Experiment 1, we also collected eye-tracking data from participants during the context time window to assess whether participants indeed looked at the screen, as instructed. In contrast to Experiment 1, Experiment 2 used a tower-mounted eye-tracker with a chin and forehead rest, resulting in more reliable data. The average percentage of time that participants looked at the screen during the context time window was comparable for the different conditions (A-only: 78%; AV: 81%; V-only: 78%) and for the two rates (fast: 80%; slow: 78%). Excluding the trials in which participants supposedly looked off screen for more than 25% of the context time window (25% data loss) did not lead to qualitatively different interpretations of results.

**Interim discussion**

The results of Experiment 2 showed, first of all, that there was an overall increase in the proportion of long /a:/ responses relative to Experiment 1 (lines in Figure 4 are higher than lines in Figure 3). This is likely due to the fact that Experiment 2 additionally included visual cues to the target words (i.e., audiovisual rather than audio-only targets). Specifically, the target videos consistently showed the talker pronouncing the long vowel /a:/, combined with various auditorily ambiguous target words. The Dutch /a-/ vowel contrast is cued by spectral and temporal differences (Escudero et al., 2009), which would presumably be visible from the articulatory movements our talker made (wider and longer lip aperture for /a:/ compared with /a/), hence accounting for the overall difference between Experiments 1 and 2. Moreover, this finding suggests that participants were indeed sensitive to the visual articulatory cues presented on screen, despite the eye-tracker only registering looks on screen approximately 80% of the time.

Second, we observed a rate effect in all three conditions in Experiment 2, including the V-only condition. That is, audiovisually presented target words were more likely to be perceived to contain the long vowel /a:/ if preceded by a fast context sentence—indeed from the modality of the context sentence (AV, A-only, V-only). This finding
contrasts with Experiment 1: while Experiment 1 did not find evidence for V-only contexts to influence audio-only target words, Experiment 2 demonstrated that V-only contexts do influence the perception of audiovisual target words. This suggests that the rate effect induced by V-only contexts in Experiment 2 operates only via the visual cues in the target word window. This will be discussed in greater detail in the “General Discussion” section.

Finally, we found that the rate effect in the A-only condition was reduced in Experiment 2 relative to Experiment 1. This may be explained in the same terms as the reduced rate effect in the AV condition in Experiment 1. That is, both the A-only condition in Experiment 2 and the AV condition in Experiment 1 included an abrupt visual change at context offset (from static fixation cross to dynamic video, and from video to fixation cross, respectively). These highly salient and sudden visual transitions may have negatively affected the perceptual binding of contexts and targets, hence reducing the size of the rate effect in these two conditions.

**General discussion**

The present two experiments addressed the question whether speech rate is encoded in a modality-independent or a modality-specific manner by testing whether and how visual articulatory cues to speech rate induce rate-dependent perception of following ambiguous target words. In a “Go Fish”–like guessing game, participants categorised ambiguous target words midway between the short vowel /a/ and the long vowel /aː/ (e.g., bal /bal/ “bale” vs. baal /baːl/ “bale”), preceded by rate-manipulated audio-only (A-only), visual-only (V-only), and audiovisual (AV) context sentences (see Figure 2). Crucially, Experiment 1 used audio-only target words (ambiguous target sounds + static fixation cross), whereas Experiment 2 used audiovisual target words (ambiguous target sounds + a video of the talker producing the target member with long /aː/). Results showed consistent rate effects in the A-only and AV conditions in both experiments: fast speech rates biased participants’ target word perception towards long /aː/. However, Experiment 1 did not find evidence for V-only contexts to influence audio-only target words. Note that participants likely could estimate the speech rate from the mute videos, because the fast vs. slow distinction was very salient (i.e., “fast” was more than twice as fast as “slow”) and listeners are generally as accurate to estimate speech rates from visual stimuli as they are from auditory speech (Green, 1987). Most participants could presumably even “reconstruct” (i.e., lipread) the spoken words from the mute videos, because the V-only block was often preceded by the A-only and/or AV blocks (for 20 out of 30 participants), and only 10 sentences were repeated throughout the experiment. Nevertheless, this was insufficient to trigger a rate effect in the V-only condition in Experiment 1. This observation is in line with findings in Bosker (2017b). Participants in that study produced fast and slow sentences themselves, after which they were presented with ambiguous /a-aː/ target words. Target categorisation data showed that overtly producing fast speech oneself did bias target perception towards long /aː/, whereas covertly producing fast speech (without any audible speech) did not. Hence, overt auditory prosodic contexts are necessary to trigger rate-dependent perception of audio-only target words.

In contrast, Experiment 2 revealed that V-only contexts did show the expected rate effect when using audiovisual target words. We take this difference between V-only conditions in Experiment 1 vs. Experiment 2 to suggest that speech rate in V-only contexts is initially encoded in a modality-specific manner, but allowing for cross-modal integration of the auditory and visual rate cues at the audiovisual integration stage (cf. Figure 1). Crucially, this cross-modal integration effect requires the audio and visual information streams to be co-present in the target window. Hence, the rate effect in the V-only condition in Experiment 2 is suggested to operate via the visual cues in the target word window. That is, the visual cues to speech rate in the V-only condition in Experiment 2 likely only influenced the perception of the visual target cues (i.e., visual cues to lip aperture). For instance, a fast moving mouth in the context window likely made the visual cues to the duration of mouth opening in the target window seem longer. In turn, these visual target cues, combined with the auditory target cues, biased participants’ categorisation responses at a later audiovisual integration stage. This suggests that rate-based prosodic context effects in perception are initially modality-specific: visual cues to speech rate (in the context window) do not cross-modally influence the perception of auditory cues to vowel duration (in the target window) in the absence of the two modalities occurring simultaneously in the target window.

This finding would seem to contrast with earlier studies on a different type of acoustic context effect, namely spectral contrast effects (also known as spectral, vowel, or talker normalisation). That is, spectral cues to vowel identity (e.g., low vs. high F1 distinguishing /ɪ/ vs. /ɛ/, respectively) are, like durational cues, also perceived relative to the surrounding acoustic context: a vowel midway between /ɪ/ and /ɛ/ is perceived as /ɛ/ after a context sentence with a relatively low F1, but as /ɪ/ after a context sentence with a high F1 (Asgari & Stilp, 2015; Bosker et al., 2019; Ladefoged & Broadbent, 1957). Interestingly, some studies have reported visually induced spectral contrast effects. For instance, listeners categorise spectrally ambiguous sound continua differently when viewing a video of a male vs. female talker (Strand & Johnson, 1996; Winn et al., 2013). Even merely telling participants they will hear a male vs. female talker can change the perception of vowel continua produced by an androgynous voice (Johnson et al., 1999). The fact that spectral contrast effects are
induced cross-modally by visual cues (Johnson et al., 1999), while rate-dependent perception effects are not (cf. Experiment 1 vs. Experiment 2), indicates that differential cognitive mechanisms may underlie the two seemingly analogous processes, as indeed suggested by recent psychoacoustic and neurobiological evidence (Bosker & Ghitza, 2018; Kösem et al., 2018; Sjerps et al., 2019). In fact, the present rate effects may speculatively be viewed in light of a predictive coding framework that assumes that listeners use the contextual speech rate to implicitly predict the duration of upcoming speech segments. This type of implicit predictive behaviour may operate mechanismically through neuronal entrainment to syllabic rhythms (Kösem et al., 2018). Future work may further relate the available neurobiological, psychoacoustic, and phonetic findings in the literature to predictive coding accounts.

We also observed that abrupt visual transitions between contexts and targets may reduce rate-dependent perception effects. That is, there was a larger rate effect in Experiment 1 in the A-only condition (static fixation cross without sudden transition) compared with the AV condition (dynamic video suddenly changing to a fixation cross), despite the latter containing additional visual cues to speech rate. When a sudden video transition was added to the A-only condition in Experiment 2 (from static fixation cross to dynamic video), the rate effect was reduced as well (relative to A-only in Experiment 1). We speculate that highly salient and sudden visual transitions may negatively affect the perceptual binding of contexts and targets, hence reducing the size of the rate effect. This observation may be considered striking given that the perceptual binding of contexts and targets was resilient to the temporal distance between the rate cues in the context sentence and the ambiguous vowels (i.e., a “buffer” of on average 323 ms separated the contextual speech rate cues from the target vowel). Mechanistically, the detrimental effect of sudden visual transitions could involve visually induced inadvertent phase resetting of low-frequency oscillations in auditory cortex (Columbic et al., 2013; Kayser et al., 2008; Schroeder et al., 2008)—the same low-frequency oscillations that would presumably underlie the behavioural rate effect (Bosker & Ghitza, 2018; Kösem et al., 2018). Future research could investigate the temporal and neurobiological factors that influence the perceptual binding between prosodic context and target sounds.

This study showed that visual cues to speech rate in a context sentence can influence the perception of audiovisually presented target words (Experiment 2), but not auditory-only target words (Experiment 1). We interpret this outcome to indicate that speech rate may be encoded in a modality-specific manner, at least initially. That is, the visual rate cues (fast vs. slow articulatory movements) presumably influenced the visual cues to the target vowel (here: lip aperture), which at an audiovisual integration stage biased target word perception (cf. Figure 1). Note, however, that the absence of evidence for modality-independent encoding of speech rate (i.e., no evidence for a rate effect in the V-only condition in Experiment 1) should not necessarily be taken as evidence against a supramodal architecture of multisensory speech comprehension in general. In fact, there is considerable neural and behavioural evidence for supramodal perception (Rosenblum, 2019; Rosenblum et al., 2017). Instead, the outcomes of Experiment 1 should be taken as inspiration for further investigation, testing, for instance, the possible conditions under which supramodal influences might be observed after all. Possible avenues could involve varying the delay between context sentence and target words, or varying how auditory cues are weighted relative to visual information by presenting target speech in visual vs. auditory noise.

A relevant question for follow-up research concerns what visual information is actually important for cuing the talker’s speech rate, and indeed the linguistic nature of these visual rate cues. We presented participants with videos of talkers from the shoulders up. Hence, we may speculate that the speed of movement of the articulators (e.g., jaw, lips, tongue) was responsible for the visual rate effect in Experiment 2. However, maybe other non-articulatory visual movements may induce similar rate effects. A potential research avenue, in this respect, could be the role of manual gestures. In natural face-to-face conversation, speakers commonly complement their speech with rapid biphasic (e.g., up and down) movements of the hands, known as beat gestures (McNeill, 1992). We know that these beat gestures tune the processing of speech through phase resetting of ongoing neural oscillations at relevant moments during natural speech comprehension (Biau et al., 2015), enhancing the perceived prominence of the word they accompany (Krahmer & Swerts, 2007). However, their potential role in manually cuing perceived speech rate is currently unknown. Future investigation of how a speaker’s mouth and hands may concurrently and interactively guide what we hear may lead to a better understanding of the multimodal nature of everyday human communication.

Acknowledgements
We would like to thank Annelies van Wijngaarden for the video-recordings, Jeroen Geerts and Iris Schmits for their help in manipulating the video stimuli, Maarten van den Heuvel for help with eye-tracking, and Iris Schmits and Esther de Kerf for testing participants.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Max Planck Society for the Advancement of Science, Munich, Germany (H.R.B., D.P., J.H.)
and the Netherlands Organisation for Scientific Research (D.P.; Veni grant 275-89-037).

**ORCID iD**
Hans Rutger Bosker. [https://orcid.org/0000-0002-2628-7738](https://orcid.org/0000-0002-2628-7738)

**Supplementary material**
The Supplementary Material is available at qjep.sagepub.com.

**References**


the words we hear. *Current Biology*, 28(18), 2867–2875. https://doi.org/10.1016/j.cub.2018.07.023