Synchronization of Speech and Gesture: Evidence for Interaction in Action

Mingyuan Chu[1] and Peter Hagoort[1,2]

1. Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

2. Donders Institute for Brain, Cognition, and Behaviour, Radboud University, 6525 HR Nijmegen, The Netherlands

Author Note

Mingyuan Chu, Neurobiology of Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; Peter Hagoort, Neurobiology of Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, and Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands.

We thank Herbert Clark, Pim Levelt, Pieter Medendorp, and Antje Meyer for providing insightful comments to this study. We thank Albert Russel and Johan Weustink for technical assistance, Livia van de Kraats, Brenda Lelie, and Manuela Schuetze for data collection and Lucy Foulkes for proofreading the manuscript.

Correspondence concerning this article should be addressed to Mingyuan Chu, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands. Email: mingyuan.chu@mpi.nl.

Abstract

Language and action systems are highly interlinked. A critical piece of evidence is that speech and its accompanying gestures are tightly synchronized. Five experiments were conducted to test two hypotheses about the synchronization of speech and gesture. According to the *interactive* view, there is continuous information exchange between the gesture and speech systems, during both their planning and execution phases. According to the *ballistic* view, information exchange occurs only during the planning phases of gesture and speech, but the two systems become independent once their execution has been initiated. In all experiments, participants were required to point to and/or name a light that had just lit up. Virtual reality and motion tracking technologies were used to disrupt their gesture or speech execution. Participants delayed their speech onset when their gesture was disrupted. They did so even when their gesture was disrupted at its late phase, and even when they received only the kinesthetic feedback of their gesture. Also, participants prolonged their gestures when their speech was disrupted. These findings support the *interactive* view and add new constraints on models of speech and gesture production.

Keywords: gesture-speech synchronization, interactive view, pointing gesture, virtual reality, visual feedback, kinesthetic feedback

Synchronization of Speech and Gesture: Evidence for Interaction in Action

Human communication is multimodal. Speakers communicate not only with speech, but also with gestures. Gesturing occurs across ages and cultures (Feyereisen & deLannoy, 1991; Kendon, 2004; Kita, 2009), and even children at the one-word stage combine speech and gesture (Iverson & Goldin-Meadow, 2005). In additiona, the speech and gesture systems are highly interactive (Kendon, 2004; McNeill, 1992). For example, when saying “rotating”, speakers often draw circles in the air with an extended index finger (Chu & Kita, 2008), and the way people describe a motion event affects the gesture they use to depict it (Kita & Özyürek, 2003).

The best evidence for the interaction between these two systems is the tight temporal coordination between them (e.g., Iverson & Thelen, 1999; Krauss, Chen, & Gottesman, 2000; McNeill, 1992). By the rule of phonological synchrony (McNeill, 1992), the gestural stroke (i.e., the forceful part of a gesture) coincides with stressed syllables. This pattern has been observed in descriptions of animated cartoons (Tuite, 1993) and in spontaneous dyadic conversations (McClave, 1998; Loehr, 2007).

The synchronization of speech and gesture requires the two systems to exchange information. There are two views on how and when this synchronization is achieved. The *ballistic* view (Levelt, Richardson, & La Heij, 1985) proposes that synchronization is established through the interaction of the two systems during their planning phases, i.e., while speakers are preparing where they are going to point and what they are going to say. Once the gesture or speech has been initiated (i.e., once the hand starts to move or speech is articulated), they act independently with no further interaction. The *interactive* view proposes that synchronization between speech and gesture is achieved through continuous interaction of the two systems not only during their planning but also during their execution phases. The two views agree that the

two systems are interactive (in fact, on some proposals, speech and gesture may be inseparable; McNeill and Ducan, 2000; McNeill, 2005, 2012) *before* they are executed, but they differ on whether or not the two action systems are still interactive *after* their execution has been initiated; that is, once the joined speech-gesture plan is implemented in the two action systems. The present study aims to contrast the two views by manipulating the gesture or speech system *after* their execution has been initiated and measuring the effect on the other system.

An experimental approach to measure the effect of gesture disruption on speech was pioneered by Levelt et al. (1985). Participants were seated in front of four lights which lit up randomly one by one, and were asked to indicate which of four lights had just lit up by pointing to it and saying “that light”. Their gesture was mechanically disrupted at unpredictable moments. A cord was tied around the participant’s wrist and a 1600 gram weight was attached to the other end of the cord. The weight was applied at either the early or the middle phase of gesture execution. The early and the middle phases were defined individually by the average gesture travelling distance in calibration trials performed before the experiment started. According to the *ballistic* view, disrupting the gesture should not affect speech production because no information exchange is possible after the gesture has been initiated. In contrast, according to the *interactive* view, disrupting the gesture at both the early and the middle phases of gesture execution should affect speech because the two systems continuously interact during their execution phases. In this study participants delayed their speech onset when their gesture was disrupted at its early phase but not at its middle phase. The speech system became ballistic between 300 to 370 milliseconds before speech onset. It was concluded that speech and deictic gesture are interactive only during their planning phases, and that two systems become ballistic almost immediately after the gesture has been initiated. These findings support the *ballistic* view.

Since Levelt's original study, models of speech production have been specified (outlined in more detail) further. There is general agreement that speech production has three major phases. During a conceptualization phase, speakers plan the content of their speech. During a formulation phase, speakers retrieve syntactic information and the phonological forms of individual words from the mental lexicon. During an articulation phase, speakers execute their speech (e.g., Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Levelt, 1989; Levelt, Roelofs, & Meyer, 1999; Rapp & Goldrick, 2000). Based on a meta-analysis of event-related potential (ERP) studies on the time course of speech production (Indefrey & Levelt, 2004), the *formulation* phase appears to start approximately 400 ms before articulation. Thus the timing data from Levelt et al. (1985) suggest that processing within the speech system becomes ballistic at the early formulation phase.

Little is known on how disrupting speech production would affect gesture execution. There is some evidence that disrupting speech could prolong gesture execution time (de Ruiter, 1998). In this study, participants were asked to point to and name pictures that had just lit up. Occasionally (on 28 out of a total of 1556 trials) participants interrupted and repaired their own speech or hesitated between words. When they did so, the duration of their gesture was 117 ms longer in these trials than in non-disrupted trials. This suggests that after gesture and speech have been initiated, the gesture system might be still open to feedback from the speech system.

The findings from Levelt et al. (1985) and de Ruiter (1998) are limited in several aspects. First, the timing for disrupting gesture could not be strictly controlled. To accommodate individual differences in gesture movement, participants were asked to make and hold the gestural movement to each light in calibration trials (Levelt et al., 1985). However, the travelling distance of pointing gestures in the calibration trials might have differed from those in the main

experiments when disrupted and non-disrupted trials were mixed together. Second, neither of the studies measured the effect of speech disruption on gesture execution. Third, neither of the studies could separate the visual and the kinesthetic feedback of gesture. When a gesture was disrupted at the early phase in Levelt et al. (1985), participants could use both visual and kinesthetic feedback to inform their speech system to delay their speech onset. It is unclear whether the interaction between the two systems can rely on kinesthetic feedback alone.

The present study used virtual reality and motion tracking technologies to overcome these limitations. First, we based the moment of gesture disruption on the mean gesture travelling distance from non-disrupted trials in the main experiment rather than in the calibration trials (e.g., Levelt et al., 1985). The mean distance was continuously updated over the course of each block of trials. This allowed us to control the moment of gesture disruption more precisely and make a better estimate of the time window within which the speech and gesture systems interact. Second, we also measured the effect of speech disruption on gesture execution. This allowed us to examine whether the interaction of the speech and gesture systems are bi-directional. Third, we assessed whether speakers could rely on kinesthetic feedback alone to adjust their speech when their gesture was disrupted. This was done in virtual reality by disabling the visual feedback of the gestures.

The task we used closely resembled the one used in Levelt et al. (1985). In virtual reality, the participant was presented with a horizontal line of four lights 20 cm apart. They were asked to point to the light that had just lit up and say either “dit lampje” (“this light”) for one of the two lights in the middle of the light panel or “dat lampje” (“that light”) for the two lights at the leftmost or the rightmost side of the light panel. A sensor was attached to the tip of the participant's right index finger (see Figure 1a). Its movement was tracked by the motion tracking

system and was displayed to the participant as a white ball in virtual reality (see Figure 1b). So the movement of the white ball provided visual feedback about the movement of the participant's right index finger. There was a minimum delay of 117 ms between the movement of participants' gestures and the movement of the white ball in virtual reality, because the motion tracking system needs 50 ms to track movements and the virtual reality system needs 67 ms to display video (see supplementary material on how this 117 ms delay was measured).

Insert Figure 1 here

We carried out five experiments. In Experiment 1, we delayed the movement of the white ball and measured the effect of this delay on speech. If gesture execution is prolonged by the delay of visual feedback, according to the *interactive* view, people should also delay their speech onset to synchronize their speech and gesture. In Experiments 2 and 3, we disrupted gesture execution either by shifting the white ball horizontally (Exp. 2) or by freezing it temporarily (Exp. 3) at the early, middle or late phase of gesture execution on randomly selected trials. If people's gesture execution is prolonged by these disruptions, according to the *interactive* view, people should also delay their speech onset to maintain the synchronization. In Experiment 4, we disabled all visual feedback. We disrupted gesture execution by shifting the position of the target light at the early, middle or late phase of gesture execution. We examined whether people could rely on the kinesthetic feedback of their gesture to adjust their speech onset to maintain the synchronization. In Experiment 5, we asked participants to point to and name the color of the light that had just lit up. We disrupted their speech by changing the color of the target light during the early, middle, or late phase of gesture execution. According to the *interactive* view, when speech is delayed, people should prolong their gesture to maintain the synchronization. For all five experiments, the *ballistic* view should predict null effects, at least when disruption

occurred at the middle or late phase of gesture execution. According to this view, interaction between the speech and gesture systems is no longer possible after the gesture or speech has been initiated.

In all experiments, we measured five dependent variables:

(1) G-init time (Gesture Initiation Time): the time between the illumination of the light and the initiation of the pointing gesture. The initiation of the gesture was defined as the moment when the speed of the gesture exceeded 11.7 cm per second. This was equal to 0.2 cm per refresh of the infrared camera (1/60 seconds).

(2) G-apex time (Gesture Apex Time): the time between the illumination of the light and the moment when the gesturing hand reached its maximal forward extension.

(3) G-exec time (Gesture Execution Time): the duration between G-init and G-apex.

(4) S-onset time (Speech Onset Time): the time between the illumination of the light and the onset of the articulation.

(5) SG-interval time: The time interval between G-apex time and S-onset time.

Experiment 1

The first goal was to replicate Levelt et al. (1985)'s finding that G-apex times were longer when people pointed to the two far lights than when they pointed to the two near lights.

The second goal was to test the assumption that participants treated the movement of the white ball as visual feedback of their own gesture. Participants were told that the movement of the white ball in virtual reality represented the movement of their right index finger. They were also told that its movement might be delayed because the computation speed of virtual reality system was not fast enough. If participants indeed treated the movement of the white ball as the visual feedback of their own gesture, they should prolong their gesture when the movement of

the white ball was delayed. This is because, when their finger reached the target light in reality, the white ball would not yet have reached the target light, so they would need to move their hand further until the white ball reached the target light. We should observe such an effect both when they pointed to and named the target light simultaneously (the gesture-and-speech condition) and when they only pointed to it (the gesture-only condition). However, if they did not treat the white ball as visual feedback of their own gesture, their gesture should not be affected by the delay of the white ball movement. This is because they should stop their gesture when they felt kinesthetically that their finger had reached the target light, and then should wait for the white ball to catch up.

The third goal was to test the *interactive* and the *ballistic* views by examining whether participants would also delay their speech when their gesture was prolonged. The *interactive* view predicts that they should delay their S-onset time when their G-exec time is prolonged. The *ballistic* view predicts that prolonging G-exec time should have no effect on S-onset time.

The final goal was to examine whether participants' speech onset was always synchronized with their gesture apex. If so, the SG-interval time should not be affected by the delay of visual feedback.

Method

Participants. The participants were 17 native Dutch speakers (12 female; mean age = 20, *SD* = 2.67). All were right-handed with normal or corrected-to-normal vision. They were paid for their participation.

Apparatus. The same equipment was used for all five experiments. The participant sat at a table about 10 cm away from the upper body. A start button was located in the centerline of the table, 25 cm away from the light panel and approximately 40 cm away from the participant's

body. Virtual reality was presented through an NVIS nVisor SX stereo Head Mounted Display (HMD). The HMD provides a stereoscopic display with a 44° horizontal (H) and 35° vertical (V) field of view, and a resolution of 1280 (H) × 1024 (V) pixels for each eye. The refresh rate of the HMD was 60 Hz. Images were rendered by a 2.66 GHz Q9400 processor with a NVidia Quadro FX5800 graphics card, using Vizard software (from WorldViz, Santa Barbara, CA). The movement of the pointing finger was tracked in three dimensions by a passive optical position sensing system using DTrack software and ARTtrack3 infrared tracking cameras with a marker attached to the tip of the right index finger. The tracker provided 3 degrees-of-freedom measurements of the sensor position at 60 Hz (within 1mm). Participants' speech was recorded by a wireless Sennheiser microphone attached to the HMD.

Design and Procedure. The basic design of the gesture-and-speech condition was the same for Experiments 1 to 4. A trial started when the participant pressed the start button, and then after a random interval one of the four red lights in virtual reality was lit up for 1000 ms. This interval had a normal distribution with a mean of 1000 ms and a standard deviation of 150 ms. The participant was told to point to the illuminated light and say "dit lampje" ("this light") if it was one of the two lights in the middle of the wooden panel or "dat lampje" ("that light") if it was one of the two lights away from the middle. They were not told anything about synchronizing their gesture and speech. They were asked to respond as accurately and as quickly as possible.

The gesture-only condition had the same design as the gesture-and-speech condition, except that the participant simply pointed to the target light without speaking.

There were six experimental blocks, with 48 trials in each block. Three blocks were in the gesture-and-speech condition and three were in the gesture-only condition. Blocks with the

two conditions alternated. Half of the participants began with the gesture-and-speech condition, and half with the gesture-only condition. Visual feedback in virtual reality was randomly delayed by 117 ms, 217 ms, 317 ms, or 417 ms. Each delay occurred twelve times in a block. The delays were created by presenting the participant's own hand movement in virtual reality (represented by the movement of the white ball) 117 ms, 217 ms, 317 ms or 417 ms after the movement of their hands. We could not implement a 0 delay condition because of the 117 ms minimum delay in the virtual reality system.

There were four practice trials before the first gesture-and-speech block and four more before the first gesture-only block. In these practice trials the delay interval was always 117 ms.

Results and Discussion

We excluded 33 error trials (1.34% of all trials) in gesture-and-speech blocks from the analyses because participants pointed to the wrong light, started pointing before the light turned on, did not point at all, produced the wrong name, did not speak, hesitated or made repairs. We also excluded 19 error trials (0.78% of all trials) in gesture-only blocks because participants pointed to the wrong light, started pointing before the illumination of the light or did not point at all.

We replicated the finding of Levelt et al. (1985) that participants take longer to point to the far lights than to the near lights. We submitted G-apex time in the gesture-and-speech condition to an Analysis of Variance (ANOVA) with delay interval (117 ms, 217 ms, 317 ms and 417 ms), field (left and right), and distance (near and far) as the independent variables. G-apex time was on average 69 ms longer when participants pointed to the far lights than when they pointed to the near lights (F (1, 16) = 29.60, $p < .01$, $\eta_p^2 = .65$; see Figure 2; Standard errors are reported in Supplementary material)[1]. So the pointing gestures in the present study were similar

to those in Levelt et al. (1985)[2]. In the rest of the paper, we combine data from the four lights, because we are mainly interested in the effect of gesture disruption on speech and speech disruption on gesture.

Insert Figure 2 here

Participants did indeed treat the movement of the white ball as visual feedback of their own gesture. We submitted G-init time and G-exec time in the gesture-and-speech condition to two ANOVAs with delay interval as the independent variable (117 ms, 217 ms, 317 ms and 417 ms)[3]. Participants were not able to predict the delay interval of each trial before their gesture was initiated, as there was no main effect of delay interval on G-init time (F (3, 48) = 0.14, p = .94). They prolonged their G-exec time when the movement of the white ball was delayed, as there was a main effect of delay interval on G-exec time (F (3, 48) = 20.31, $p < .01$, $\eta_p^2 = .56$; See Figure 3). A trend analysis showed that the G-exec time increased linearly as the delay interval increased ($p < .01$).

Participants also delayed their speech onset when visual feedback of their gesture was delayed. However, the synchronization of speech and gesture was not affected by the delay interval of visual feedback. We submitted S-onset time and SG-interval time (S-onset time – G-apex time) to two ANOVAs with delay interval as the independent variable. There was a main effect of delay interval on S-onset time (F (3, 48) = 8.48, $p < .01$, $\eta_p^2 = .35$; see Figure 3). A trend analysis showed that the S-onset time increased linearly as the delay interval increased ($p < .01$). There was no main effect of delay intervals on SG-interval time (F (3, 48) = 0.75, p = .47; see Figure 3).

Insert Figure 3 here

Thus when participants received delayed visual feedback, they prolonged their gesture execution time and delayed their speech onset time. These results are consistent with the *interactive* view that speech and gesture can interact with each other after they have been initiated.

In Experiment 1, participants could detect the delay of visual feedback immediately after they initiated their gesture. Would they delay their speech when their gesture was disrupted at its middle or late phases? This question was addressed in Experiments 2, 3 and 4.

Experiment 2

The first goal of Experiment 2 was to examine whether participants would delay speaking when their gesture execution was disrupted at its early, middle, or late phase. Participants were told that the movement of the white ball represented the movement of their right index finger, but we disrupted their gesture by shifting visual feedback (i.e., the white ball in virtual reality) to the left or to the right. We informed participants that sometimes the white ball might shift to the left or to the right because the tracking system was unstable and that when this happened they should still point to and name the illuminated light as quickly and accurately as possible. We expected that shifting the white ball would lead to a prolonged G-exec time because participants would try to "correct" their gesture to point to the target light. The *interactive* view predicts that they should delay their speech regardless of when the disruption occurred, whereas the *ballistic* view predicts that they would not delay their speech.

The second goal was to whether participants' speech onset was always synchronized with their gesture apex. If so, the SG-interval time should not be affected by the disruption of visual feedback and by whether the disruption occurred at the early, middle, or late phase of

gesture execution. Furthermore, the more their gesture apex was delayed, the more their speech onset should be delayed.

Method

Participants. The participants were 17 native Dutch speakers. All were right-handed with normal or corrected-to-normal vision. They were paid for their participation. We excluded one participant because she always started speaking after she had completed the retraction of her gesture (i.e., after returning her hand to the table). The final sample consisted of 16 participants (13 female) with an average age of 21 years (*SD* = 1.88).

Design and Procedure. The basic design was identical to the gesture-and-speech condition in Experiment 1. There were six experimental blocks, with 40 trials in each block. We did not disrupted visual feedback in the first eight trials of each block. Based on these trials, we calculated the mean straight line distance between the gesture initiation position and gesture apex position for each light. We defined the *early, middle, and late* phases as 25%, 50%, and 75% of the straight line distance between these two positions. The remaining 32 trials of each block consisted of 20 non-shifted trials and 12 ball-shifted trials that were randomly intermixed. We added the gesture distances of the 20 non-shifted trials into the calculation of the early, middle and late phases, so that these figures were continuously updated. In the 12 ball-shifted trials, the white ball was randomly shifted 5 cm horizontally left or right (parallel to the four lights) at the early, middle or late phase of gesture execution. Shifting at each phase occurred randomly four times in each block, and each time occurred on a different light. There were four non-shifted practice trials before the first block. In all trials the visual feedback of gesture in virtual reality was delayed 117 ms from the actual gesture due to system computation time.

An example video is provided in Supplementary Material illustrating the trajectory of gesture execution in a late ball-shifted trial on the far left light.

Results and Discussion

We excluded 48 error trials (1.25% of all trials) from the analyses by the same criteria we used for the gesture-and-speech condition of Experiment 1. We excluded an additional 227 trials (5.91% of all trials) where the white ball was shifted after participants' gesture reached apex or after participants started speaking, because in these trials the white ball shift could no longer affect gesture execution time or speech onset time. The time interval between the light illumination and the shift of white ball was on average 529 ms (*SD* = 49 ms) in the early ball-shifted trials, 623 ms (*SD* = 65 ms) in the middle ball-shifted trials, and 730 ms (*SD* = 83 ms) in the late ball-shifted trials. We submitted G-init time, G-exec time, S-onset time and SG-interval time to four ANOVAs with trial type as the independent variable (non-shifted, early, middle and late ball-shifted trials).

Participants were not able to predict the type of each trial before they initiated their gesture, as there was no main effect of trial type on G-init time (F (3, 45) = 1.75, p = .17). They prolonged their G-exec time when visual feedback was shifted regardless whether it occurred at the early, middle or late phase of gesture execution. There was a main effect of trial type on G-exec time (F (3, 45) = 5.26, $p < .01$, $\eta_p^2 = .26$; see Figure 4). Bonferroni post hoc tests showed that G-exec time was longer in all ball-shifted trials than in the non-shifted trials (ps < .05).

Participants also delayed their speech when visual feedback was shifted, regardless of whether it occurred at the early, middle, or late phase of gesture execution. There was a main effect of trial type on S-onset time (F (3, 45) = 20.51, $p < .01$, $\eta_p^2 = .58$; see Figure 4).

Bonferroni post hoc tests showed that S-onset time was longer in all ball-shifted trials than in the non-shifted trials (*p*s < .05).

We then calculated the minimum time interval needed for interaction between the speech and gesture systems. In the late ball-shifted trials, visual feedback was disrupted on average 730 ms after the light was lit up, and participants' planned S-onset time was on average 829 ms after the light lit up. Thus, the speech system was still open to feedback from the gesture system approximately 99 ms before the estimated speech onset. This is substantially shorter than the 300 ms to 370 ms window estimated by Levelt et al. (1985). Note that it is impossible to know participants' planned S-onset time in the ball-shifted trials, so we used S-onset time in the non-shifted trials as an estimate. This should be an unbiased estimate because participants could not predict whether or not the white ball would be shifted in a given trial.

The synchronization of gesture and speech was not affected by the shift of visual feedback or by whether it occurred at the early, middle, or late phase of gesture execution. There was no main effect of trial type on SG-interval time (F (3, 45) = 2.71, p = .06, η_p^2 = .15; see Figure 4). Bonferroni post hoc tests showed that all pairwise comparisons failed to reach significance (*p*s > .23).

Insert Figure 4 here

We then examined whether the delay of gesture apex was positively correlated with the delay of speech onset. We first calculated the average G-apex time and S-onset time from the non-shifted trials of each participant. For each participant, we then subtracted the G-apex time in each ball-shifted trial from the average G-apex time in the non-shifted trials. We also subtracted the S-onset time in each ball-shifted trial from the average S-onset time in the non-shifted trials. Finally, we pooled all difference scores across all participants and computed three correlations,

namely, for the early, the middle and the late ball-shifted trials. The G-apex time difference was positively correlated with the S-onset time difference in all three types of ball-shifted trials (early: r (362) = .33, p < .01; middle: r (318) = .27, p < .01; late: r (228) = .51, p < .01; see Figure 5 for the scatter plots of the correlations). The results suggest that the more a gesture apex was delayed, the more the speech onset was delayed.

Insert Figure 5 here

In Experiment 2, visual feedback was disrupted in a quite salient way. Would people still delay their speech when visual feedback was disrupted in a less salient way? This question was addressed in Experiment 3.

Experiment 3

The goal was to replicate the findings of Experiment 2 with a less salient disruption of visual feedback. To do so, we temporarily froze visual feedback at the early, middle or late phase of gesture execution. Participants were told, as before, that the movement of the white ball represented the movement of their own right index finger. They were also told that sometimes the white ball might freeze for a short period of time because the computation power of the virtual reality system was not strong enough to support continuous movement of an object in virtual reality, and that they should ignore this.

Method

Participants. The participants were 17 native Dutch speakers. All were right-handed with normal or corrected-to-normal vision. They were paid for their participation. We excluded one participant because she always started speaking after she had completed the retraction of her gesture. The final sample consisted of 16 participants (13 female) with an average age of 21 years (SD = 3.18).

Design and Procedure. The procedure was the same as in Experiment 2, except that the ball-shifted trials in Experiment 2 were replaced by the ball-frozen trials in which the white ball was temporarily frozen for 200 ms and then jumped forward in space to be synchronized with resumed synchronization with the participant's right index finger. In all trials the visual feedback of gesture in virtual reality was delayed 117 ms from the actual gesture due to system computation time.

An example video is provided in Supplementary Material illustrating the trajectory of gesture execution in a late ball-frozen trial on the far left light.

Results and Discussion

We excluded 64 error trials (1.67% of all trials) from the analyses by the same criteria we used for the gesture-and-speech condition in Experiments 1 and 2. We excluded an additional 250 trials (6.51% of all trials) where the white ball was frozen after participants' gesture reached apex or after participants started speaking, because in these trials freezing the white ball could no longer affect gesture execution time or speech onset time. The time interval between the light illumination and the freeze of the white ball was on average 583 ms (*SD* = 107 ms) in the early ball-shifted trials, 671 ms (*SD* = 122 ms) in the middle ball-shifted trials, and 757 ms (*SD* = 143 ms) in the late ball-shifted trials. We submitted G-init time, G-exec time, S-onset time and SG-interval time to four ANOVAs with trial type as the independent variable (non-frozen; early, middle and late ball-frozen trials).

We essentially replicated all findings in Experiment 2. Participants were not able to predict the type of each trial before they initiated their gesture, as there was no main effect of trial type on G-init time (F (3, 42) = 0.02, p = .99). They prolonged their G-exec time when visual feedback was frozen at the late phase of gesture execution. There was a main effect of trial

type on G-exec time (F (3, 45) = 5.44, $p < .01$, $\eta_p^2 = .28$; see Figure 6). Bonferroni post hoc tests showed that G-exec time was longer in the late frozen trials than in the non-frozen trials ($p < .05$). G-exec time of early ($p = .10$) and middle ($p = .07$) frozen trials was not significantly longer than G-exec time of non-frozen trial. This might be because in the late ball-frozen trials the white ball was frozen on average from 757 ms to 975 ms after the light illumination. Participants' gesture reached the planned apex (calculated from the non-frozen trials) on average 892 ms after the light illumination. This means that there was no valid visual feedback 135 ms before their gesture reached the planned apex. The absence of visual feedback significantly slowed down gesture execution. When the white ball was frozen in the early or the middle ball-frozen trials, the ball resumed moving before the gesture reached the planned apex. Therefore, freezing the white had a smaller impact on gesture execution on the early and middle ball-frozen trials than on the late ball-frozen trials.

Participants delayed their speech when visual feedback was frozen regardless of whether it occurred at the early, middle, or late phase of gesture execution. There was a main effect of trial type on S-onset time (F (3, 45) = 10.62, $p < .01$; $\eta_p^2 = .43$; see Figure 6). Bonferroni post hoc tests showed that S-onset time was longer in all ball-frozen trials than in the non-frozen trials (ps < .05).

In the late ball-frozen trials, visual feedback was disrupted on average 757 ms after the light was lit up, and participants' planned S-onset time was on average 864 ms after the light was lit up. Thus, the speech system was still open to feedback from the gesture system at around 107 ms before the estimated speech onset.

The synchronization of gesture and speech was not affected by the frozen of visual feedback or by whether it occurred at the early, middle, or late phase of gesture execution. There

was no main effect of trial type on SG-interval time (F (3, 45) = 2.72, p = .09, η_p^2 = .16; see Figure 6). Bonferroni post hoc tests showed that all pairwise comparisons failed to reach significance (ps > .12).

Insert Figure 6 about here

Finally, we observed, again, a positive correlation between delays in gesture apex and speech onset times. The G-apex time difference was positively correlated with the S-onset time difference in all three types of ball-frozen trials (early: r (349) = .57, p < .01; middle: r (310) = .43, p < .01; late: r (230) = .47, p < .01; see Figure 7 for the scatter plots of the correlations).

Insert Figure 7 about here

In all three experiments reported earlier, participants could not only see the disruption of their gesture (visual feedback) but also felt the kinesthetic change of their gesture (kinesthetic feedback). Could people use kinesthetic feedback alone to inform their speech system about the disruption of their gesture and delay their speech accordingly? This question was addressed in Experiment 4.

Furthermore, in all three experiments reported earlier, we manipulated gesture execution by delaying or disrupting visual feedback of gesture in virtual reality. One might argue that gesture and speech were affected independently by surprising visual inputs (i.e., the delayed or the disrupted visual feedback of gesture in virtual reality), and there was no interaction between the gesture and speech systems after both gesture and speech had been initiated. This alternative explanation is unlikely because the target utterance (“this light” or “that light”) remained unchanged when visual feedback of gesture was delayed or disrupted. In addition, Experiment 2 and Experiment 3 showed that the more participants’ gesture apex was delayed, the more their speech onset was delayed. However, to completely rule out this alternative explanation, one

needs to include an additional speech-only condition, where participants name the target light without pointing to it, and to show that the surprising visual input does not affect speech onset time. Experiment 4 addressed this issue.

Experiment 4

The goal was to examine whether people could rely on kinesthetic feedback alone to delay their speech when their gesture was disrupted at different phases. We did not provide any visual feedback in this experiment. To disrupt gesture execution, we shifted the target light randomly to the left or to the right horizontally by 5 cm at the early, middle, or late phase of gesture execution. Participants were told that sometimes the position of the target light might be shifted to the left or to the right, and they needed to point to its new position. We expected that shifting the position of the target light should prolong participants' gesture, because they would need to correct their gesture to point to the new light position. If people are able to rely on kinesthetic feedback alone to inform their speech production system to maintain gesture-speech synchronization, participants should delay their speech when their gesture was disrupted and when there was no visual feedback. Also, the synchronization of gesture and speech should not be affected by the disruption of gesture or by whether it occurred at the early, middle or late phase of gesture execution.

We also included a speech-only condition in which participants named the light without pointing. This allowed us to assess the effect of the target light shifting on S-onset time independent of any effect from the disruption of gesture execution. As the name of the target light did not change when it was shifted, people should not delay their speech in these trials. Strictly speaking, however, there was no pure speech-only condition because people would direct

their gaze or head towards the target light, which could be seen as a form of pointing with the eyes or head (Levelt et al., 1985).

Method

Participants. The participants were 25 native Dutch speakers. All were right-handed with normal or corrected-to-normal vision. They were paid for their participation. We excluded 7 participants due to a programming error. The final sample consisted of 18 participants (14 female) with an average age of 21 years (*SD* = 2.35).

Design and Procedure. The procedure for the gesture-and-speech condition was identical to that used in Experiment 2, except that there was no visual feedback in all trials, and the ball-shifted trials in Experiment 2 were replaced by the light-shifted trials in which the target light was shifted 5 cm horizontally to the left or to the right at the early, middle or late phase of gesture execution. The light shift took 300 ms.

The procedure for the speech-only condition was the same as the one used in the gesture-and-speech condition, except that the participant only said “dit lampje” or “dat lampje”, without pointing, and the three types of light-shifted trials (early, middle and late light-shifted trials) were based on the pre-articulation period (i.e., the period between the moment of the light illumination and the speech onset) from non-disrupted trials.

Two example videos are provided in Supplementary Material illustrating the light shifting in the virtual environment and a participant’s gesture in the real environment; in both cases for a late light-shifted trial on the far left light.

In this experiment, there were six experimental blocks, with 40 trials in each block. Three blocks were in the gesture-and-speech condition and three were in the speech-only condition. Blocks with the two conditions alternated. Half of the participants began with the gesture-and-

speech condition, and half with the speech-only condition. There were four practice trials before the first gesture-and-speech block and four more before the first speech-only block. In these trials the target lights were not shifted.

Results and Discussion

In the gesture-and-speech blocks, we excluded 51 error trials (2.36% of all trials) from the analyses by the same criteria we used for the gesture-and-speech condition of Experiment 1. We excluded an additional 134 trials (6.20% of all trials) where the target light was shifted after participants' gesture reached apex or after participants started speaking, because in these trials the target light shift could no longer affect gesture execution time or speech onset time. In the speech-only blocks, we excluded 67 error trials (3.10% of all trials) from the analyses because participants produced a wrong name, because their speech was hesitant or repaired, or because they failed to produce a response at all. In the gesture-and-speech blocks, the interval between the light illumination and the shift of the target light was on average 583 ms (*SD* = 77 ms) in the early light-shifted trials, 652 ms (*SD* = 59 ms) in the middle light-shifted trials, and 760 ms (*SD* = 67 ms) in the late light-shifted trials. In the speech-only blocks, the interval between the light illumination and the shift of the target light was on average 226 ms (*SD* = 27 ms) in the early light-shifted trials, 379 ms (*SD* = 51 ms) in the middle light-shifted trials, and 530 ms (*SD* = 80 ms) in the late light-shifted trials.

To examine whether people could rely on kinesthetic feedback alone to synchronize their speech and gesture, we first analyzed the gesture-and-speech blocks. We submitted G-init time, G-exec time, S-onset time and SG-interval time to four ANOVAs with trial type as the independent variable (non-shifted, early, middle and late light-shifted). To examine whether shifting the target light had an effect on S-onset time when participants did not produce any

gesture, we then analyzed the speech-only blocks. We submitted the S-onset time to an ANOVA with trial type as the independent variable (non-shifted, early, middle and late light-shifted trials).

In the gesture-and-speech blocks, participants were not able to predict the type of each trial before they initiated their gesture, as there was no main effect of trial type on G-init time (F (3, 51) = 2.12, p = .11). They prolonged their G-exec time when the target light was shifted regardless whether it occurred at the early, middle or late phase of gesture execution. There was a main effect of trial type on G-exec time (F (3, 51) = 15.72, $p < .01$, $\eta_p^2 = .48$; see Figure 8). Bonferroni post hoc tests showed that G-exec time was longer in all light-shifted trials than in the non-shifted trials (ps < .01).

Participants delayed their speech when the target light was shifted at the middle or the late phase of gesture execution. There was a main effect of trial type on S-onset time (F (3, 51) = 11.44, $p < .01$, $\eta_p^2 = .40$; see Figure 8). Bonferroni post hoc tests showed that S-onset time was significantly longer in the middle and the late light-shifted trials than in the non-shifted trials (ps < .05).

The synchronization of speech and gesture was not significantly affected by the shift of the target light or by the moment when the target light was shifted. Although there was a main effect of trial type on SG-interval time (F (3, 51) = 3.83, $p < .05$, $\eta_p^2 = .18$; See Figure 8), Bonferroni post hoc tests showed that none of the pairwise comparisons reached significance (ps > .11).

Insert Figure 8 about here

In the speech-only blocks, participants delayed their speech when the target light was shifted at the early, but not at the middle or the late phase, of gesture execution. There was a main effect of trial types on S-onset time (F (3, 51) = 14.39, $p < .01$, $\eta_p^2 = .46$; see Figure 9).

Bonferroni post hoc tests showed that S-onset time was longer in the early light-shifted trials than in the non-shifted trials ($p < .05$). We did not expect S-onset time to differ between the light-shifted and the non-shifted trials. It is possible that participants had not finished directing their eyes or head towards the target light when the target light was shifted at the early phase, and they redirected their eye gaze or head movement to the new position. This might lead to the delay of speech onset. However, at the middle or late phase, participants had presumably completed directing their eyes and head towards the original position of the target light and were less likely to expect the light position to be shifted, and therefore S-onset time was not delayed in these trials.

Insert Figure 9 about here

Most importantly, in the middle and the late light-shifted trials of the gesture-and-speech blocks, participants delayed their speech because they prolonged their gestures but not because the light was shifted. These results ruled out the possibility that gesture and speech was affected independently by a surprising visual input (i.e., the shift of the target light), and showed that the delayed speech onset was caused by prolonged gesture execution. One might argue that shifting the light in the speech-only condition might have occurred too late to affect speech, but this was not the case. In the speech-only condition, the light-shifting occurred on average 317 ms (in the middle light-shifted trials) and 165 ms (in the late light-shifted trials) before the planned speech onset (calculated from the non-shifted trials). In the gesture-and-speech condition, the light-shifting occurred on average 219 ms (in the middle light-shifted trials) and 111 ms (in the late light-shifted trials) before the planned speech onset (calculated from the non-shifted trials).

In the experiments reported so far, when gesture execution was disrupted, people delayed speaking to synchronize their gesture and speech. They did so even when gesture was disrupted

at the late phase and when there was no visual feedback. The synchronization was never significantly affected by the disruption of gesture execution, by the phase at which disruption occurred, or by the absence of visual feedback. Would people prolong their gesture when their speech was disrupted after they had initiated their gesture? This question was addressed in Experiment 5.

Experiment 5

The first goal of Experiment 5 was to examine whether people prolonged their gesture when their speech was disrupted at either the early, middle or late phase of gesture execution. The procedure was slightly different from Experiments 1 to 4. Participants were asked to point to the target light and name the light with its color ("dit blauwe lampje" or "dit gele lampje"; in English: "this blue light" or "this yellow light")[4]. To disrupt speech, we changed the color of the target light at the early, middle or late phase of gesture execution. Participants were told that sometimes the color of the target light might change, and when this happened, they should name the new color. To prevent them from deliberately delaying their speech until the color of the light had been changed, they were asked to point to and name the light as quickly and accurately as possible, and not to delay their speech. According to the interactive view, they should prolong their gesture when their speech was disrupted, whereas according to the *ballistic* view, they should not prolong their gesture.

We also included a gesture-only condition in which participants pointed to the target light without naming it. This allowed us to assess the effect of changing the light color on gesture execution independent of any effect from disrupting the speech. As the location of the target light did not change when its color was changed, people's gesture execution should not be affected.

Method

Participants. The participants were 25 native Dutch speakers. All were right-handed with normal or corrected-to-normal vision. They were paid for their participation. We excluded 7 participants: one failed to complete the experiment due to sickness induced by virtual reality; one only lifted her index finger without pointing clearly to the target light; five produced the wrong color name (and did not correct themselves) in at least 1/3 of the color-changed trials. The final sample consisted of 18 participants (15 female) with an average age of 21 years (*SD* = 2.63).

Design and Procedure. The participant sat at a table. Each trial started when the participant pressed the start button, and then after a random interval one of the four lights turned blue or yellow randomly for 1300 ms. This interval had a normal distribution with a mean of 1000 ms and a standard deviation of 150 ms.

There were six experimental blocks, with 40 trials in each block. Three blocks were in the gesture-and-speech condition and three were in the gesture-only condition. Blocks with the two conditions alternated. Half of the participants began with the gesture-and-speech condition, and half with the gesture-only condition. There were eight practice trials before the first gesture-and-speech block and eight more before the first gesture-only block. We did not change the color of the target lights in these trials.

In the gesture-and-speech blocks, participants were asked to point to the illuminated light and say “dit blauwe lampje” or “dit gele lampje” (“this blue light” or “this yellow light”). In this experiment, visual feedback was provided in virtual reality. The early, middle and late phases of gesture execution were calculated by the same method described in Experiment 3.

The remaining 32 trials of each block consisted of 20 non-color-changed trials and 12 color-changed trials that were randomly mixed. In the 12 color-changed trials, the color of the target light was randomly changed from blue to yellow or vice versa at either the early, middle or

late phase of gesture execution. The light color change at each phase occurred randomly four times in each block, and each time occurred on a different light.

An example video is provided in Supplementary Material illustrating the light color change and the trajectory of gesture in a late color-changed trial on the far left light.

The procedure of the gesture-only condition was the same as the one used in gesture-and-speech condition, except that the participant simply pointed to the target light without naming it.

In all trials the visual feedback of gesture in virtual reality was delayed by 117 ms from the actual gesture due to system computation time.

Results and Discussion

In the gesture-and-speech blocks, we excluded 41 error trials (1.90% of all trials) from the analyses by the same criteria we used for the gesture-and-speech condition of Experiment 1. We excluded an additional 18 trials (0.83% of all trials) where the color of the target light changed after the gesture had reached its apex because in these trials changing the target light color could no longer affect gesture execution time. In the gesture-only blocks, we excluded 5 error trials (0.23% of all trials) by the same criteria we used for the gesture-only condition of Experiment 1. In the gesture-and-speech blocks, the interval between the light illumination and the change of light color was on average 568 ms (*SD* = 54 ms) in the early color-changed trials, 647 ms (*SD* = 68 ms) in the middle color-changed trials, and 757 ms (*SD* = 85 ms) in the late color-changed trials. In the gesture-only blocks, the interval between the light illumination and the change of light color was on average 573 ms (*SD* = 53 ms) in the early color-changed trials, 666 ms (*SD* = 74 ms) in the middle color-changed trials, and 787 ms (*SD* = 80 ms) in the late color-changed trials.

To examine whether people prolong their gesture when their speech is disrupted, we first analyzed the gesture-and-speech blocks. We first submitted S-onset time of the correct color name[5], G-init time, G-exec time and SG-interval time to four ANOVAs with trial type as the independent variable (non-color-changed trials, early, middle, and late color-changed trials). To examine the effect of light color change on G-exec time independent of any influence from the disruption of the speech system, we then analyzed the gesture-only trials. We submitted G-init time and G-exec time to two ANOVAs with trial type as the independent variable.

In the gesture-and-speech blocks, participants started producing the correct color word later when the color of the target light was changed, as there was a main effect of trial type on S-onset time of the correct color word (F (3, 51) = 76.88, $p < .01$, $\eta_p^2 = .82$; see Figure 10). Bonferroni post hoc tests showed that S-onset time was longer in all color-changed trials than in the non-color-changed trials (ps < .01). Participants were not able to predict the type of each trial before they initiated their gesture, as there was no main effect of trial type on G-init time (F (3, 51) = 1.47, $p = .23$).

Participants prolonged their G-exec time when their speech was disrupted at both the early and the late phases of gesture execution. There was a main effect of trial type on G-exec time (F (3, 51) = 5.04, $p < .01$, $\eta_p^2 = .23$; See Figure 10). Bonferroni post hoc tests showed that G-exec time was longer in the early and the late color-changed trials than in the non-color-changed trials (ps < .05).

The synchronization of speech and gesture was affected by the disruption of speech. Participants did not prolong their G-exec time long enough to synchronize gesture apex with the onset of the correct color word. There was a main effect of trial type on SG-interval time (F (3, 51) = 60.44, $p < .01$, $\eta_p^2 = .78$; See Figure 10). Bonferroni post hoc tests showed that SG-

interval time was longer in all color-changed trials than in the non-color-changed trials (*p*s < .01). In addition, SG-interval time was longer in the late color-changed trials than in the early and the middle color-changed trials (*p*s < .05).

Insert Figure 10 about here

In some color-changed trials, participants started their speech with the wrong color name and then repaired it (e.g., “this blue- yellow light” when the color changed from blue to yellow). Among these trials, we selected those in which participants started articulating the word "dit" ("this") before gesture apex (423 trials in total). If participants prolonged their gesture in these trials compared to in non-color changed trials, it would provide the strongest support for the *interactive* view because the gesture and speech systems could still exchange information when both of them were at their execution phases. We compared the G-exec and the G-apex times in these self-repaired color-changed trials with those in the non-color-changed trials. We collapsed the data from the early, middle and late color-changed trials to increase statistical power. Participants prolonged their gesture when speech execution was disrupted during the gesture execution phase. S-onset time of the correct color name, G-execution time, and G-apex time were longer in these self-repaired color-changed trials than in the non-color-changed trials (S-onset time: t (17) = 9.45, p < .01, d = 1.14; G-exec time: t (17) = 2.90, p < .05, d = .82; G-apex time: t (17) = 2.58, p < .05, d = .64; see Figure 11).

Insert Figure 11 about here

In the gesture-only blocks, participants were not able to predict the type of each trial before they initiated their gesture, as there was no main effect of trial type on G-init time (F (3, 51) = .94, p = .43). They also did not prolong their G-exec time when the color of the target light changed. Although there was a main effect of trial type on G-exec time (F (3, 51) = 5.13, p < .01,

η_p^2 = .23; See Figure 12), Bonferroni post hoc tests revealed that this main effect was driven by the fact that G-exec time of middle color-changed trials were significantly longer than G-exec time of late color-changed trials ($p < .05$). G-exec time of non-color-changed trials was not different from any color-changed trials ($ps > .16$)[5]. These results showed that prolonged gesture execution was caused by speech disruption.

Insert Figure 12 about here

General Discussion

We contrasted two competing views on how people synchronize their speech and gesture. According to the *ballistic* view (Levelt et al., 1985), people establish the synchronization *before* the execution of gesture and speech, and once they have initiated their gesture or speech the two systems cannot interact with each other. According to the *interactive* view, people achieve synchronization through information exchange between the two systems both *before* and *after* they initiate their gesture or speech. To test these two alternative views, we disrupted either gesture or speech after gesture or speech had been initiated, and we measured the effect of one type of disruption on the other action type.

Effect of Gesture Disruption on Speech Production

Experiments 1 to 3 showed that people delayed their speech when visual feedback of their gestures was disrupted. Synchronization was not significantly affected by the disruption of gesture execution in any of these experiments. Experiments 1 showed that people started speaking later when the visual feedback of their gesture was delayed. The longer the feedback was delayed, the longer speech was delayed. Experiments 2 and Experiment 3 showed that people delayed their speech when visual feedback of their gesture was disrupted. They did so even when the disruption came at the late phase of gesture execution. Experiment 4 showed that

gesture disruption could delay speech onset even when visual feedback was not available. Based on these results, the speech planning and formulation phases are open to feedback from the gesture system during all phases of gesture execution. These results support the *interactive* view that the speech and gesture systems can exchange information even *after* gesture has been initiated.

Our results are consistent with evidence that people adapt their speech to the time course of their gesture to synchronize gesture and speech production. For example, people start their speech later when they point to and name a far object than when they point to and name a near object, but not when they only name the objects (de Ruiter, 1998; Levelt et al., 1985). The present study extends previous findings by showing that: (1) People can delay their speech even when their gesture is disrupted at its late phase, which was on average around 100 ms before they started speaking in these experiments. This is well beyond the 300 to 370 milliseconds window before speech onset reported by Levelt et al. (1985). According to the speech production literature, the formulation phase (i.e., the phase in which speakers retrieve syntactic information and phonological forms of words from the mental lexicon) starts around 400 ms before articulation (Hagoort & van Turennout, 1997; Indefrey & Levelt, 2004). So the speech system can adapt to the gesture system even during the late phase of the formulation process. (2) People can use kinesthetic feedback alone to inform the speech system about the disruption of gesture and can delay their speech accordingly. (3) The synchronization of speech and gesture, as indicated by the time interval between gesture apex and speech onset, was hardly affected by a disruption of gestures and by whether the disruption occurred at the early, middle, or late phase of gesture execution. The more a gesture apex was delayed, the more the speech onset was delayed.

Why could participants in the present studies still delay their speech when their gesture was disrupted at the late phase, whereas in Levelt et al. (1985) participants could not delay their speech anymore when their gesture was disrupted at the middle phase? In Levelt et al. (1985), gestures were disrupted by applying a mass to the wrist of the gesture hand, whereas in our study gestures were disrupted either by shifting or freezing visual feedback (Exp 2 & 3) or by shifting the target light (Exp 4). Presumably when people's gesture is disrupted, they must generate a new motor plan, and then synchronize their speech with the adapted motor plan. When their gestures were disrupted by an external force (as in Levelt et al., 1985), they would need to reset all the usual kinesthetic parameters and compute the amount of force needed to compensate for the impedance from the unexpected external force. This process presumably took time, especially because people usually rely on visual feedback more than on kinesthetic feedback in estimating their hand movement (Welch & Warren, 1986). When gestures were disrupted in the middle phase in Levelt et al. (1985), it might have been too late for the motor system to generate a new plan to influence the speech system before articulation started. In the present study, participants' kinesthetic feedback was never disrupted by an external force. They could quickly generate a new motor plan based on the new position of the white ball or the target light. So under these more favorable conditions, there was still enough time to generate a new motor plan and influence the speech system.

Although our findings support the view that the gesture and speech systems are still interactive after gesture and speech have been initiated, there exists an alternative explanation for the fact that people delayed their speech when their gesture execution was disrupted. That is, speech production and gesture execution may compete for resources, so disrupting one modality may increases the processing load in that modality and consequently result in slower processing

in the other modality as well (the *competition* hypothesis). According to this hypothesis, participants in our studies may not have delayed their speech onset to synchronize it with the delayed gesture apex, but they may simply have slowed down speech production while they calculated the new motor plan after their gesture execution was disrupted. However, we find this alternative explanation unlikely. The design of Experiment 1 and 5 allowed us to test this *competition* hypothesis directly. In both experiments, we had a gesture-and-speech condition and a gesture-only condition. If there was competition between the two systems during the gesture execution phase, one should expect that in the normal non-disrupted trials the gesture execution time should be longer in the gesture-and-speech condition than in the gesture-only condition. However, we found the opposite pattern. In the non-disrupted trials, the gesture execution time was shorter in the gesture-and-speech condition (597 and 690 ms for Experiment 1 and 5, respectively) than in the gesture-only condition (678 and 733 ms for Experiment 1 and 5, respectively; *p*s < .05). These results are consistent with the hypothesis that gesture production can be facilitated by speech production (Feyereisen, 1997), but they do not support the *competition* hypothesis.

Effect of Speech Disruption on Gesture Production

Experiment 5 showed that people prolonged their gesture when their speech was disrupted. Participants did so even when their speech was disrupted after they have initiated their articulation. These results suggest that there is a speech-to-gesture influence after both gesture and speech have been initiated. Our results are consistent with previous evidence showing that people adapt their gesture to speech. For example, de Ruiter (1998) showed that gesture execution time was longer when the location of the stressed syllable occurred later in speech than when it occurred earlier. Participants also prolonged their gesture when they made a speech error

(e.g., when repairing their speech or hesitating between words) than when they did not. In our experiments, people prolonged their gesture when their speech was disrupted at all phases of gesture execution and after articulation has been initiated.

Furthermore, although disrupting gesture had no significant effect on the synchronization of gesture and speech, disrupting speech did affect the synchronization. Experiment 5 showed that the interval between gesture apex and the onset of the correct color word was larger in the color-changed trials than in the non-color-changed trials. The later the speech was disrupted, the larger the interval was. In deictic expressions, gesture tends to start earlier than speech. So when gesture is disrupted, there is still enough time for the speech system to adapt to the gesture system to reach full synchronization. In contrast, although disrupting speech had some effect on gesture, the gesture system might not have enough time to calculate how long a gesture should be prolonged to reach full synchronization with speech, especially when the disruption occurred at the late phase of gesture execution.

Implications for Speech and Gesture Production Models

Our study has implications for the computational and psycholinguistic models of speech and gesture production. Recent computational models of speech production (e.g., Hickok, Houde & Rong, 2011) and action production (e.g., Franklin & Wolpert, 2011) propose that when producing speech or performing a hand action, the motor control system generates both a motor command to the motor execution system and a corollary discharge to an internal model, which estimates the sensory consequences of a motor command. After speech or action has been initiated, the internal model is used to compare the predicted sensory consequences with the actual sensory consequences of the motor command. A mismatch between them will generate an error signal that can be used to update the internal model and to provide corrective feedback to

the motor control system. We showed that there exists a bidirectional link between the speech and action production systems. When the internal model of one system detects a mismatch between the predicted and the actual sensory feedback, it can inform the other system to generate a new motor plan.

Psycholinguistic models of speech production typically assume that speech production involves a *conceptualizing* phase (i.e., the speech planning phase), a *formulating* phase (the syntactic and phonological retrieval phase), and an *articulation* phase (the speech execution phase; e.g., Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Levelt, 1989; Levelt, Roelofs, & Meyer, 1999; Rapp & Goldrick, 2000). Psycholinguistic models of gesture production typically assume that gesture production involves two stages: a *motor planning* phase for generating motor programmes and a *motor execution* phase for gesture execution (e.g., de Ruiter, 1998; Kita & Ozyurek, 2003; Krauss, Chen, & Chawla, 1996). It has been proposed that interaction can occur between the planning phases of both systems (de Ruiter, 1998; Kita & Ozyurek, 2003)[6] or between the planning phase of gesture production and the formulating phase of speech production (Krauss, Chen, & Chawla, 1996). Our results indicate that there is also interaction between the motor execution phase of gesture production and the formulating and articulation phases of speech production. When gesture or speech was disrupted, execution of the other modality was affected. Figure 13 depicts a model that illustrates our findings. Arrow 1 shows the finding from previous studies that the gesture and speech systems interact during their planning phases (De Ruiter, 1998; Feyereisen, 1997; Levelt et al., 1985). Arrow 2 shows our finding that when gesture execution is disrupted, the sensory discrepancy can feed into the speech planning or the formulation phase and delay speech onset (Experiment 1 to 4)[6]. Arrow 3 shows our finding that when speech production is disrupted, the discrepancy between the predicted speech

output and the actual speech output detected before or during speech execution can feed into the gesture execution phase and prolong gesture execution (Experiment 5).

Insert Figure 13 here

Some gesture production theories propose that gesture and speech are inseparable (e.g., McNeill and Ducan, 2000; McNeill, 2005, 2012). Although our current data are consistent with this view and it is possible that speech and gesture originate from the same conceptualization process, there is ample evidence that the motor preparation and execution of gesture and speech are controlled by distinct brain networks (e.g., Gazzaniga, Ivry, & Mangun, 2008; Kandel, Schwartz, Jessell, Siegelbaum & Hudspeth, 2013). As the focus of the current study is the interaction between gesture and speech during their execution phases, we propose a model featuring gesture and speech as two independent but highly interactive systems.

Conclusion

We used virtual reality and motion tracking technologies to investigate the mechanism underlying the synchronization of speech and gesture. When gesture was disrupted, people delayed their speech to synchronize their gesture and speech. When speech was disrupted, people prolonged their gesture. Thus the two systems appear to exchange information even after both gesture and speech have been initiated, supporting the *interactive* view that the synchronization is achieved through continuous interaction between the two systems both *before* and *after* they have been initiated.

Our study focused on the synchronization between two action modalities, namely gesture and speech. Although we studied deictic gestures, tight temporal coordination with speech is also required for other types of co-speech gestures, such as iconic gestures (Habets et al. 2011). In many other cases, different effector systems have to be coordinated and synchronized for

complex, conjoined actions. For example, hand and foot actions need to be coordinated in complex movements, such as jumping. We would not be surprised if synchronization between other action modalities obeys the same principles that guide the coordination of speech and pointing gestures.

References

Chu, M., & Kita, S. (2008). Spontaneous gestures during mental rotation tasks: Insights into the microdevelopment of the motor strategy. *Journal of Experimental Psychology: General, 137,* 706–723.

De Ruiter, J. P. A. (1998). *Gesture and speech production.* Unpublished Ph.D. dissertation. University of Nijmegen, Nijmegen.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104,* 801–838.

Feyereisen, P. (1997). The competition between gesture and speech production in dual-task paradigms. *Journal of Memory and Language , 36,* 13–33.

Feyereisen, P., & de Lannoy, J. (1991). *Gestures and speech: Psychological investigations.* New York, NY: Cambridge University Press.

Franklin, D. W., & Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control. *Neuron, 72,* 425–442.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2008). *Cognitive Neuroscience: The Biology of the Mind. (3rd ed..)* New York: W. W. Norton & Company, Inc.

Habets, B., Kita, S., Shao, Z., Ozyurek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience, 23,* 1845-1854.

Hagoort, P., & Van Turennout, M. (1997). The electrophysiology of speaking: Possibilities of event-related potential research for speech production. In W. Hulstijn, H. Peters, & P. Van Lieshout (Eds.), *Speech motor production and fluency disorders: Brain research in speech production* (pp. 351-361). Amsterdam: Elsevier.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron, 69,* 407–22.

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition, 92,* 101-144.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science, 16,* 367–371.

Iverson, J. M., & Thelen, E. (1999). Hand, mouth, and brain: The dynamic emergence of speech and gesture. *Journal of Consciousness Studies, 6,* 19–40.

Kandel, E., Schwartz, J., Jessell, T., Siegelbaum, S., & Hudspeth, A. (2012) *Principles of neural science (5th edition).* New York, NY: McGraw Hill Medical.

Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge, UK: Cambridge University Press.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes, 24,* 145–167.

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,* 16–32.

Krauss, R. M., Chen, Y., & Chawla, P. (1996). Nonverbal behaviour and nonverbal communication: What do conversational hand gestures tell us? In M. Zanna (Eds.), *Advances in experimental social psychology* (pp. 389–450). San Diego, CA: Academic Press.

Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Eds.), *Language and Gesture* (pp. 261–283). Cambridge, UK: Cambridge University Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Levelt, W. J. M., Richardson, G., & La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language, 24,* 133–164.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–37.

Loehr, D. (2007). Aspects of rhythm in gesture and speech. *Gesture*, *7,* 179–214.

McClave, E. (1998). Pitch and manual gestures. *Journal of Psycholinguistic Research, 27,* 69–89.

McNeill, D. (1992). *Hand and Mind.* Chicago: University of Chicago Press.

McNeill, D. (2005). *Gesture and Thought.* Chicago: University of Chicago Press.

McNeill, D. (2012). *How Language Began: Gesture and Speech in Human Evolution.* Cambridge, UK: Cambridge University Press.

McNeill, D. & Duncan, S. D. (2000). Growth points in thinking-for-speaking. In D. McNeill (Eds.), *Language and Gesture,* (pp. 141-161). Cambridge, UK: Cambridge University Press.

Miall, R. C., Weir, D. J., Wolpert, D. M., & Stein, J. F. (1993). Is the cerebellum a Smith predictor? *Journal of Motor Behaviour, 25,* 203–216.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107,* 460–499.

Tuite, K. (1993). The production of gesture. *Semiotica, 93,* 83–105.

Welch, R. B., & Warren, D. H. (1986). Intersensory interactions. In J. P. Thomas (Eds.), *Handbook of perception and human performance, Vol. 1: Sensory processes and perception* (pp. 25.1-25.36). New York: Wiley.

Footnotes

[1] When G-apex time in the gesture-only condition was submitted to the same ANOVA, the results were similar those obtained in the gesture-and-speech condition.

[2] In the supplementary material, we reported an extra experiment showing that participants' behaved similarly regardless whether there was 117 ms delay between the gesture movement and visual feedback presented in virtual reality and whether there was visual feedback presented in virtual reality.

[3] When G-exec time in the gesture-only condition was submitted to the same ANOVA, the results were similar to those obtained in the gesture-and-speech condition.

[4] We did not differentiate "this" and "that" in this experiment since we were only interested in the effect of changing the required color name on gesture production. However, the word "this" was kept as part of the speech response because it is natural to include a referring word in a pointing situation.)

[5] The correct color name refers to the name of the target light color in the non-color-changed trials and the name of the new target light color in the color-changed trials. Trials in which an incorrect color was given, and not repaired, were counted as error trials and excluded from analysis.

Both S-onset time of the word "dit" and the color word were recorded. However, S-onset time of the color word was used for analysis because the color of the light was manipulated in this experiment.

[6] The color change occurred at similar time in the gesture-and-speech condition and in the gesture-only condition. We submitted the time between the moment of light color changing and

G-apex time to an ANOVA with condition (gesture-and-speech condition and gesture-only condition), and trial type (early, middle, late color-changed trials) as independent variables. There was no main effect of condition (p = .20) or interaction between condition and trial type (p = .24).

[7] Since Experiment 1 to 4 investigated the effect of gesture disruption on the speech onset time (i.e., the moment that articulation starts), the results cannot tell us whether disrupting gesture execution affects the articulation phase (i.e., the period between the start and the end of articulation).

Figure 1a.

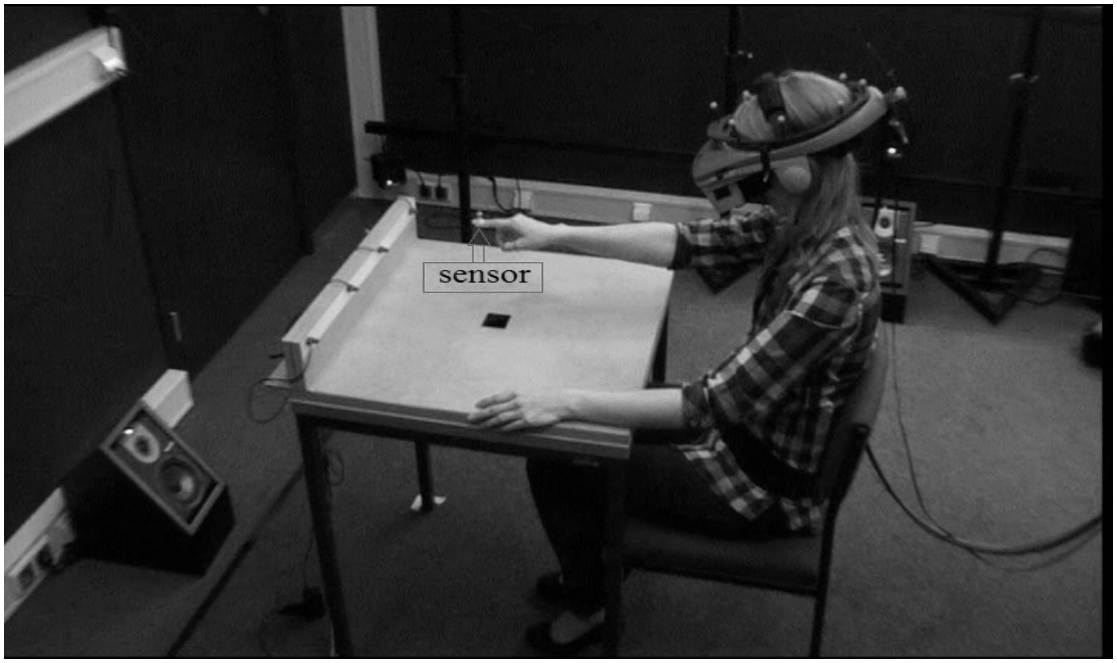

Figure 1b.

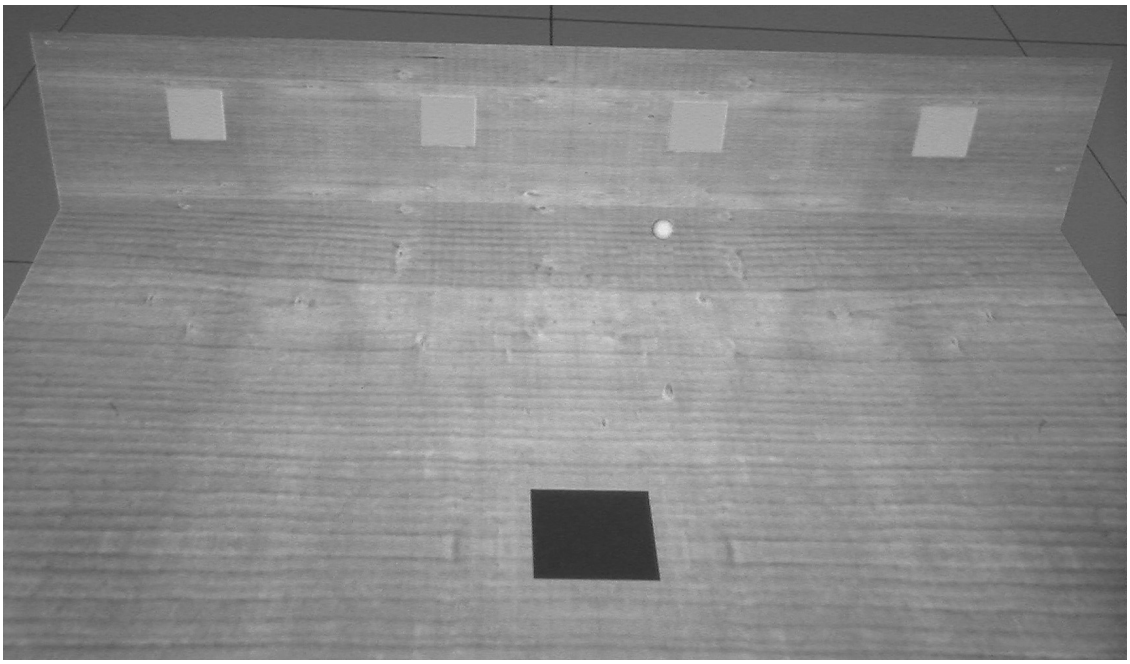

Figure 1. (a) A marker is attached to the tip of the participant's right index finger. The start button was located in the center of the table. The distance was 32 cm between the center of start button and the bottom edge of the table and was about 10 cm between the bottom edge of table and the participant's upper body. (b) The center button represented the start button. The white ball represented the tip of the participant's right index finger. The four white squares represented the four lights and the distance between each light was 20 cm.

Figure 2.

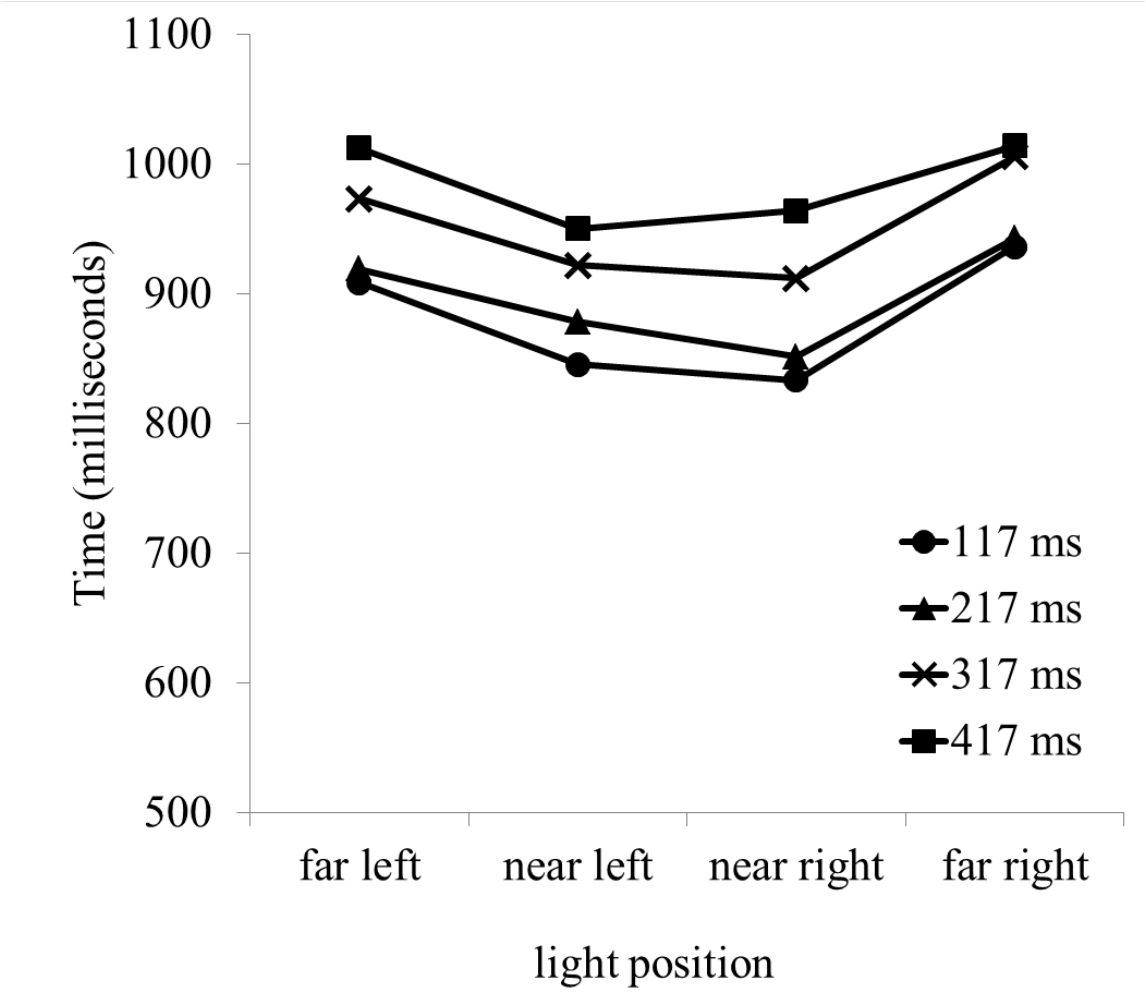

Figure 2. Mean G-apex time of far left, near left, near right, and far right lights in the 117 ms, 217 ms, 317 ms, and 417 ms delay interval trials in the gesture-and-speech condition of Experiment 1.

Figure 3.

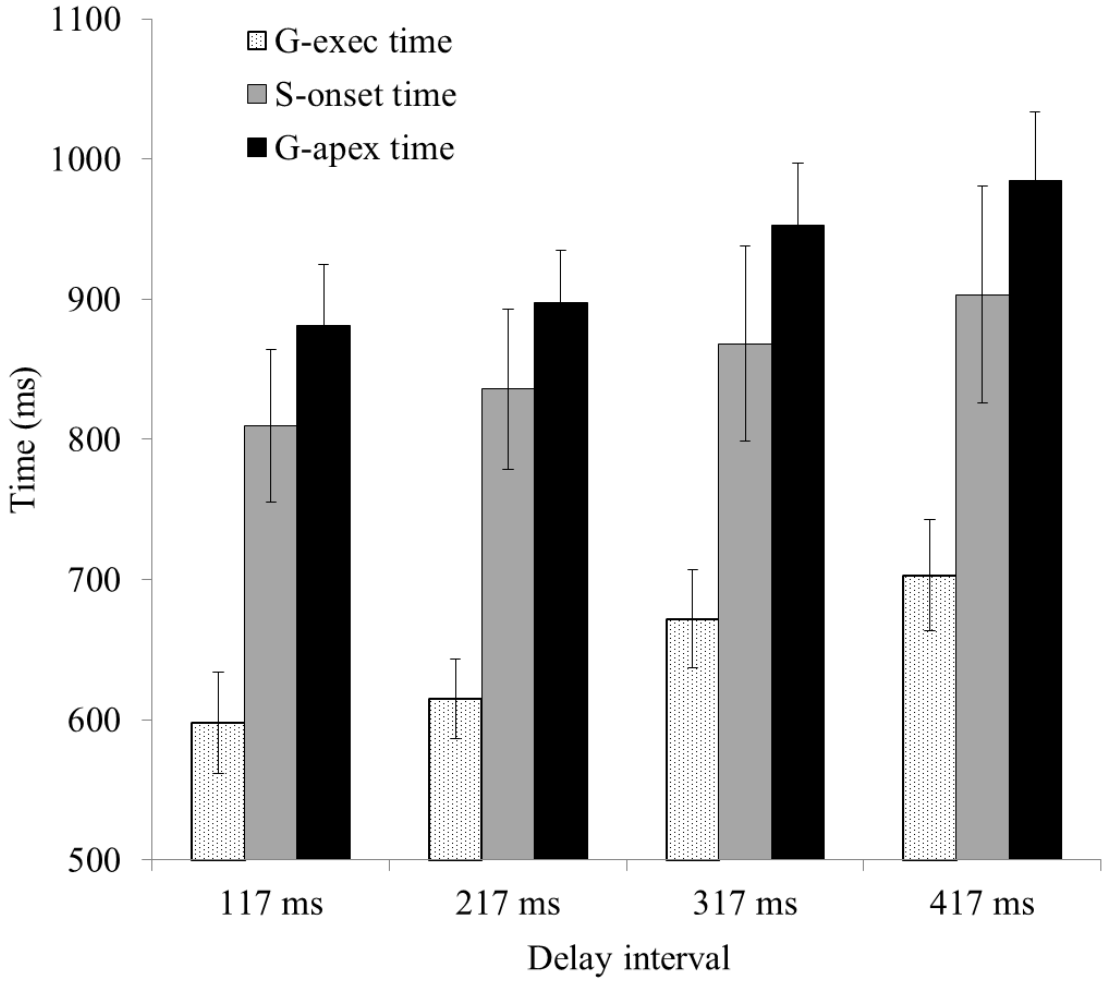

Figure 3. Mean gesture execution time (G-exec time), speech onset time (S-onset time), and gesture apex time (G-apex time) of the 117 ms, 217 ms, 317 ms, and 417 ms delay interval trials in the gesture-and-speech condition of Experiment 1. The error bars represent standard errors.

Figure 4.

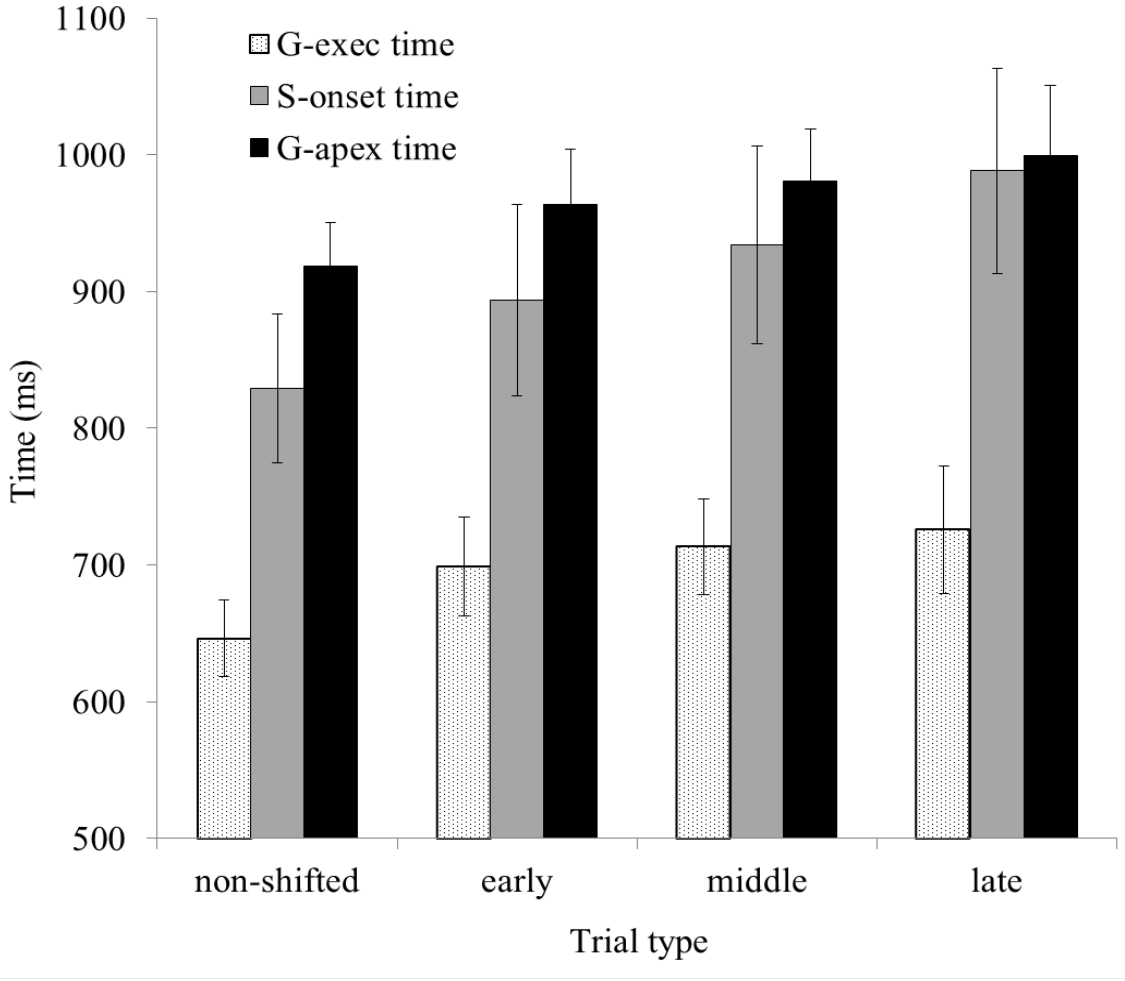

Figure 4. Mean gesture execution time (G-exec time), speech onset time (S-onset time), and gesture apex time (G-apex time) of the non-shifted, early, middle, and late ball-shifted trials in Experiment 2. The error bars represent standard errors.

Figure 5

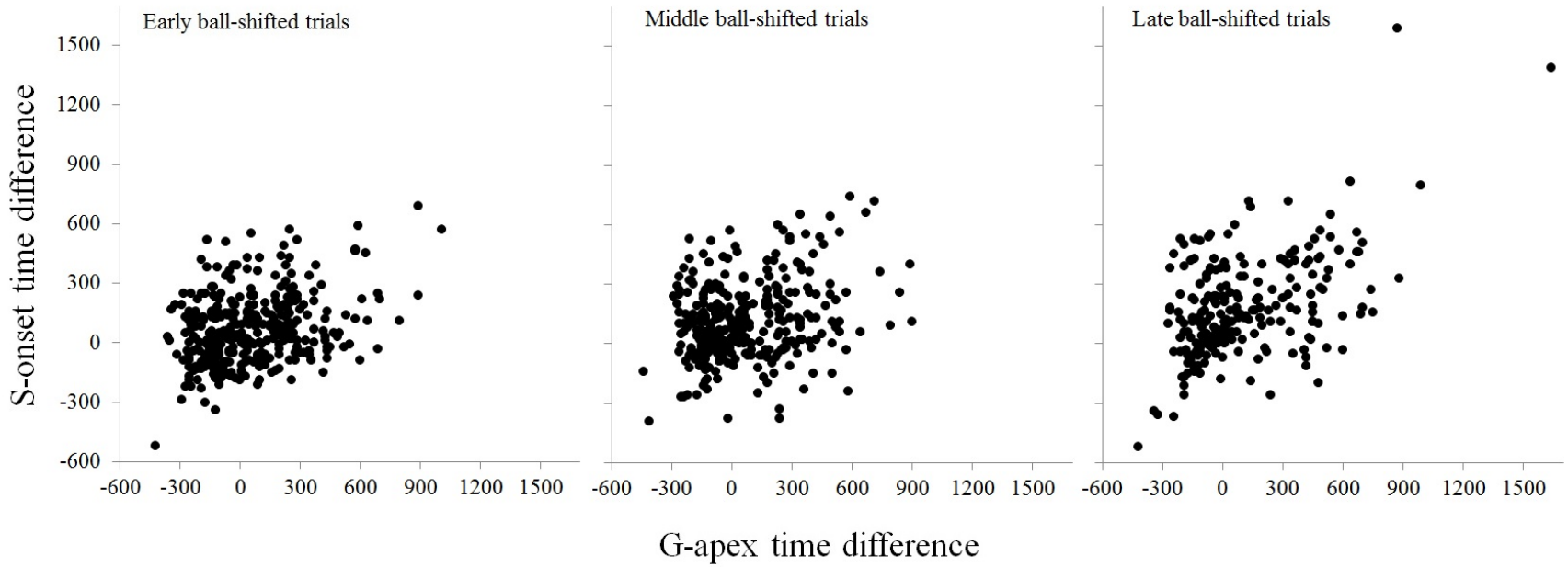

Figure 5. Scatter plot of the correlations between G-apex time difference (G-apex time in the ball-shifted trials *minus* G-apex time in the non-shifted trials) and S-onset time difference (S-onset time in the ball-shifted trials *minus* S-onset time in the non-shifted trials) in the early, middle, and late ball-shifted trials.

Figure 6.

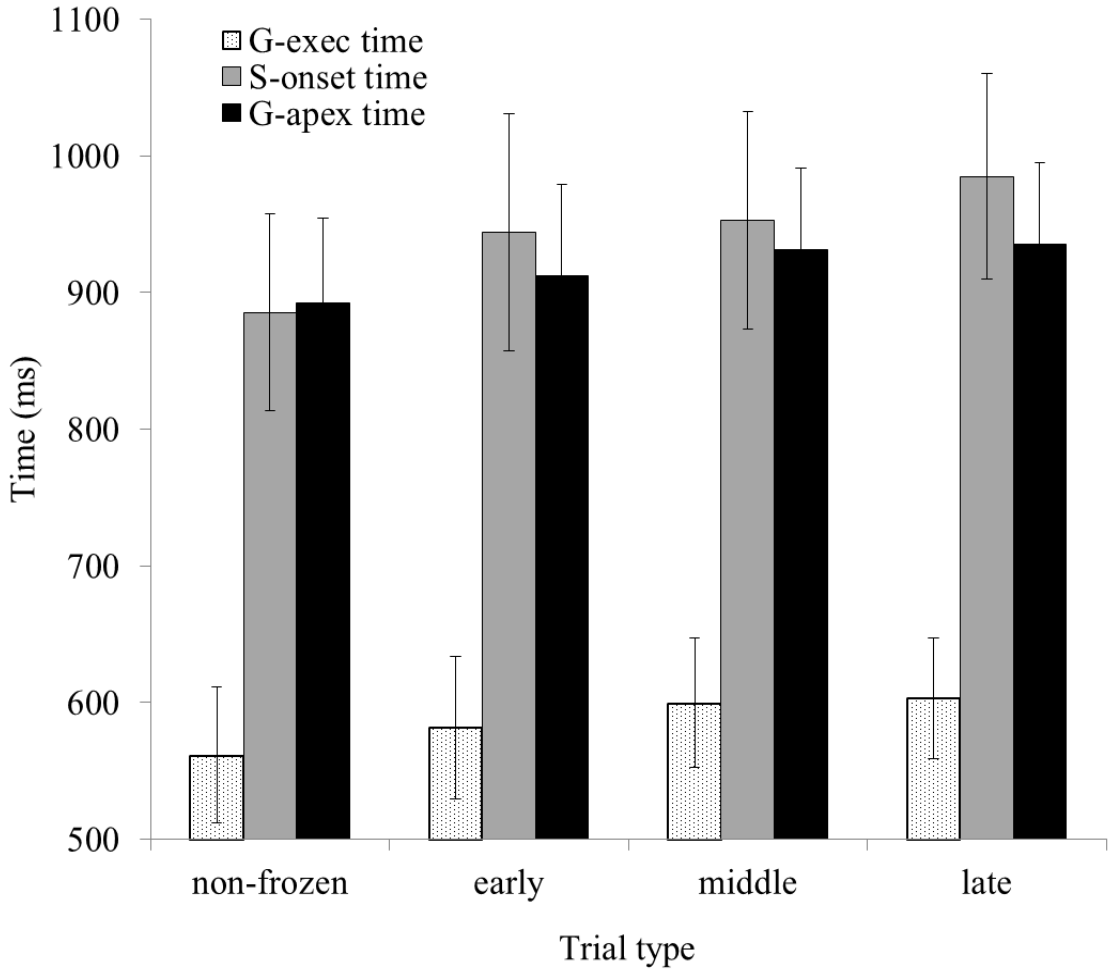

Figure 6. Mean gesture execution time (G-exec time), speech onset time (S-onset time), and gesture apex time (G-apex time) of the non-frozen, early, middle, and late ball-frozen trials in Experiment 3. The error bars represent standard errors.

Figure 7

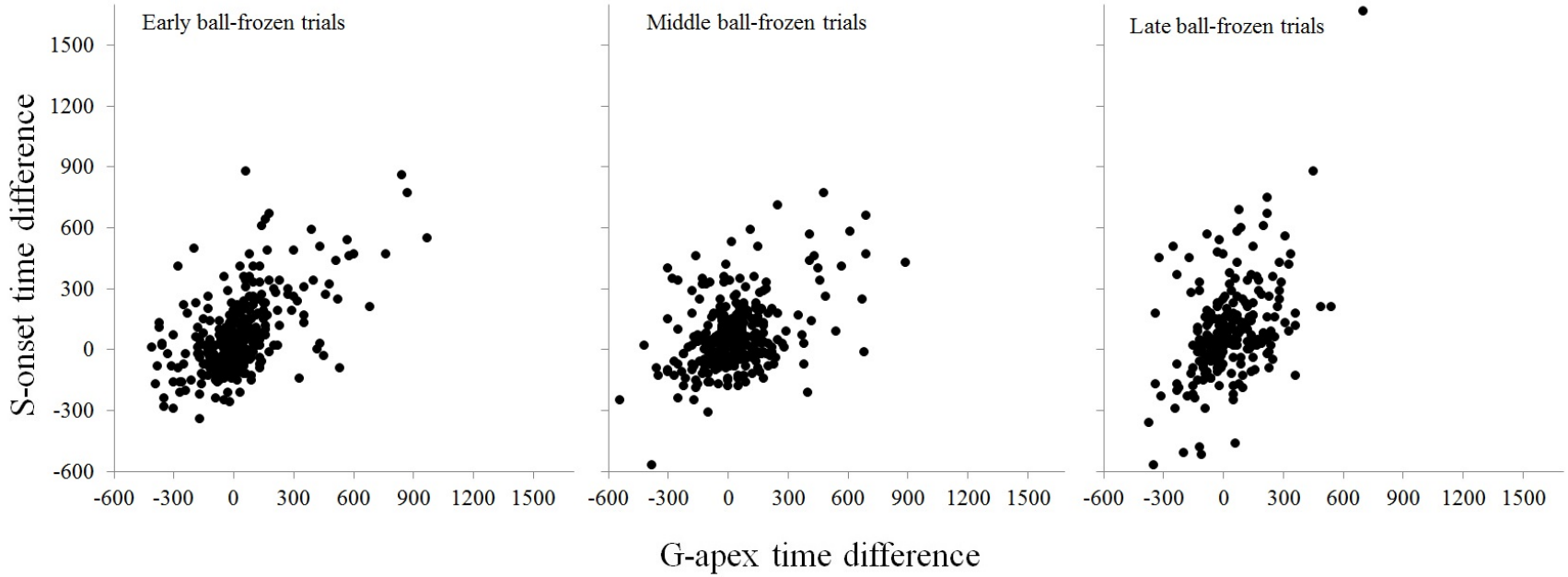

Figure 7. Scatter plot of the correlations between G-apex time difference (G-apex time in the ball-frozen trials *minus* G-apex time in the non- frozen trials) and S-onset time difference (S-onset time in the ball- frozen trials *minus* S-onset time in the non- frozen trials) in the early, middle, and late ball-frozen trials.

Figure 8.

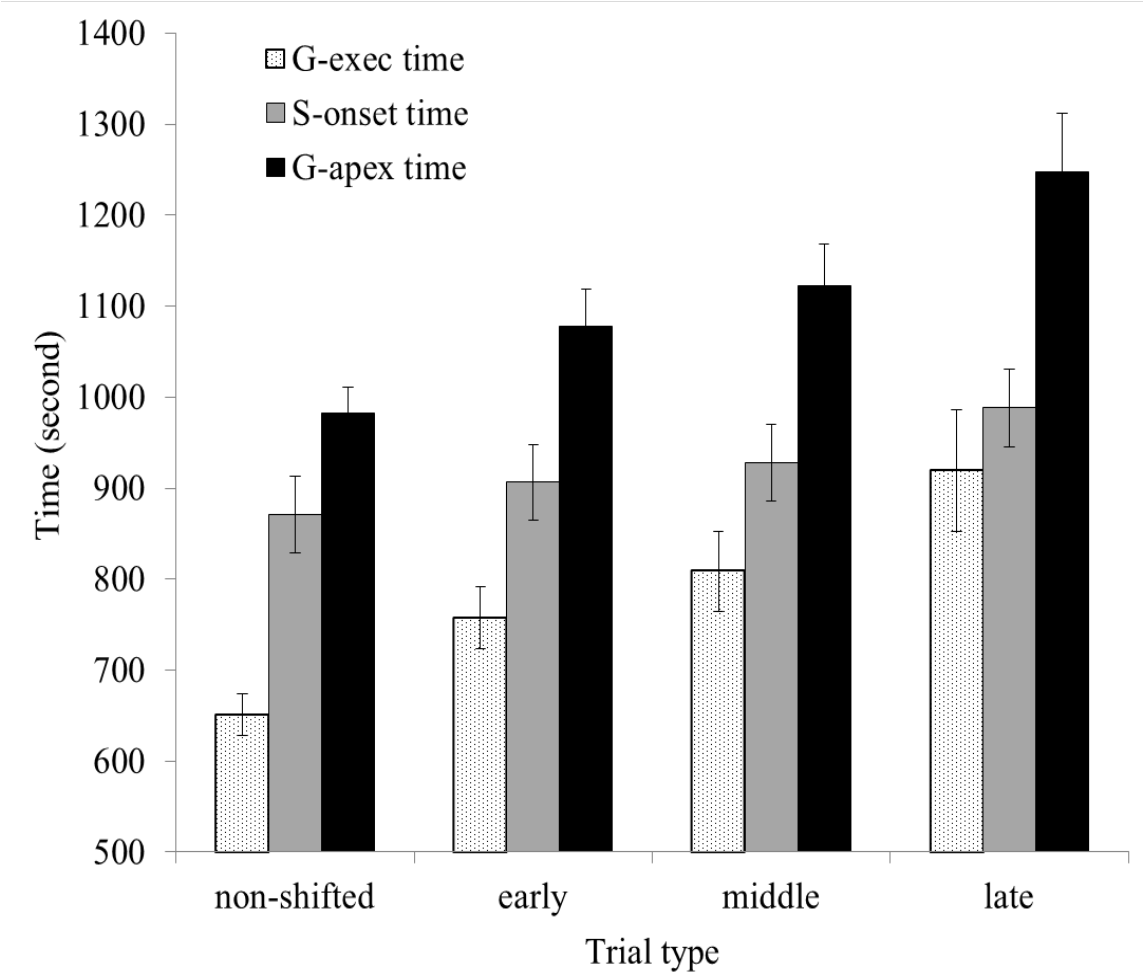

Figure 8. Mean gesture execution time (G-exec time), speech onset time (S-onset time), and gesture apex time (G-apex time) of the non-shifted, early, middle, and late light-shifted trials in the gesture-and-speech condition of Experiment 4. The error bars represent standard errors.

Figure 9.

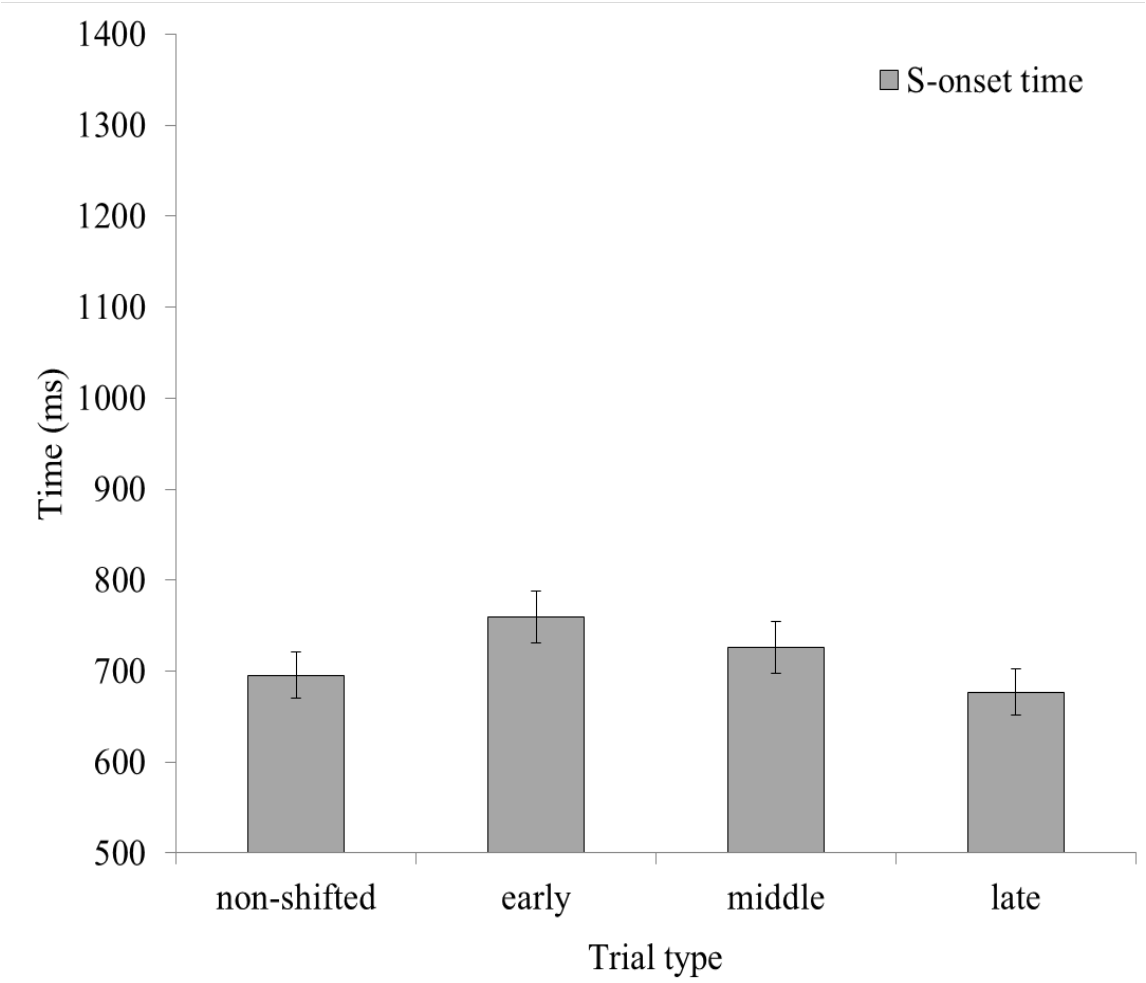

Figure 9. Mean speech onset time (S-onset time) of the non-shifted, early, middle, and late light-shifted trials in the speech-only condition of Experiment 4. The error bars represent standard errors.

Figure 10.

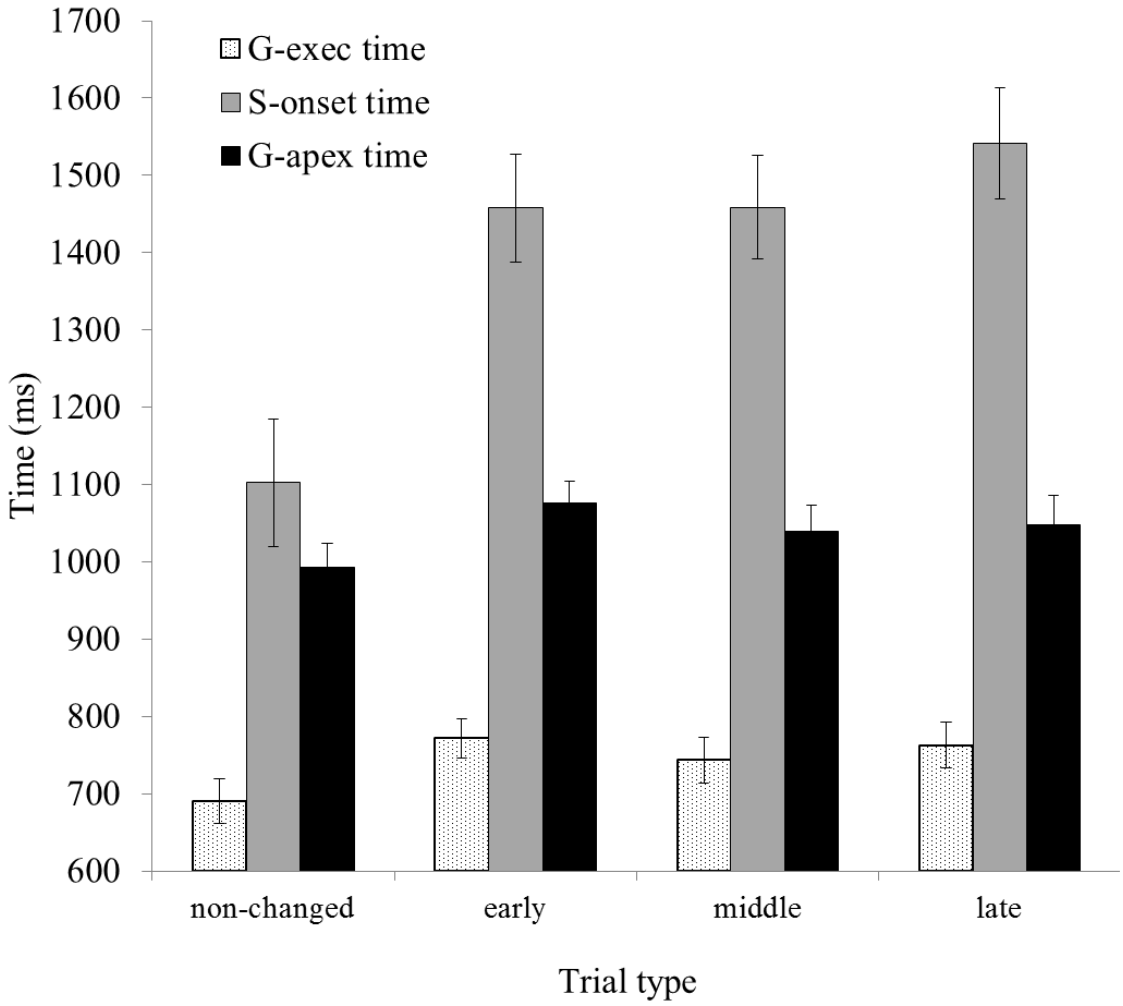

Figure 10. Mean gesture execution time (G-exec time), speech onset time (S-onset time), and gesture apex time (G-apex time) of the non-color-changed, early, middle, and late color-changed trials in the gesture-and-speech condition of Experiment 5. The error bars represent standard errors.

Figure 11.

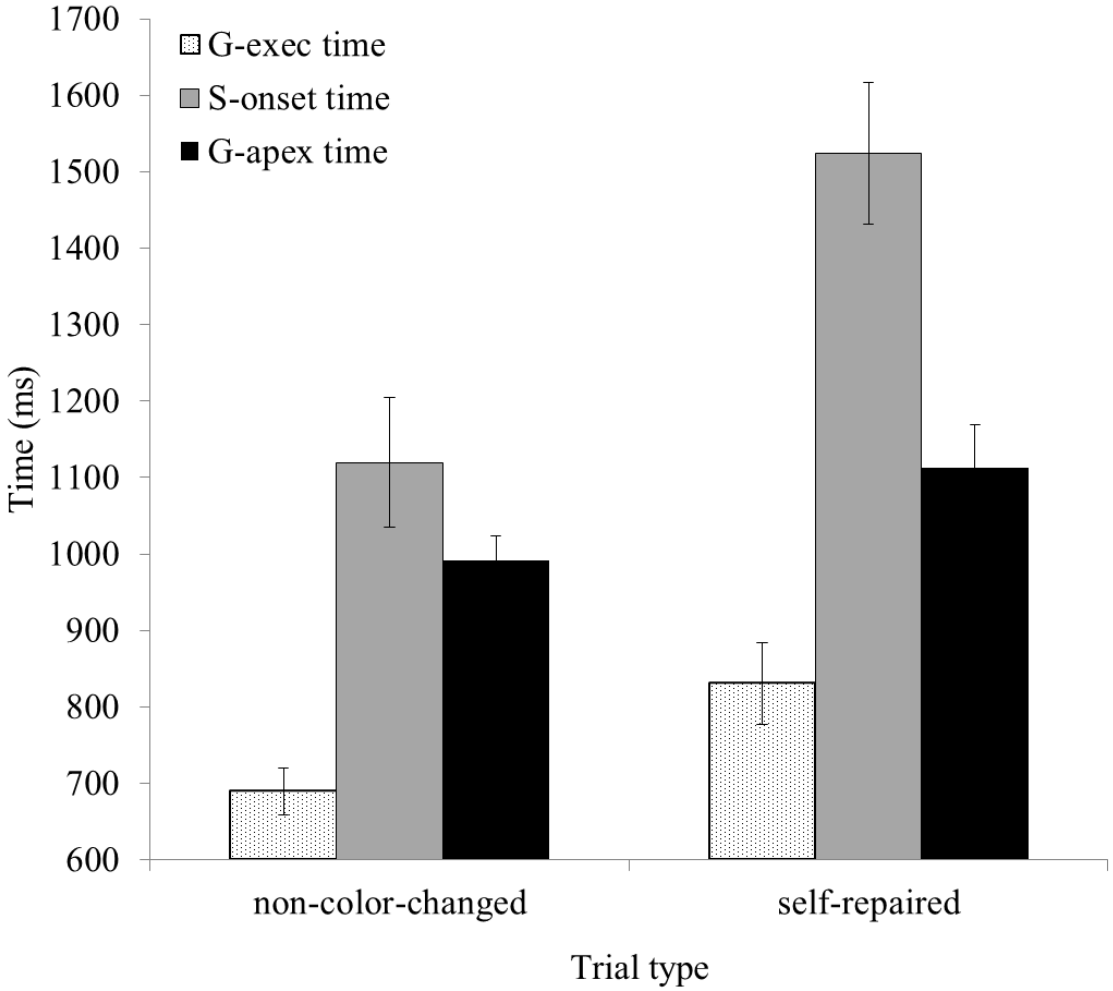

Figure 11. Mean gesture execution time (G-exec time), gesture apex time (G-apex time), and speech onset time (S-onset time) of the non-color-changed trials (with no self-repaired speech) and the color-changed trials (with self-repaired speech and in which "dit" was articulated before gesture apex) of the gesture-and-speech condition of Experiment 5. The error bars represent standard errors.

Figure 12.

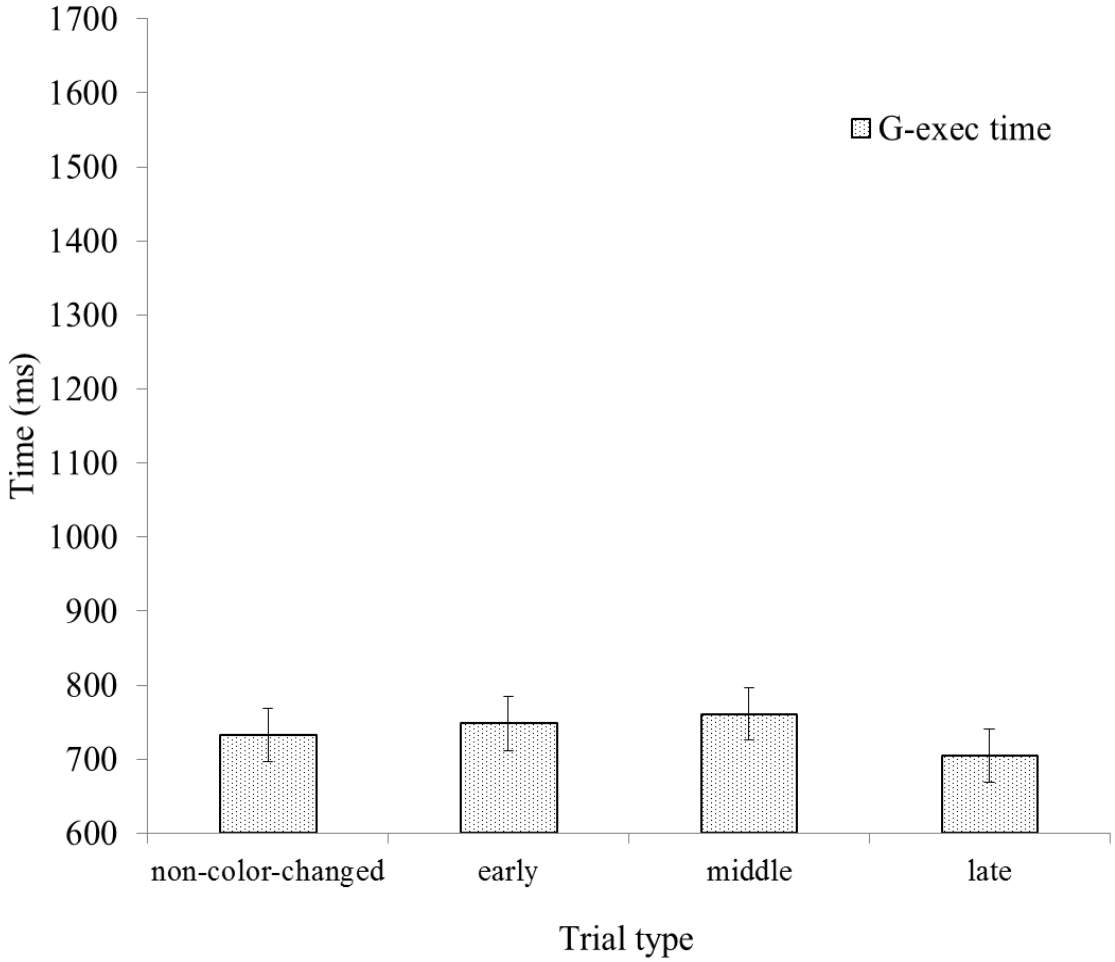

Figure 12. Mean gesture execution time (G-exec time) of the non-color-changed, early, middle, and late color-changed trials in the gesture-only condition of Experiment 5. The error bars represent standard errors.

Figure 13.

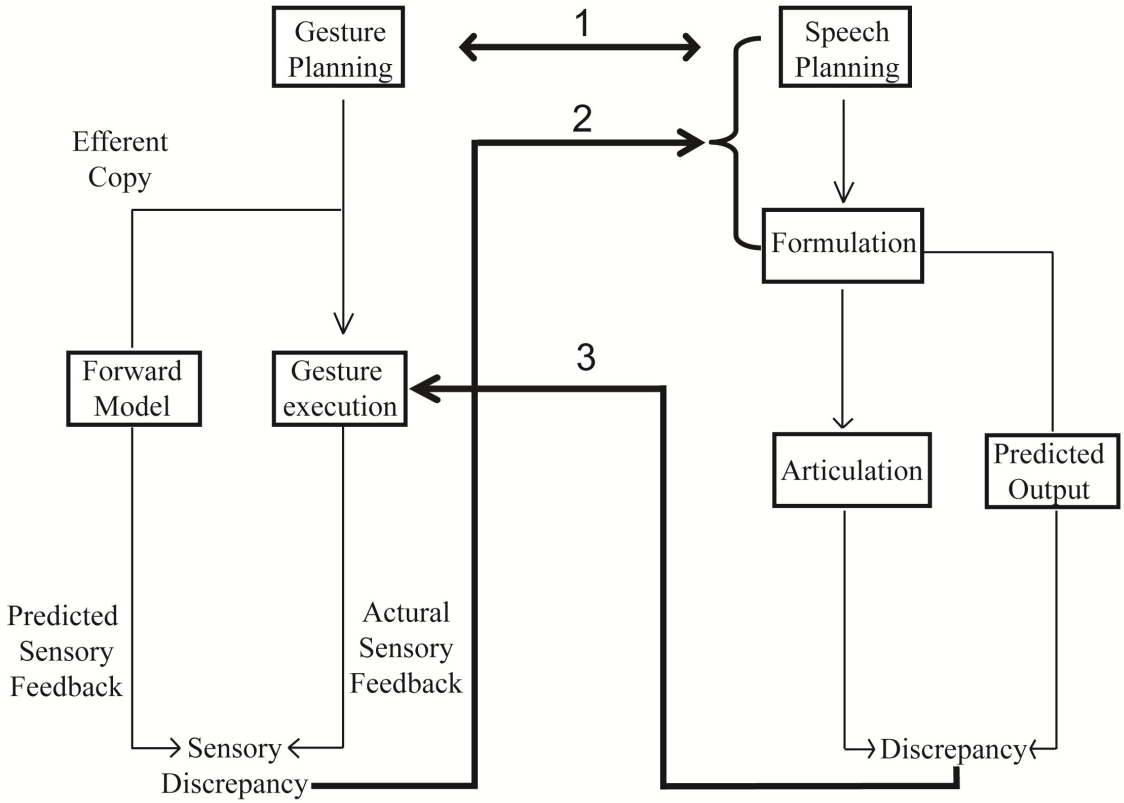

Figure 13. The interaction between gesture and speech (the left diagram is adapted from Miall, Weir, Wolpert, and Stein (1993); the right diagram is adapted from Indefrey and Levelt (2004) and Hickok, Houde, and Rong (2011)). Previous studies showed that gesture and speech interact during their planning phases (indicated by arrow 1). In Experiment 1 to 4, the sensory feedback of gesture was manipulated, resulting in a discrepancy between the predicted sensory feedback from the forward model and the actual sensory feedback. This discrepancy is used to generate a new gestural motor command *and* to inform the speech system to adapt its time course. The time course of speech can be adjusted during the planning and the formulation phases of speaking. In Experiment 5, when a discrepancy between the predicted speech output and the actual speech output was detected before or during speech execution, this discrepancy is used to generate a new speech plan *and* to inform the gesture system to adapt the time course of gesture execution.