LUND UNIVERSITY

Attention to speech-accompanying gestures: Eye movements and information uptake

Gullberg, Marianne; Kita, Sotaro

Published in:
Journal of Nonverbal Behavior

DOI:
10.1007/s10919-009-0073-2

2009

Link to publication

Citation for published version (APA):
Gullberg, M., & Kita, S. (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behavior*, *33*(4), 251-277. https://doi.org/10.1007/s10919-009-0073-2

Total number of authors:
2

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Download date: 12. Oct. 2024

Editorial Manager(tm) for Journal of Nonverbal Behavior
Manuscript Draft

Manuscript Number: JONB210R2

Title: Attention to Speech-Accompanying Gestures: Eye Movements and Information Uptake

Article Type: Original Research

Keywords: gesture; interaction; eye gaze; fixation; multimodal information processing

Corresponding Author: Dr Marianne Gullberg, Ph.D.

Corresponding Author's Institution: Max Planck Institute for Psycholinguistics

First Author: Marianne Gullberg, PhD

Order of Authors: Marianne Gullberg, PhD; Sotaro Kita, Ph.D.

Gullberg & Kita Response to Editors

Editor's comments in *italics*, and responses in plain font.

Please rename the section The Current Study to The current research (don't use capital letters in headings) and make the corresponding change in the text. Then rename Experiment 1 etc to Study 1 etc.
Done.

Please note the gender distribution and the mean age of all participants, as well as how much participants were paid
Done.

Please make sure that you report means and SD for all analyses.
Done.

Page 9, line 43 "Each video clip contained a whole, unedited story retelling containing sequences of gestures" -> something is wrong with this sentence
The sentence has been amended.

Psge 12, line 28,30: proportions -> the proportion
Fixed.

Page 13, line 57, please add a sentence or two describing McNeill's schema?
Done.

Page 16, line 55, Although the mean duration of the real speaker-fixations in the original Speaker-fixation condition in Experiment 1 was 980 ms, the Artificial speaker-fixations had to be shorter for the manipulation to align with the shorter gesture strokes of the original Central, No hold, No-speaker-fixated gestures. -> please specify the mean duration of the artificial fixations in the text
Done.

Page 17, line 37, please add the mean and SD
These numbers appeared in lines 31 and 33 for the two relevant conditions.

I would also like to ask you to carefully check the references in the reference section against the text as in my experience this is an area where errors can easily slip in especially during a revision.
Done.

*Manuscript
Click here to download Manuscript: Gullberg_Kita_rev_fin_details.doc Click here to view linked References

Attention to Speech-Accompanying Gestures: Eye Movements and Information Uptake

Marianne Gullberg[i] & Sotaro Kita[ii]

[i]Radboud University Nijmegen & Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

[ii]School of Psychology, University of Birmingham, Birmingham, United Kingdom

Corresponding author:

Marianne Gullberg

Max Planck Institute for Psycholinguistics

PO BOX 310

6500 AH Nijmegen

The Netherlands

Email: marianne.gullberg@mpi.nl

Telephone: +31-24-352 1271

Abstract

There is growing evidence that addressees in interaction integrate the semantic information conveyed by speakers' gestures. Little is known, however, about whether and how addressees' attention to gestures and the integration of gestural information can be modulated. This study examines the influence of a social factor (speakers' gaze to their own gestures), and two physical factors (the gesture's location in gesture space and gestural holds) on addressees' overt visual attention to gestures (direct fixations of gestures) and their uptake of gestural information. It also examines the relationship between gaze and uptake. The results indicate that addressees' overt visual attention to gestures is affected both by speakers' gaze and holds but for different reasons, whereas location in space plays no role. Addressees' uptake of gesture information is only influenced by speakers' gaze. There is little evidence of a direct relationship between addressees' direct fixations of gestures and their uptake.

Keywords: gesture, interaction, eye gaze, fixation, multimodal information processing

Acknowledgements

We gratefully acknowledge financial and technical support from the Max Planck Institute for Psycholinguistics. We also thank Wilma Jongejan for help with the video manipulations, and Martha Alibali, Kenneth Holmqvist, and members of the Max Planck Institute's Gesture project for useful discussions.

Attention to Speech-Accompanying Gestures: Eye Movements and Information Uptake

When we talk, we typically also gesture, that is, we perform manual movements as part of the expressive effort (Kendon, 2004; McNeill, 1992). Such speech-accompanying gestures typically convey meaning (e.g. size, shape, direction of movement), which is related to the ongoing talk. The communicative role of these gestures is somewhat controversial. It is debated both whether speakers actually intend gestural information for their addressees (e.g. Holler & Beattie, 2003; Melinger & Levelt, 2004), and whether addressees attend to and integrate the gestural information. This paper focuses on the latter issue.

There is growing evidence that speech and speech-accompanying gestures are processed and comprehended together, forming an ‘integrated’ system or a ‘composite signal’ (e.g. Clark, 1996; Kendon, 2004; McNeill, 1992). Gestural information is integrated with speech in comprehension and influences the interpretation and memory of speech (e.g. Beattie & Shovelton, 1999a, 2005; Kelly, Barr, Breckinridge Church, & Lynch, 1999; Langton & Bruce, 2000; Langton, O'Malley, & Bruce, 1996). For instance, information expressed only in gestures re-surfaces in retellings, either as speech, as gesture, or both (Cassell, McNeill, & McCullough, 1999; McNeill, Cassell, & McCullough, 1994). Further, neurocognitive studies show that incongruencies between information in speech and gesture yield electrophysiological markers of integration difficulties such as the N400 (e.g. Özyürek, Willems, Kita, & Hagoort, 2007; Wu & Coulson, 2005). However, surprisingly few studies have attempted to examine directly whether attention to gestures and uptake of gestural information is deterministic and unavoidable or whether such attention is modulated in human interaction, and if so by what factors. Furthermore, surprisingly little is known about the role of gaze in this context. This study therefore aims to examine what factors influence overt, direct visual attention to gestures and uptake of gestural information, focusing on one social factor, namely speakers’ gaze at their own gestures, and two physical properties of gestures, namely their place in gesture space and the effect of gestural holds. The study also examines what the relationship is between addressees’ gaze and uptake.

Visual attention to gestures

Gestures are visuo-spatial phenomena, and so the role of vision and gaze for attention is important. However, addressees seem to gaze directly at speakers' gestures relatively rarely. Addressees

mainly look at the speaker's face in interaction (Argyle & Cook, 1976; Argyle & Graham, 1976; Bavelas, Coates, & Johnson, 2002; Fehr & Exline, 1987; Kendon, 1990; Kleinke, 1986). Studies using eye-tracking techniques in face-to-face interaction have further demonstrated that addressees spend as much as 90-95% of the total viewing time fixating the speaker's face and thus fixate only a minority of gestures (Gullberg & Holmqvist, 1999; 2006).

However, the likelihood of an addressee directly fixating a gesture increases under the following three circumstances (Gullberg & Holmqvist, 1999; 2006; Nobe, Hayamizu, Hasegawa, & Takahashi, 1998; 2000). The first is when speakers first look at their own gestures (speaker-fixation) (Gullberg & Holmqvist, 1999; 2006). This tendency is stronger in live face-to-face interaction than when observing speakers on video (Gullberg & Holmqvist, 2006). This suggests that the overt shift of visual attention to the target of a speaker's gaze is essentially social in nature rather than an automatic response. The second circumstance is when a gesture is produced in the periphery of gesture space in front of the speaker's body (cf. McNeill, 1992). The third is when a gestural movement is suspended momentarily in mid-air and goes into a hold before moving on (cf. Kendon, 1980; Kita, Van Gijn, & Van der Hulst, 1998; Seyfeddinipur, 2006). Holds are often found between the dynamic movement phase of a gesture, the stroke, and the so-called retraction phase, which marks the end of a gesture. It is currently not clear whether these three factors – speaker-fixation, peripheral articulation, and holds – all contribute independently to the increased likelihood of the addressee's fixation on gesture. The evidence for the influence of these three factors mostly comes from observational studies on naturalistic conversations, in which the three factors often co-occur (Gullberg & Holmqvist, 1999; 2006; see also Nobe et al., 1998; 2000). Therefore, one of the goals of this study is to experimentally manipulate these factors and assess their contributions to the likelihood of addressees' fixations of gesture.

The three factors may draw the addressee's attention either for bottom-up, stimulus-related reasons or for top-down, social-cognitive reasons. Gestures in peripheral gesture space or with a hold may elicit the addressee's fixation for bottom-up reasons, namely, because these gestures challenge peripheral vision. Firstly, the acuity of peripheral vision decreases the further away from the fovea the image is projected, and secondly, peripheral vision, which is good at motion detection, cannot process information about a static hand in a hold efficiently. In contrast, gestures with speaker-fixations may

elicit the addressee's fixation for top-down social reasons, namely to manifest social alignment or joint attention. The difference between bottom-up and top-down processes should be reflected in different onset-latencies of fixations to gestures (cf. Gullberg & Holmqvist, 2006). Fixation onsets that are bottom-up driven should be short, whereas fixations driven by top-down concerns should have longer onsets (e.g. Yantis, 1998; 2000). Thus, another goal of the study is to compare the onset-latency for fixations on gestures triggered by the three factors to further elucidate the reasons for fixation.

Uptake of gestural information

Only a few studies have attempted to directly examine whether attention to and uptake of information from gestures is unavoidable or whether it is ever modulated and if so by what factors. Rogers (1978) manipulated noise levels showing that addressees pick up more information from gestures the less comprehensible the speech signal is. Beattie & Shovelton (1999a; 1999b) demonstrated that addressees decode information about relative position and size better when presented with speech and gesture combined than with either gesture or speech alone. Interestingly, this study also indicated that not all gestural information was equally decodable. Addressees reliably picked up location and size information pertaining to objects, but did worse with information such as direction. These studies indicate that the comprehensibility of speech affects addressees' attention to gestures and also that the type of gestural information matters.

Other factors may also modulate addressees' attention to gestures. Speakers' gaze to their own gestures, a factor of a social nature, is a likely candidate. It is well-known that humans are extremely sensitive to the gaze direction of others (e.g. Gibson & Pick, 1963), and that gaze plays a role in the establishment of joint attention (e.g. Langton, Watt, & Bruce, 2000; Moore & Dunham, 1995; Tomasello, 1999; Tomasello & Todd, 1983). It has been suggested that speakers look at their own gestures as a means to draw addressees' attention to them in face-to-face interaction (e.g. Goodwin, 1981; Streeck, 1993; 1994). Such behavior could increase the likelihood of addressees' uptake of gestural information, although this has not been tested with naturalistic, dynamic gestures that are not pointing gestures.

Physical properties of gestures may also affect addressees' uptake of gestural information. First, the location of the gesture in gesture space may matter (cf. McNeill, 1992). Speakers often bring gestures up into central gesture space, that is, to chest height and closer to the face, when they want

to highlight the relevance of gestures in interaction (e.g. Goodwin, 1981; Gullberg, 1998; Streeck, 1993; 1994). The information expressed by such a gesture seems more likely to be integrated than that of a gesture articulated for instance on the speaker's lap in lower, peripheral gesture space.

A second potentially important physical property is the gestural hold. The functional role of holds is somewhat debated, but holds have been implicated in turn taking and floor holding in interaction. Transitions between speaker turns in interaction are likelier once a gesture is terminated or when a tensed hand position is relaxed (e.g. Duncan, 1973; Fornel, 1992; Goodwin, 1981; Heath, 1986). If holds are a first indication that speakers are about to give up their turn, it would be communicatively useful for addressees to attend to them. This in turn may increase the likelihood of information uptake from a gesture with a hold. A further goal of this study, then, is to examine the impact of these three factors on addressees' uptake of gesture information.

The relationship between fixations and information uptake

As indicated above, most gestures are perceived through peripheral vision. Although peripheral vision is powerful, optimal image quality with detailed texture and color information is achieved only in direct fixations, that is, if the image falls directly on the small central fovea. Outside of the fovea, parafoveal or peripheral vision gives much less detailed information (Bruce & Green, 1985; Latham & Whitaker, 1996). Consequently, it is generally assumed that an overt fixation equals attention in the sense of information uptake. If addressees shift their gaze from the speaker's face to a gesture in interaction, this might indicate that they are attempting to integrate the gestural information (e.g. Goodwin, 1981; Streeck, 1993; 1994).

However, addressees' tendency to gaze directly at an information source is modulated in face-to-face interaction by culture-specific norms for maintained or mutual gaze to indicate continued attention (e.g. Rossano, Brown, & Levinson, in press; Watson, 1970). In cultures where mutual gaze is socially important, face-to-face interaction may emphasize the reliance on peripheral vision for gesture processing and dissociation between overt and covert attention. Addressees can fixate a visual target without attending to it ("looking without seeing"), and conversely, attend to something without directly fixating it ("seeing without looking"). If the speaker's face is the default location of visual attention in interaction, then most gestures must be attended to covertly. It is therefore not entirely clear what the relationship between overt fixation and information uptake might be in

interaction from information sources like gestures. A final goal of this study is therefore to examine the relationship between overt fixation of and uptake of information form gestures.

The current research

This study aims to examine what factors modulate addressees' visual attention to and information uptake from gestures in interaction by asking the following questions:

1) Do social and physical factors influence addressees' fixations of speakers' gestures? Furthermore, do different factors trigger qualitatively different fixations, reflecting the difference between top-down vs. bottom-up processes? We expect top-down driven fixations to have longer onset latencies than bottom-up driven fixations.

2) Do social and physical factors influence addressees' uptake of gesture information?

3) Are addressees' fixations a good index of information uptake from gestures?

To examine these questions we present participants ('addressees') with video recordings of naturally occurring gestures embedded in narratives. We examine the effect of a social factor, namely the presence/absence of speakers' fixations of their own gestures (Study 1), and the effect of two physical properties of gestures, namely gestures' location in gesture space (central/peripheral) and the presence/absence of holds (Study 2). In Studies 1 and 2, we manipulate the independent variables by selecting gestures with the relevant properties from a corpus of video recorded gestures. In a second set of control experiments, we present participants with digitally manipulated versions of the gesture stimuli used in Study 1 and 2, examining the effect of presence/absence of speakers' artificial fixations of their own gestures (Study 3) and the presence/absence of artificial holds (Study 4). These studies are undertaken to control for any other unknown variables that may have differed between the stimulus gestures used in the conditions in Studies 1 and 2.

In all studies, participants were presented with brief narratives that included a range of gestures, but our analyses focus on one "target gesture" in each narrative. Each target gesture conveyed information about the direction of a movement. This information was only encoded in the target gesture, and not in other gestures or in speech. Overt visual attention to gestures was operationalized as direct fixations of gestures. Participants' eye movements were recorded during the presentation of the narratives using a head-mounted eye-tracker. Further, information uptake was operationalized as the extent to which participants could reproduce the information conveyed in the

target gesture in a drawing task following stimulus presentation. Participants were asked to draw an event in the story that crucially involved the movement depicted by the target gesture. The match between the directionality of the movement in the drawing and in the target gesture was taken as indicative of information up-take.

Study 1 Speaker-fixations

The first study examines the effect of a social factor on addressees' overt visual attention to and uptake of information from gestures, namely the presence/absence of speakers' fixations of their own gestures.

Method

Participants

Thirty Dutch undergraduate students from Radboud University Nijmegen participated in this study (*M* age = 22, *SD* = 3), 23 females and 7 males. They were paid 5 euros for their participation.

Materials

The stimuli were taken from a corpus of videotaped face-to-face story retellings in Dutch (Kita, 1996). The video clips showed speakers facing an addressee or viewer retelling short stories. The video clips did not show the original live addressee, but only the speaker seated en face. Each video clip contained a whole, unedited story retelling. Each clip therefore contained multiple gestures, only one of which was treated as a target gesture. Consequently, the target gesture appeared within sequences of other gestures so as not to draw attention as a singleton. The stimulus videos were selected from the corpus because they contained one target gesture displaying the appropriate properties. For Study 1, each target gesture displayed either presence or absence of speaker-fixation, that is, the speakers either looked at their own gestures or not. The target gestures were otherwise similar, and performed in central gesture space without holds. All target gestures were representational gestures encoding the movement of a protagonist in the story from an observer viewpoint (McNeill, 1992), meaning that the speaker's hand represented a protagonist in the story as seen from outside. The target gestures, typically expressing a key event in the story lines, encoded the direction of the protagonist's motion left or right. Although the movement itself was an important

part of the storyline, the direction of the movement was not. The directional information was only present in the target gesture and not in co-occurring speech. Further, the directional information could not be inferred from other surrounding gestures. Care was taken to ensure that the gestural information was not highlighted in any other way. Co-occurring speech did not contain any deictic expressions referring to and therefore drawing attention to the gesture (e.g., 'that way'). Moreover, the target gesture did not co-occur with hesitations in speech, with the story punch line or with first mention of a protagonist, as all of these features might have lent extra prominence to a co-occurring gesture. Descriptions of the animated cartoons used to elicit the narratives and the target scenes therein are provided in Appendix 1. Outlines of the spatio-temporal properties of the target gestures across conditions (and all studies) are provided in Appendix 2, and speech co-occurring with target gestures is listed in Appendix 3.

In Study 1, the target gestures consisted of gestures that were either fixated or not by the speaker in the video (Speaker-fixation vs. No-speaker-fixation). Location in gesture space and presence/absence of hold was held constant (central space, no hold). There were 4 items in each condition. The mean durations of the target gestures in each condition in Study 1 are summarized in Table 1.

INSERT TABLE 1 HERE

Apparatus

We used a head-mounted SMI iView© eye-tracker, which is a monocular 50 Hz pupil and corneal reflex video imaging system. The eye-tracker records the participant's eye movements with the corneal reflex camera. The eye-tracker also has a scene-camera on the headband, which records the field of vision. The output data from the eye-tracker consist of a merged video recording showing the addressee's field of vision (i.e. the speaker on the video), and an overlaid video recording of the addressee's fixations as a circle overlay. Since the scene-camera moves with the head, the eye-in-head signal indicates the gaze point with respect to the world. Head movements therefore appear on the video as full-field image motion. The fixation marker represents the foveal fixation and covers a visual angle of 2°. The output video data allow us to analyze both gesture and eye movements with a temporal accuracy of 40 ms.

Procedure

Participants were randomly assigned to one of the two conditions: Speaker-fixation (central space, no hold, speaker-fixation) and No-speaker-fixation (central space, no hold, no speaker-fixation). The participants were seated 250 cm from the wall and fitted with the SMI iView© headset. A projector placed immediately behind the subject projected a nine-point matrix calibration screen on the wall of the same size as the subsequent stimulus videos. After calibration, four stimulus video clips were projected against the wall. The speakers appearing in the videos were thus life-sized, and their heads were level with the participants' heads. Life-sized projections have been shown to yield fixation behavior towards gestures that is similar to behavior in live interaction (Gullberg & Holmqvist, 2006). A black screen appeared between each video clip for a duration of 10 seconds.

Participants were instructed to watch the videos carefully to be able to answer questions about them subsequently. The instructions did not mention gestures or the direction of the movements in the story. Participants' eye movements were recorded as they watched the video clips. After watching all four videos, participants answered questions about the target events of each video, exemplified in (1), by drawing pictures of the protagonists in the story (see Appendix 4 for the complete set of questions).

(1) *De muis heeft moeite met roeien. Hoe komt hij toch vooruit?* 'The mouse has trouble rowing. How does it make progress?'

The participants did not know the contents of the questions until they had finished watching all four videos. A drawing task was chosen because it allows directionality to be probed implicitly: The participant must apply a perspective on the event and the protagonist in order to draw them, a perspective which in turn will reveal the direction of the protagonist (see Fig. 1). The drawing task thus avoids the well-known difficulties involved in overt labelling of left-right directionality (e.g., Maki, Grandy & Hauge, 1979). A post-test-questionnaire ensured that gesture was not identified as the target of study.

Coding

The eye movement data were retrieved from the digitized video output from the eye-tracker. The merged video data of the participants' gaze positions on the scene image were analyzed frame-by-frame and coded for fixation of target gesture (Yes or No) and for matched reply (Yes or No). A target

gesture was coded as fixated if the fixation marker was immobile on the gesture, i.e. moved no more than 1 degree, for a minimum of 120 ms (=3 video frames) (cf. Melcher & Kowler, 2001). Note that fixations on gestures were spatially unambiguous. Either a gesture was clearly fixated, or the fixation marker stayed on the speaker's face (cf. Gullberg & Holmqvist, 1999; 2006). A drawing was coded as a matched reply if the direction of the motion in the drawing matched the direction of the target gesture on the video as seen from the addressee's perspective (see Fig. 1).[1] Only responses that could be coded as matched or non-matched were included in the analysis. When drawings did not depict a lateral direction of any kind, the data point was discarded. Chance performance therefore equals 50%.

INSERT FIG. 1 HERE

Analysis

The dependent variables were (a) the proportion of trials with fixations on target gestures, and (b) the proportion of matched responses as defined above. We employed non-parametric Mann-Whitney tests to analyze the fixation data because the dependent variable, proportions of trials with fixation on gesture, had a skewed distribution with clustering of data at zero. We analyzed the information uptake data using parametric, independent samples analyses of variance and single sample *t*-tests. Throughout, the alpha level for statistical significance is $p = .05$.

Results and discussion

INSERT FIG. 2 A AND 2 B HERE

The proportion of trials in which the addressee fixated gestures were significantly higher in the Speaker-fixation condition ($M = .08$, $SD = .12$) than in the No-speaker-fixation condition ($M = 0$, $SD = 0$), Mann-Whitney, $Z = -2.41$, $p = .016$ (see Fig. 2a). The proportion of trials in which the addressees'

[1] There is no evidence that addressees reversed the directions in the drawings in order to represent the direction as expressed from the speaker's viewpoint. Had addressees been reversing the viewpoints, we would have expected within-subject consistency of such reversals. There is no such consistency in the data, however.

drawn direction and the gesture direction matched (an index of information uptake) was higher in the Speaker-fixation condition (*M* = .86, *SD* = .19) than in the No-speaker-fixation condition (*M* = .63, *SD* = .32), $F(1, 28) = 5.59$, $p = .025$, $\eta^2 = .17$ (see Fig. 2b). Furthermore, the proportion of trials in which addressees' drawing and gestures matched was above chance level (=.50) in the Speaker-fixation condition, one-sample t-test, $t(14) = 7.33$, $p < .001$, but not in the No-speaker-fixation condition, $t(14) = 1.61$, $p = .13$.

The results show that speakers' fixation of their own gestures increase the likelihood of addressees fixating the same gestures. Furthermore, Speaker-fixations also increase the likelihood of addressees' uptake of gestural information, even when it is of little narrative significance and embedded in other directional information. Overall, the combined fixation and uptake findings suggest that speakers' gaze at their own gestures constitute a very powerful attention directing device for addressees influencing both their overt visual attention and their uptake.

Study 2 Location in space and Holds

The second study examines the effect of two physical gestural properties on addressees' overt visual attention to and uptake of information from gestures, namely gestures' location in space (central vs. peripheral) and the presence vs. absence of holds.

Method

Participants

Forty-five new Dutch undergraduate students from Radboud University Nijmegen participated in this study (*M* age = 21, *SD* = 2), 41 females and 4 males. They were paid 5 euros for their participation.

Materials

Three new sets of stimulus videos were selected from the aforementioned corpus using the same criteria as previously, targeting narratives containing different target gestures. For Study 2, the target gestures consisted of gestures performed in central vs. peripheral gesture space with presence vs. absence of hold, with four items in each condition. We used McNeill's (1992) schema to code gesture space. McNeill divides the speaker's gesture space into central and peripheral gesture space, where central space refers to the space in front of the speaker's body, delimited by the elbows, the

shoulders, and the lower abdomen, and peripheral gesture space is everything outside this area. Although McNeill makes more fine-grained distinctions within central and peripheral space, we collapsed all cases of centre-centre and centre space, and all cases of peripheral space, leaving two broad categories: Central and Peripheral. To code for holds (the momentary cessation of a gestural movement), we considered post-stroke holds, that is, cessations of movement after the hand has reached the endpoint of a trajectory of a gesture stroke (Kita et al., 1998; Seyfeddinipur, 2006). The speakers never fixated the target gestures. The mean durations of the target gestures in each condition are summarized in Table 2. As before, descriptions of the animated cartoons used to elicit narratives and the target scenes are provided in Appendix 1, outlines of the spatio-temporal properties of the target gestures across conditions in Appendix 2, and speech co-occurring with target gestures in Appendix 3.

INSERT TABLE 2 HERE

Apparatus, procedure, coding and analysis

Participants were randomly assigned to one of the three conditions (15 participants in each condition): Central Hold, Peripheral No Hold, Peripheral Hold. The data from the No-Speaker-fixation condition from Study 1 was used as the fourth condition, Central No Hold, in the analysis. The apparatus, procedure, coding, and analyses were otherwise identical to Study 1.

Results and discussion

INSERT FIG. 3A AND 3 B HERE

We examined the effect of location and hold on fixations in separate Mann-Whitney tests. The proportion of trials in which the addressees fixated gestures was significantly higher for gestures with Hold (M = .11, SD = .16) than for gestures with No Hold (M = 0, SD = 0), Mann-Whitney, Z = -3.63, p < .001 (see Fig. 3a). In contrast, there was no significant difference in fixation rate between Central gestures (M = .07, SD = .13) and Peripheral gestures (M = .04, SD = .12), Z = -.957, p = .339 (see Fig. 3a). The proportion of trials in which the addressees' drawn directions matched the gesture directions was significantly higher for Central gestures (M = .65, SD .28) than for Peripheral gestures (M = .50, SD = .26), $F(3, 56) = 4.32$, p = .042, $\eta^2 = .072$ (see Fig. 3b). There was no significant effect of Hold, F < 1, and no significant interaction, F < 1. Moreover, the proportion of trials where the

drawings and the gestures matched was only above chance in the Central Hold condition, one sample *t*-test $t(14) = 2.54$, $p = .023$.

The results show that, when location in gesture space and holds were teased apart, only holds increased the likelihood of addressees fixating gestures, whereas the location in gesture space where gestures were produced did not influence addressees' fixations. Moreover, surprisingly, only information conveyed by gestures performed in central, neutral gesture space was taken up and integrated by addressees. However, this result seems to be due to properties of a single item in the Central Hold condition, viz. the "trashcan" item (cf. Appendix 2). 80% of the participants (12/15) had a matched response on this item. Closer inspection of the stimulus showed that the speaker in this stimulus item had looked at another gesture immediately preceding the target gesture. The item therefore inadvertently became similar to the items in the Speaker-fixation condition. When this item was removed from the analysis, uptake for the Central Hold condition dropped to chance level, (M = .59, SD = .32) $t(14) = 1.17$, $p = .262$. Therefore, we conclude that location in gesture space and holds do not modulate the likelihood of information uptake from gestures.

Post-hoc analysis of fixation onset latencies from studies 1 and 2

To examine whether different gestures are fixated for different reasons, we analyzed the fixation onset latencies for those gestures that drew fixations, that is, gestures with speaker-fixations, and gestures with holds (collapsing central and peripheral hold gestures). We measured the time difference between the onset of the relevant cue (speaker-fixation or gestural hold) and the onset of the addressees' fixations of the gestures. Fixation onset latencies for gestures with Speaker-fixations were significantly longer (M = 800 ms, SD = 400 ms) than onset latencies for gestures with Holds (M = 102 ms, SD = 88 ms), Mann-Whitney, $Z = -3.14$, $p = .01$.

These differences suggest that addressees' fixations of gestures are driven by different mechanisms. Onset latencies in the realm of 800 ms indicate that top-down concerns involving higher cognitive mechanisms are driving the fixation behavior. Onset latencies around 100 ms instead suggest that fixations of gestural holds may be bottom-up responses driven by the inner workings of the visual system (cf. Yantis, 2000).

Study 3 Artificial Speaker-Fixations

The unexpected effect of an individual stimulus item in Study 2 raises a general concern that the independent variables may have been confounded with other unknown variables, given that the stimulus gestures differed across the conditions. For instance, the target gesture in the "plank" item had a more complex trajectory than the other items, and the gesture in the "pit" item was performed closer to the face than other target gestures (cf. Appendix 2). Although it is a strength of these studies that they draw on ecologically valid stimuli where the target gestures are naturally produced, dynamic gestures, embedded in discourse and among other gestures, it is important to ascertain that the fixation and uptake findings were not caused by other factors. To test whether speaker-fixations and holds do account for the fixation and uptake data, we therefore created minimal pairs of the most neutral, baseline test items, the centrally produced gestures with no hold or speaker-fixation, by artificially introducing speaker-fixation (Study 3) and holds (Study 4) on these neutral gestures through video editing.

The third study examines the effect of artificially induced speaker-fixations on addressees' overt visual attention to and uptake of information from gestures.

Method

Participants

Fifteen new Dutch undergraduate students from Radboud University Nijmegen participated in this study (*M* age = 21, *SD* = 3), 11 females and 4 males. They were paid 5 euros for their participation.

Materials

Four stimulus items from Study 1, characterized as Central, no Hold, no Speaker-fixation, were used to create four new test items. Each of these was digitally manipulated in Adobe® After Effects® to create minimal pairs of gestures with or without an Artificial speaker-fixation. A section in the video was identified where the speaker's eyes seemed to be directed towards her hands. This set of eyes was cut out and pasted over the real eyes starting at the onset of the stroke of the target gesture and maintained for a duration of 7 frames or 480 ms to form an artificial speaker-fixation (see Fig. 4). The speech stream and the synchronisation between the auditory and the visual parts of the stimulus

videos were not manipulated. This procedure allowed for a speaker-fixation to be imposed on a gesture while keeping the gesture, mouth movements, etc., constant. Although the mean duration of the real speaker-fixations in the original Speaker-fixation condition in Study 1 was 980 ms, the Artificial speaker-fixations had to be shorter (i.e., 480 ms) for the manipulation to align with the shorter gesture strokes of the original Central, No hold, No-speaker-fixated gestures. However, the Artificial speaker-fixations were still within the range of the naturally occurring speaker-fixations. The four digitally manipulated items constitute the Artificial speaker-fixation condition.

INSERT FIG. 4 HERE

Apparatus, procedure, coding, and analysis

These were identical to Study 1. The data from the Artificial speaker-fixation condition were compared to the data from the original No-speaker-fixation condition reported in Study 1, henceforth the Control condition.

Results and discussion

INSERT FIG. 5A AND 5 B HERE

There was no significant difference between the proportion of fixated trials in the Artificial speaker-fixation condition (M = .03, SD = .09) and the Control condition (M = 0, SD = 0), Mann-Whitney, Z = -1.44, p = .15. Furthermore, there was no significant difference in the proportion of trials with uptake in the Artificial speaker-fixation condition (M = .71, SD = .31) and the Control condition (M = .63, SD = .32), $F(1, 28) < 1$, p = .536. However, the proportion of trials with uptake was reliably above chance (.50) in the Artificial speaker-fixation condition, one-sample t-test, $t(14) = 2.58$, p = .022, but not in the Control condition, $t(14) = 1.61$, p = .13.

Both for fixation and uptake, the differences between the Artificial speaker-fixation and Control condition went in the same direction as predicted by the results from Study 1, but neither difference reached statistical significance. The comparison against chance nevertheless indicated uptake above chance from gestures in the Artificial speaker-fixation, in line with the effect of natural speaker-fixations on uptake found in Study 1.

There are two possible explanations for the weaker fixation results in this study than in Study 1. First, for practical reasons the duration of the Artificial speaker-fixations was significantly shorter

(480 ms) than the average authentic duration (*M* = 980 ms, *SD* = 414 ms), Mann Whitney, *Z* = -2.46, *p* = .014. It is likely that the longer the speaker's gaze on a gesture, the more likely the addressee is to also look at it. A closer inspection of the results from Study 1 revealed a tendency for longer speaker-fixations to yield more addressee-fixations than shorter ones. Second, the duration of the gesture stroke itself may also have played a role. Again, the average duration of the authentic gestures with speaker-fixations was significantly longer (*M* = 2,410 ms, *SD* = 437 ms) than the strokes of the control gestures on which we imposed the Artificial speaker-fixation (*M* = 1,310 ms, *SD* = 305), Mann Whitney, *Z* = -2.31, *p* = .021. However, the influence of the stroke duration is debatable because peripheral gestures, which by virtue of their spatial expanse also have longer duration than centrally produced gestures, did not draw fixations. Indirectly, then, these findings suggest that speakers' fixations of their own gestures increase the likelihood of addressees' shifting overt visual attention to gestures, and this effect is enhanced the longer the speakers' fixation.

Study 4 Artificial Holds

The fourth study examines the effect of artificially induced gestural holds on addressees' overt visual attention to and uptake of information from gestures.

Method

Participants

Fifteen new Dutch undergraduate students from Radboud University Nijmegen participated in this study (*M* age = 21, *SD*= 2), 11 females and 4 males. They were paid 5 euros for their participation.

Materials

As in Study 3, the four items characterized as Central, no Hold, no Speaker-fixation from Study 1 were digitally manipulated in Adobe® After Effects® to create minimal pairs of gestures with or without an artificial hold. The hand shape of the last frame of the original target gesture stroke was isolated and then pasted and maintained over the original retraction phase of the gesture for 5 frames or 200 ms, using the same procedure as illustrated in Fig. 4. The pasted hand shape was then moved spatially for a number of transition frames to fit onto the original, underlying location of the hand without creating a jerky movement. As before, speech and the synchronisation between the auditory

and the visual parts of the stimulus videos were not manipulated. The procedure allowed head and lip movements to remain synchronized with speech. Note that the original mean duration of natural holds (central and peripheral) was 575 ms. As in Study 3, a shorter hold duration (i.e., 200 ms), although still within the range of naturally occurring holds, was chosen to avoid too large a spatial discrepancy between the location of the artificially held hand, and the underlying retracted gesture. Such a discrepancy would have made the manipulation impossible to conceal. The four digitally manipulated items constitute the Artificial hold condition.

Apparatus, procedure, coding, and analysis

These were identical to Study 1. The data from the Artificial hold condition were compared to the data from the original No-speaker-fixation condition reported in Study 1, henceforth the Control condition.

Results and discussion

INSERT FIG. 6A AND 6B HERE

The proportion of fixated trials was significantly higher in the Artificial-hold condition (M = .08, SD = .12) than in the Control condition (M = 0, SD = 0), Mann-Whitney, Z = -2.41, p = .016. There was no significant difference in uptake between the Artificial hold (M = .59, SD = .35) and the Control conditions (M = .63, SD = .32), $F(1, 28) < 1$, p = .75. Moreover, the proportion of matched trials was at chance both in the Artificial hold condition, one-sample t-test, $t(14) = 1.03$, p = .319, and in the Control condition, $t(14) = 1.61$, p = .13.

To summarize, both the fixation and the uptake findings from Study 2 were replicated. Holds made addressees more likely to fixate speakers' gestures, but they did not seem to contribute to uptake of gestural information.

Post-hoc analysis of fixation onset latencies from Studies 3 and 4

As in Studies 1 and 2, we measured the time difference between the onset of the relevant cue (Artificial speaker-fixation or Artificial hold) and the onset of the addressees' fixations of the gestures. Fixation onset latencies for Artificial speaker-fixations were generally longer (M = 100 ms, SD = 85 ms) than onset latencies for gestures with Artificial holds (M = 40 ms, SD = 0 ms), although there

were too few data points to undertake a statistical analysis. These differences in fixation onset latencies nevertheless display the same trends as for natural speaker-fixations and holds.

Post-hoc analyses of the relationship between addressees' fixations and uptake

One of the research questions concerned the relationship between fixations and uptake of gestural information. To address this issue, we examined whether information uptake differed between fixated versus non-fixated gestures.

All trials from Studies 1 through 4 were combined for this analysis to compare the likelihood of uptake in a within-subject comparison for those 20 participants who had codable trials with and without addressee-fixation (n = 15 from the Hold conditions, n = 5 from the Speaker-fixation condition). The proportion of matched responses was not significantly different between trials with addressee-fixation (M = .70, SD = .47) and without addressee-fixation (M = .62, SD .42), $F(1, 19) < 1$, p = .576.

When the data were broken down according to the two cue types (speaker-fixation and holds), the proportion of matched responses in the two types of trials were still not significantly different from each other: uptake from speaker-fixated trials with addressee-fixation (M = .60, SD = .55) did not differ from speaker-fixated trials without addressee-fixations (M = .40, SD = .55), $F(1, 4) < 1$, p = .621. Similarly, uptake from hold-trials with addressee-fixation (M = .73, SD = .46) did not significantly differ from hold-trials without addressee-fixations (M = .69, SD = .36), $F(1, 14) < 1$, p = .783. Thus, there is little evidence that addressees' fixations of gestures are associated with uptake of the gestural information.

General discussion

This study investigated what factors influence addressees' overt visual attention to (direct fixation of) gestures and their uptake of gestural information, focusing on one social factor, namely speakers' gaze at their own gestures, and two physical properties of gestures, namely their place in gesture space and the effect of gestural holds. We also examined the relationship between addressees' fixations of gesture and their uptake of gestural information. We explored these issues drawing on examples of natural gestures expressing directional information left or right embedded in narratives.

The results concerning fixations of gestures can be summarized in four points. First, in line with previous studies (Gullberg & Holmqvist, 1999; 2006; Nobe et al., 1998; 2000), addressees looked directly at very few gestures. Second, they were more likely to fixate gestures that speakers themselves had first fixated (speaker-fixation) than others. This tendency held also for gestures with artificially introduced speaker-fixations, although it did not reach statistical significance. Moreover, addressees were also more likely to fixate gestures with a post-stroke hold than gestures without. This held both for natural and artificial holds. Third, contrary to expectation, the locations of gestures in gesture space (central vs. peripheral) did not affect addressees' tendency to fixate gestures. Fourth, the onset latency of fixations differed across gesture types. Fixations of gestures with post-stroke holds had shorter onset latencies than those of speaker-fixated gestures, suggesting that addressees look at different gestures for different reasons. Holds are fixated for bottom-up reasons and speaker-fixated gestures for top-down reasons.

There were three main findings concerning uptake of gestural information. First, addressees did not generally process and retain directional gestural information uniformly in all situations. Second, addressees were more likely to retain the directional information in gesture when speakers themselves had first fixated the gesture than when they had not. Third, there was no evidence that the presence or absence of post-stroke holds or the location in gesture space affected information uptake when an item with inadvertent speaker-fixation on a previous gesture was removed.

Finally, regarding the relationship between addressees' fixations and their information uptake, a post-hoc analysis based on the pooled data from all the studies showed no evidence that addressees' information uptake from gestures was associated with their fixations of gestures.

In previous studies of fixation behavior towards gestures (Gullberg & Holmqvist, 1999; 2006; Nobe et al., 1998; 2000), the three factors investigated here have been conflated. The current study demonstrates the individual contributions of two of these factors, the social factor Speaker-fixation, and one of the physical factors, namely post-stroke Holds. It also shows that the other physical property, location in gesture space, does not matter. Moreover, the data suggest that addressees fixate different gestures for different reasons. The effect of speaker-fixations on addressees' gaze behavior is compatible with suggestions that humans automatically orient to the target of an interlocutor's gaze (e.g., Driver et al., 1999). Notice, however, that speaker-fixations only lead to *overt*

gaze-following or addressee-fixations 8% of the time (Study 1; this rate is similar to that reported in Gullberg & Holmqvist, 2006). This suggests that overt gaze-following is not an automatic process but rather a socially mediated process, where social norms for maintaining mutual gaze is the default, and overt gaze-following to a gesture rather signals social alignment (Gullberg & Holmqvist, 2006). The longer onset latencies of addressee-fixations following speaker-fixations support this notion, as longer onset latencies are likely to reflect top-down processes such as social alignment. In contrast, addressees' tendency to fixate gestures with holds may result from holds constituting sudden change in the visual field, or from holds challenging peripheral vision, which is best at motion detection. With no motion to detect, an addressee needs to shift gaze and fixate the gesture in order to extract any information at all. Both accounts assume that fixations to holds should be driven by low-level bottom-up processes. The fixation onset latency data support this account. The very short fixation onset latencies to gestural holds suggest a stimulus-driven response by the visual system.

The uptake results strongly suggest that all gestural information is not uniformly processed and integrated. That is, it is not the case that addressees cannot help but integrate gesture information (e.g., Cassell et al., 1999). The findings indicate that directional gesture information is not well integrated in the absence of any further highlighting, which is in line with Beattie & Shovelton's results (1999a; 1999b) showing that directional gesture information is less well retained than information about size and location. However, the social factor (speaker-fixation) modulated uptake of such information such that addressees retained gestural information about direction when speakers had looked at gestures first. The physical properties of gestures played no role for uptake.

The comparison of fixation behavior and uptake showed that uptake from gestures was greatest in a condition where gestures were first fixated by the speaker (86%), although the addressees only fixated these gestures 8 % of the time (Exp.1). Addressees' attention to gestures was therefore mostly covert. It seems that addressees' uptake of gestural information may be independent of whether they fixate the target gesture or not provided that speakers have highlighted the gesture with their gaze first. Although this finding must be consolidated in further studies, it suggests that although overt gaze-following is not automatic, covert attention shift to the target of a speaker's gaze location may well be, allowing fine-grained information extraction in human interaction.

An important implication of these findings for face-to-face communication is that addressees' gaze is multifunctional and not necessarily a reliable index of attention locus, information uptake or comprehension. Addressees clearly look at different things for different reasons and one cannot assume that overt visual attention to something – like a gesture with a post-stroke hold – necessarily implies that the target is processed for information. This is primarily a caveat to studies on face-to-face interaction where a mono-functional view of gaze is often in evidence. In interaction addressees will typically maintain their gaze on the speaker's face as a default. Addressees' overt gaze shift may be an act of social alignment to show speakers that they are attending to the their focus of attention (e.g., their gestures), rather than an act of information seeking which is often possible through peripheral vision. Conversely, the fact that addressees' attention to gestures is not uniform means that speakers can manipulate it, highlighting gestures strategically as a relevant channel of information in various ways. For instance, speakers can use spoken deictic expressions such as 'like this' to draw direct attention to gestures, or use their own gaze (speaker-fixation) to do the same thing visually. Other possibilities include distributing information across the modalities in complementary fashion, such as saying 'this big' and indicating size in gesture (also an example of a deictic expression).

This study has raised a number of further issues to explore. An important question is what other factors might affect addressees' attention to gestures. Other physical properties of gestures are likely candidates, such as gestures' size and duration, the difference between simple and complex movement trajectories, etc. A social factor that is likely to play a role concerns the knowledge shared by participants, also known as common ground (e.g., Clark & Brennan, 1991; Clark, Schreuder, & Buttrick, 1983). The more common ground is shared between interlocutors, the more reduced the gestures tend to be in form and the less likely information is to be expressed in gesture at all (e.g., Gerwing & Bavelas, 2004; Holler & Stevens, 2007; Holler & Wilkin, 2009). This opens for the possibility that attention to gesture is modulated by discourse factors with heightened attention to gesture when information is new and first introduced, and mitigated attention as information grows old. Another discourse effect concerns the relevance of information. The information probed in this study was deliberately chosen to be unimportant to the gist of the narratives. It is important to test whether these findings generalize to discursively vital information.

To conclude, this study has taken a first step towards a more fine-grained understanding of how and when addressees take gestural information into account and of the factors that govern attention allocation – both overt and covert – to such gestural information.

References

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Argyle, M., & Graham, J. A. (1976). The Central Europe experiment: Looking at persons and looking at things. *Journal of Environmental Psychology and Nonverbal Behavior, 1*, 6-16.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication, September 2002*, 566-580.

Beattie, G., & Shovelton, H. (1999a). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? *Semiotica, 123*, 1-30.

Beattie, G., & Shovelton, H. (1999b). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology, 18*, 438-462.

Beattie, G., & Shovelton, H. (2005). Why the spontaneous images created by the hands during talk can help make TV advertisements more effective. *British Journal of Psychology, 96*, 21-37.

Bruce, V., & Green, P. (1985). *Visual perception. Physiology, psychology and ecology* (2nd 1987 ed.). Hillsdale, NJ: Erlbaum.

Cassell, J., McNeill, D., & McCullough, K.-E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition, 7*, 1-33.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, H. H., & Brennan, S. A. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine & S. D. Teasley (Eds.), *Perspectives on socially shared cognition*. Washington: APA Books.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior, 22*, 245-258.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition, 6*, 509-540.

Duncan, S. J. (1973). Toward a grammar for dyadic conversation. *Semiotica, 9*, 29-47.

Fehr, B. J., & Exline, R. V. (1987). Social visual interaction: A conceptual and literature review. In A. W. Siegman & S. Feldstein (Eds.), *Nonverbal behavior and communication* (pp. 225-326). Hillsdale, NJ: Erlbaum.

Fornel, M. (1992). The return gesture: Some remarks on context, inference, and iconic gesture. In P. Auer & A. di Luzio (Eds.), *The contextualization of language* (pp. 159-176). Amsterdam: Benjamins.

Gerwing, J., & Bavelas, J. B. (2004). Linguistic influences on gesture's form. *Gesture, 4*, 157-195.

Gibson, J. J., & Pick, A. D. (1963). Perception of another person's looking behavior. *American Journal of Psychology, 76*, 386-394.

Goodwin, C. (1981). *Conversational organisation: Interaction between speakers and hearers*. New York: Academic Press.

Gullberg, M. (1998). *Gesture as a communication strategy in second language discourse. A study of learners of French and Swedish*. Lund: Lund University Press.

Gullberg, M., & Holmqvist, K. (1999). Keeping an eye on gestures: Visual perception of gestures in face-to-face communication. *Pragmatics & Cognition, 7*, 35-63.

Gullberg, M., & Holmqvist, K. (2006). What speakers do and what listeners look at. Visual attention to gestures in human interaction live and on video. *Pragmatics and Cognition, 14*, 53-82.

Heath, C. (1986). *Body movement and speech in medical interaction*. Cambridge: Cambridge University Press.

Holler, J., & Beattie, G. (2003). How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica, 146*, 81–116.

Holler, J., & Stevens, R. (2007). The effect of common ground on how speakers use gesture and speech to represent size information. *Journal of Language and Social Psychology, 26*, 4-27.

Holler, J. & Wilkin, K. (2009). Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task. *Language and Cognitive Processes*, *24,* 267-289.

Kelly, S. D., Barr, D. J., Breckinridge Church, R., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language, 40*, 577-592.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207-227). The Hague: Mouton.

Kendon, A. (1990). *Conducting interaction*. Cambridge: Cambridge University Press.

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

Kita, S. (1996). Listeners' up-take of gestural information. *MPI Annual report, 1996*, 78.

Kita, S., Van Gijn, I., & Van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and Sign Language in Human-Computer interaction* (pp. 23-35). Berlin: Springer.

Kleinke, C. L. (1986). Gaze and eye contact: A research review. *Psychological Bulletin, 100*, 78-100.

Langton, S. R. H., & Bruce, V. (2000). You must see the point: Automatic processing of cues to the direction of social attention. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 747-757.

Langton, S. R. H., O'Malley, C., & Bruce, V. (1996). Actions speak no louder than words: Symmetrical cross-modal interference effects in the processing of verbal and gestural information. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1357-1375.

Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences, 4*, 50-59.

Latham, K., & Whitaker, D. (1996). A comparison of word recognition and reading performance in foveal and peripheral vision. *Vision Research, 37*, 2665-2674.

Maki, R. Grandy, C. A. & Hauge, G. (1979). Why is telling right from left more difficult than telling above from below? *Journal of Experimental Psychology: Human Perception and Performance*, 5, 52-67.

McNeill, D. (1992). *Hand and mind. What the hands reveal about thought*. Chicago: University of Chicago Press.

McNeill, D., Cassell, J., & McCullough, K.-E. (1994). Communicative effects of speech mismatched gestures. *Research on Language and Social Interaction, 27*, 223-237.

Melcher, D., & Kowler, E. (2001). Visual scene memory and the guidance of saccadic eye movements. *Vision Research, 41*, 3597-3611.

Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture, 4*, 119-141.

Moore, C., & Dunham, P., J. (Eds.). (1995). *Joint attention*. Hillsdale, NJ: Erlbaum.

Nobe, S., Hayamizu, S., Hasegawa, O., & Takahashi, H. (1998). Are listeners paying attention to the hand gestures of an anthropomorphic agent? An evaluation using a gaze tracking method. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and Sign Language in human-computer interaction* (pp. 49-59). Berlin: Springer.

Nobe, S., Hayamizu, S., Hasegawa, O., & Takahashi, H. (2000). Hand gestures of an anthropomorphic agent: Listeners' eye fixation and comprehension. *Cognitive Studies. Bulletin of the Japanese Cognitive Science Society, 7*, 86-92.

Özyürek, A., Willems, R. M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience, 19*, 605-616.

Rogers, W. T. (1978). The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances. *Human Communication Research, 5*, 54-62.

Rossano, F., Brown, P., & Levinson, S. C. (in press). Gaze, questioning and culture. In J. Sidnell (Ed.), *Conversation Analysis: Comparative perspectives*. Cambridge: Cambridge University Press.

Seyfeddinipur, M. (2006). *Disfluency: Interrupting speech and gesture.* Unpublished PhD diss, Radboud University, Nijmegen.

Streeck, J. (1993). Gesture as communication I: Its coordination with gaze and speech. *Communication Monographs, 60*, 275-299.

Streeck, J. (1994). Gesture as communication II: The audience as co-author. *Research on Language and Social Interaction, 27*, 239-267.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Tomasello, M. & Todd, J. (1983). Joint attention and lexical acquisition style. *First Language, 4*, 197-211.

Watson, O. M. (1970). *Proxemic behavior: A cross-cultural study*. The Hague: Mouton.

Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology, 42*, 654-667.

Yantis, S. (1998). Control of visual attention. In H. Pashler (Ed.), *Attention* (pp. 223-256). Hove: Psychology Press Ltd.

Yantis, S. (2000). Goal-directed and stimulus-driven determinants of attentional control. In S. Monsell & J. Driver (Eds.), *Attention and performance XVIII* (pp. 73-103). Cambridge, MA: MIT Press.

Table 1. Mean duration (ms) of target gestures with and without Speaker-fixation

	Mean duration ms (SD)
Speaker-fixated gesture stroke	2,410 (437)
Speaker-fixation on gesture	980 (414)
No-speaker-fixated gesture stroke	1,310 (305)

Table 2. Mean duration (ms) of Central and Peripheral target gestures with and without Holds

	Mean duration ms (SD)
Central stroke No Hold	1,385 (493)
Peripheral stroke No Hold	1,370 (696)
Central stroke + Hold	1,460 + 570 (677 + 428)
Peripheral stroke + Hold	1,490 + 580 (962 + 334)

Figure captions

Figure 1. Example of a match between the (gesture) direction seen on the stimulus video (left) and the direction indicated as a response on the subsequent drawing task (left).

Figure 2. (a) Mean proportion of fixated target gestures in the Speaker-fixation and No-speaker-fixation conditions, (b) mean proportion of matched responses in the Speaker-fixation and No-speaker-fixation conditions, i.e. responses where the direction in the drawing matched that of the target gesture (chance = .5). Error bars indicate SDs.

Figure 3. (a) Mean proportion of fixated target gestures across the four conditions Location (Central vs. Peripheral) and Hold (presence vs. absence), (b) mean proportion of matched responses across the four conditions Location (Central vs. Peripheral) and Hold (presence vs. absence), i.e. responses where the direction in the drawing matched that of the target gesture (chance = .5). Error bars indicate SDs.

Figure 4. Example of the minimal pair creation (Artificial Speaker-fixation) used in Study 3. The top panel shows example frames of the original target gesture. The bottom panel shows a set of eyes seemingly directed towards the target gesture pasted over the original eyes for a certain number of frames

Figure 5. (a) Mean proportion of fixated target gestures in the Control and Artificial speaker-fixation conditions, (b) mean proportion of matched responses in the Control and Artificial speaker-fixation conditions, i.e. responses where the direction in the drawing matched that of the target gesture (chance = .5). Error bars indicate SDs.

Figure 6. (a) Mean proportion of fixated target gestures in the Control and Artificial hold conditions, (b) mean proportion of matched responses in the Control and Artificial hold conditions, i.e. responses where the direction in the drawing matched that of the target gesture (chance = .5). Error bars indicate SDs.

Figure 1

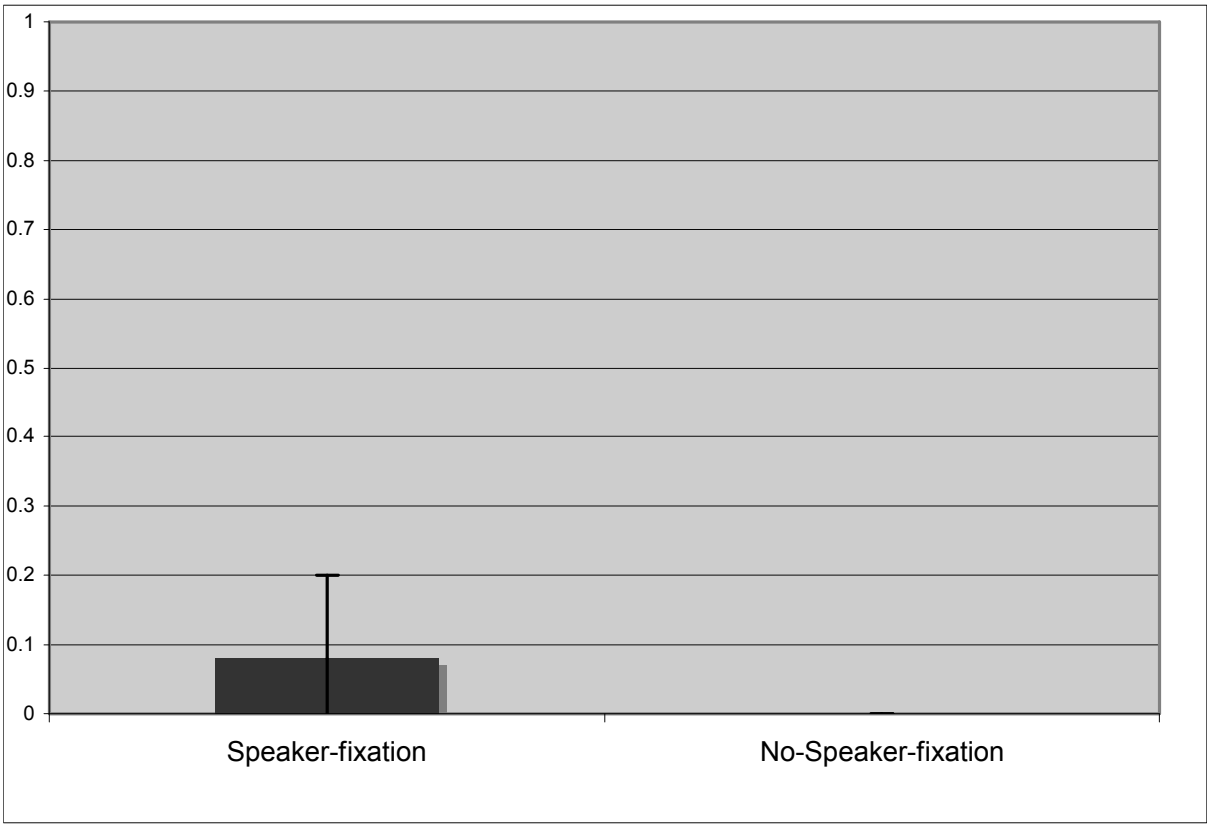

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Speaker-fixation
No-Speaker-fixation

Figure 2a

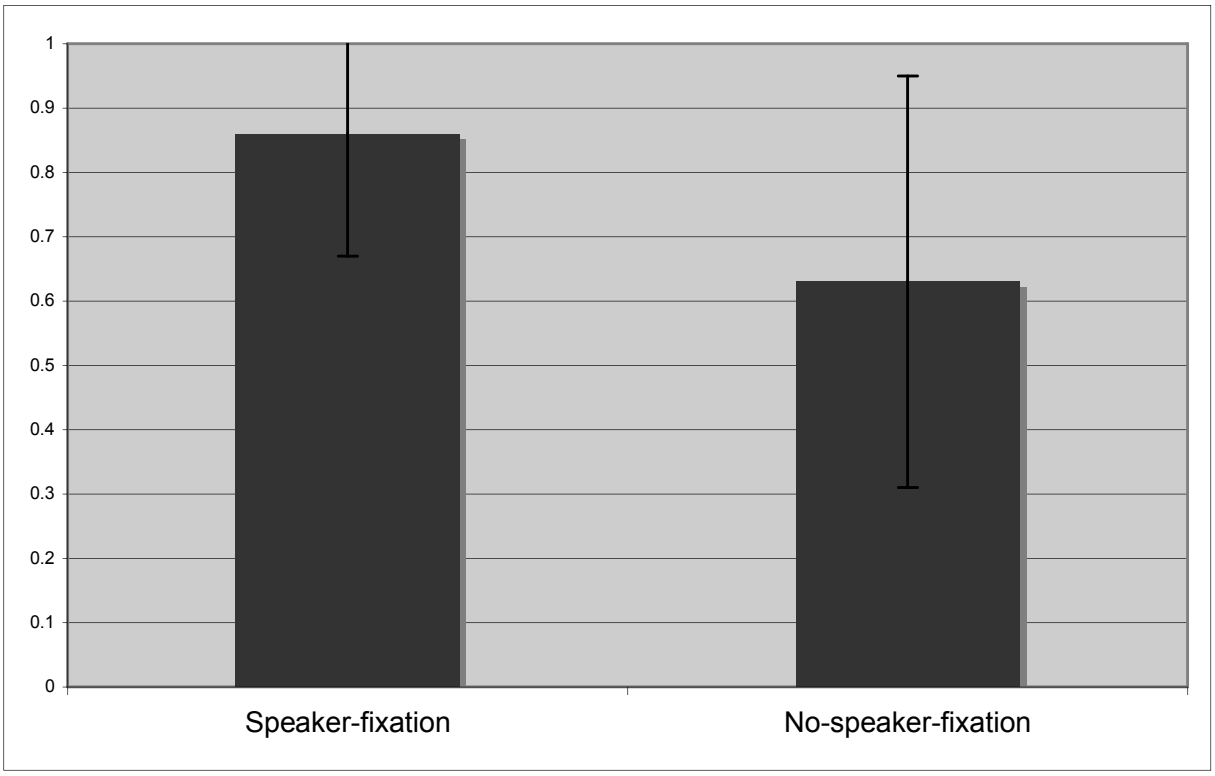

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Speaker-fixation
No-speaker-fixation

Figure 2b

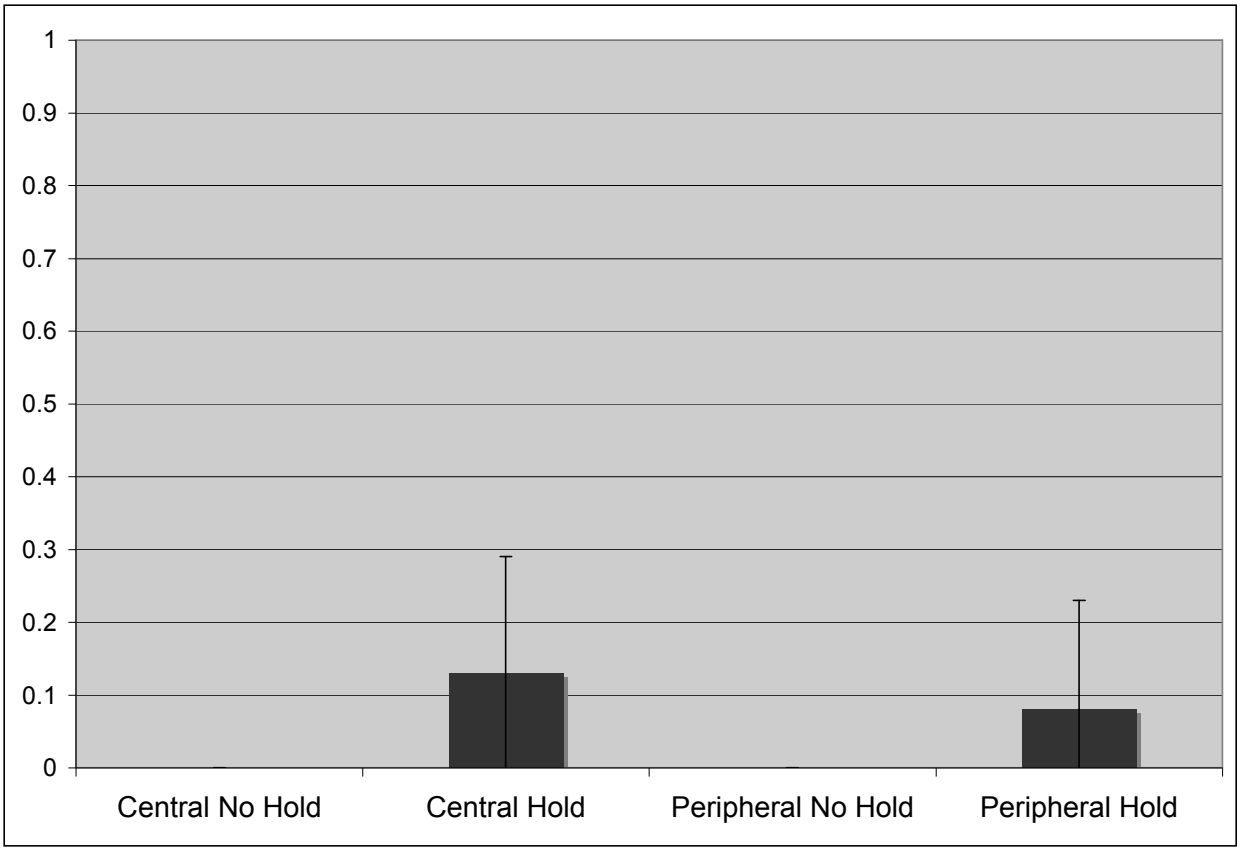

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Central No Hold
Central Hold
Peripheral No Hold
Peripheral Hold

Figure 3a

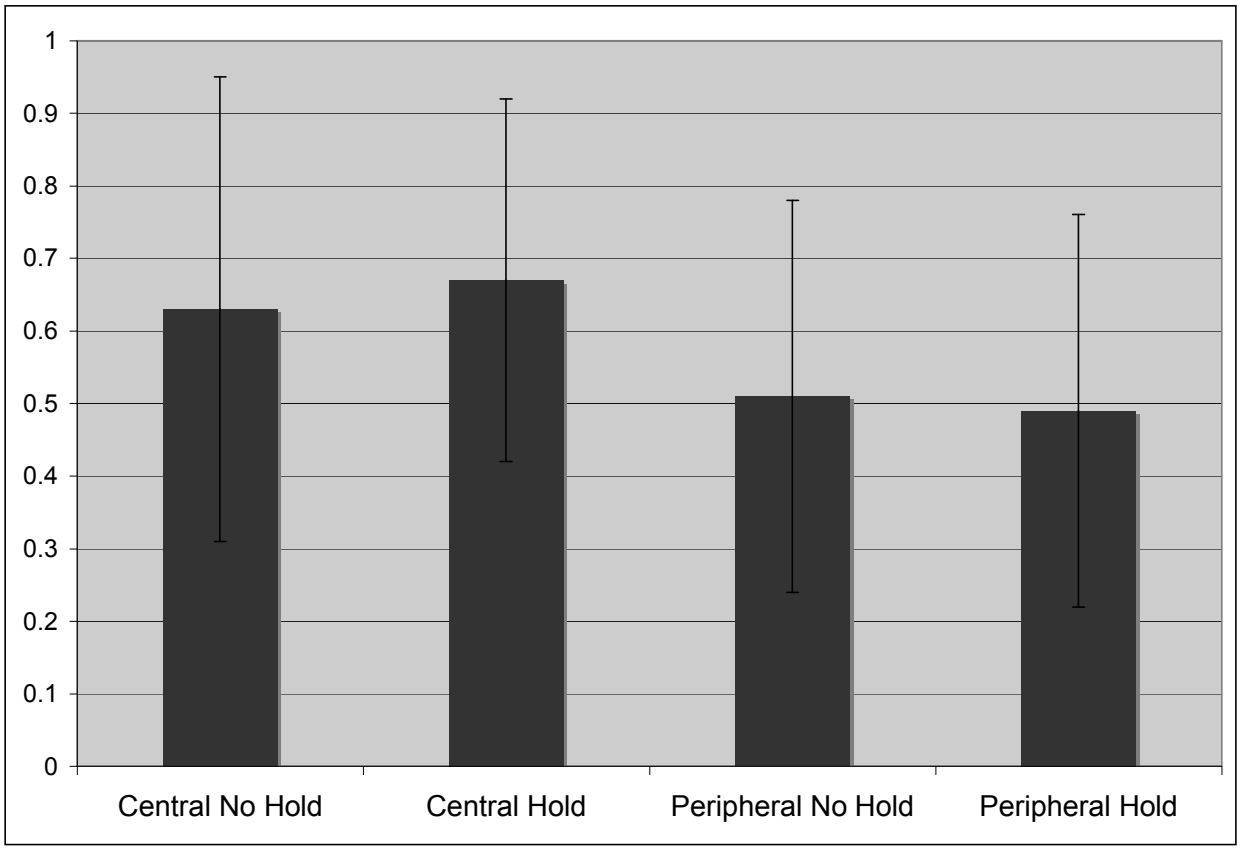

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Central No Hold
Central Hold
Peripheral No Hold
Peripheral Hold

Figure 3b

00:12:18
00:12:19
00:12:20
00:12:21
00:12:22
original frames
added layer
00:12:18
00:12:19
00:12:20
00:12:21
00:12:22
manipulated frames
onset manipulation
onset target gesture

Figure 4

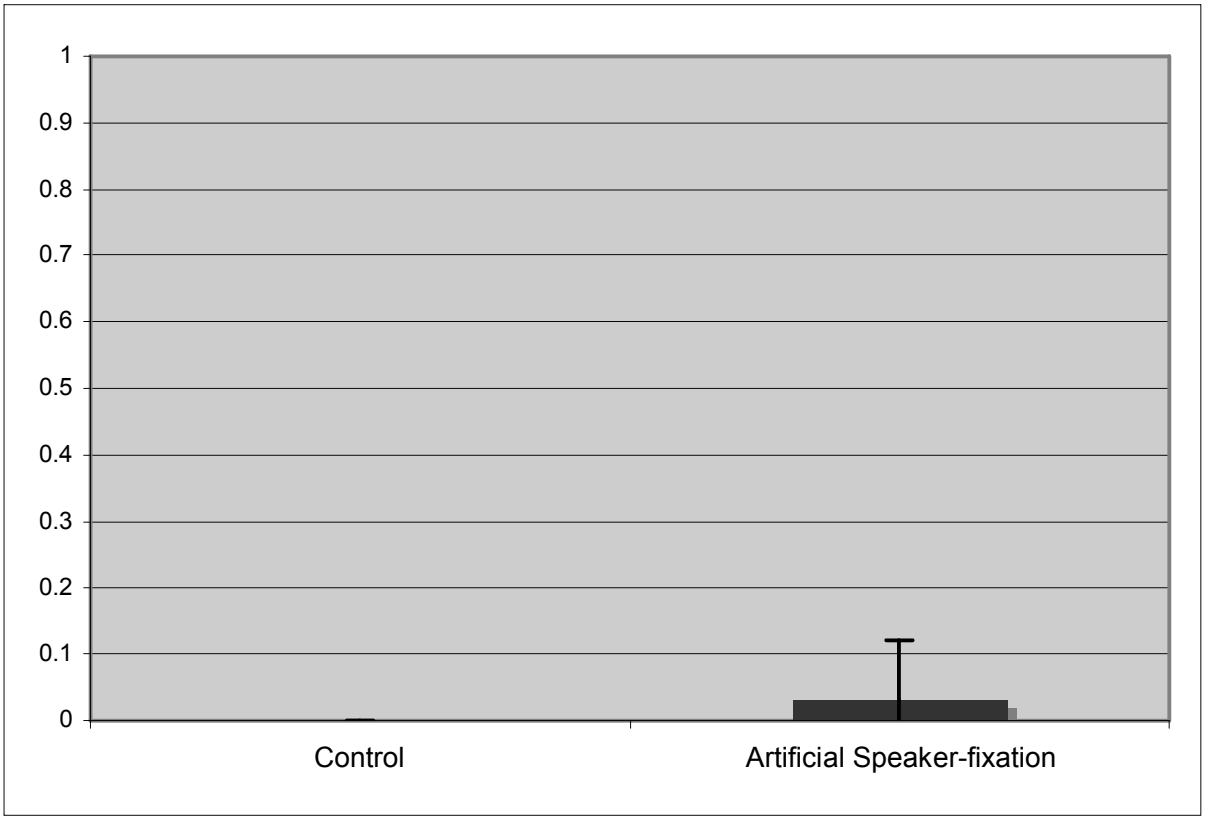

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Control
Artificial Speaker-fixation

Figure 5a

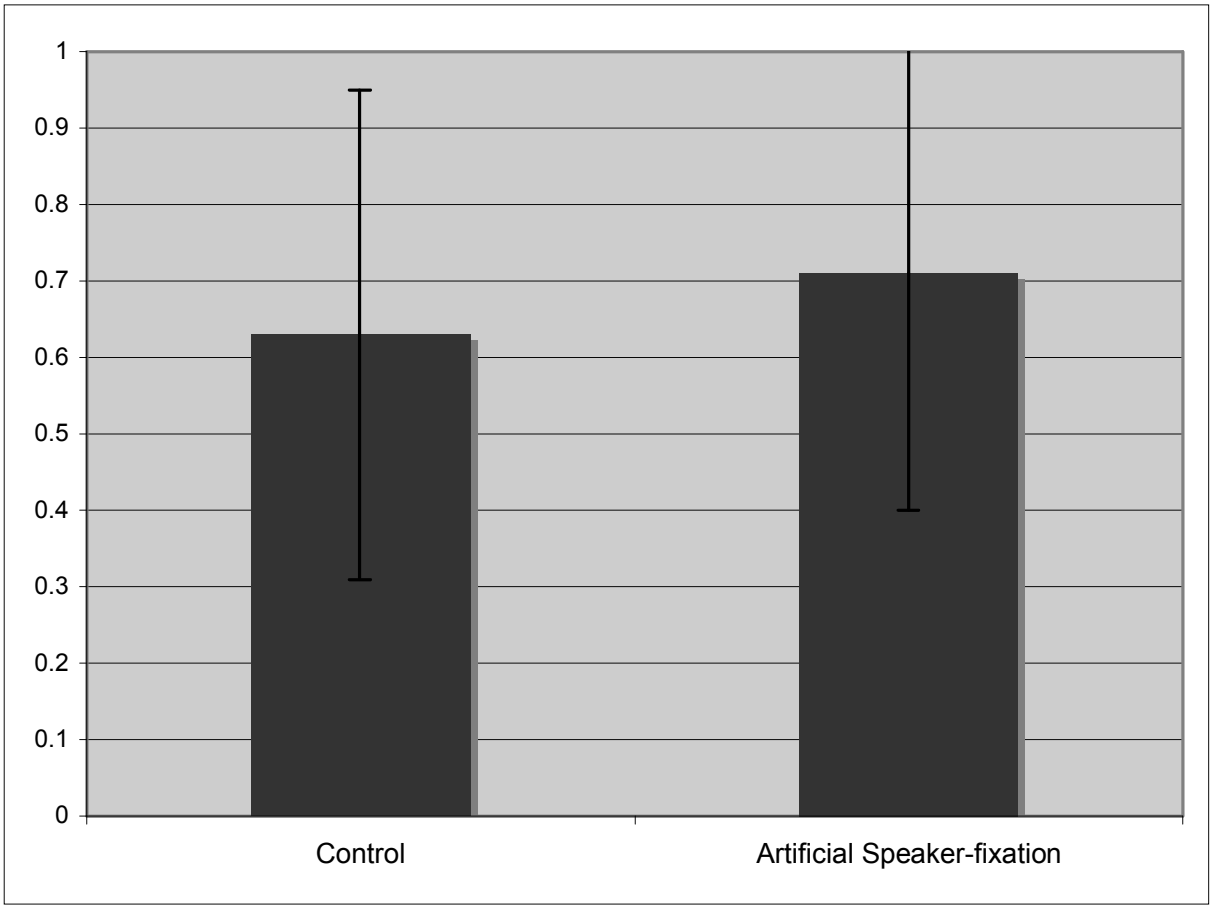

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Control
Artificial Speaker-fixation

Figure 5b

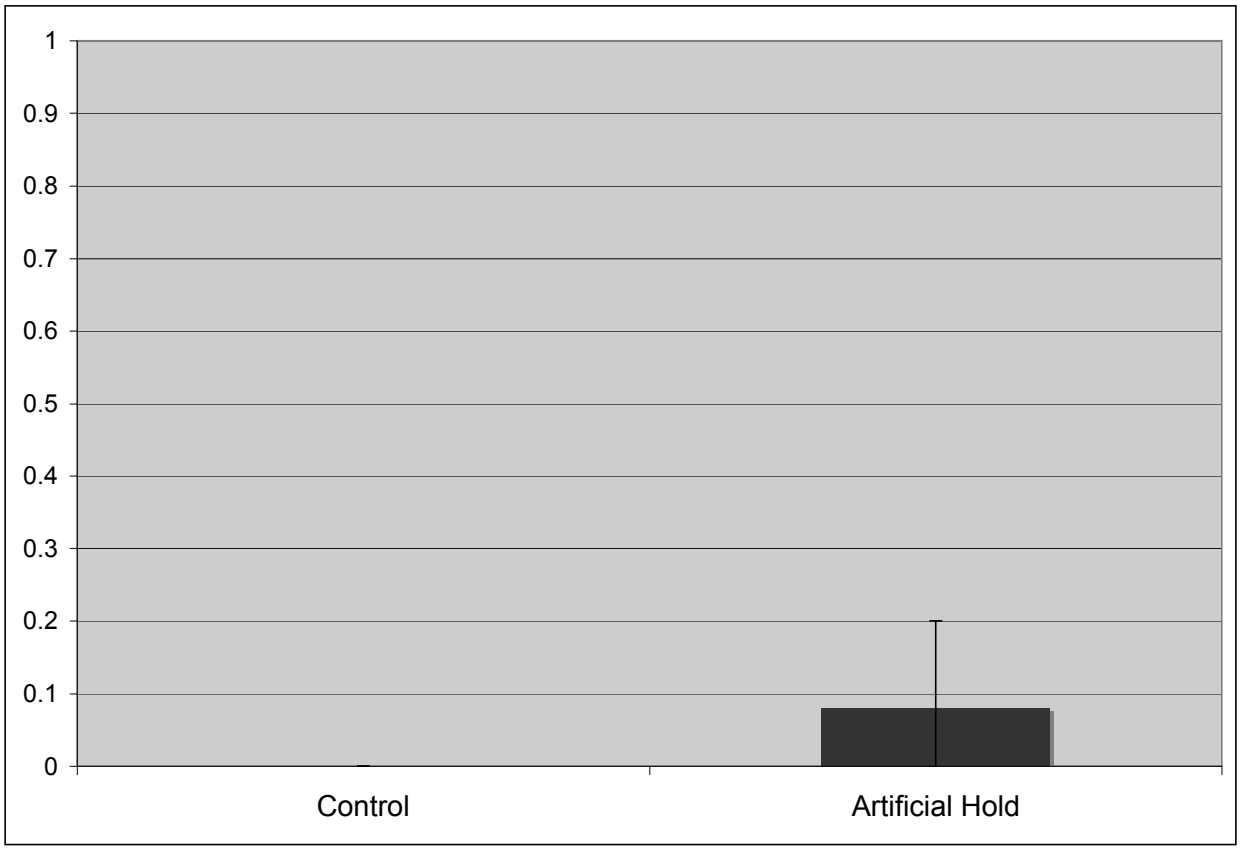


Figure 6a

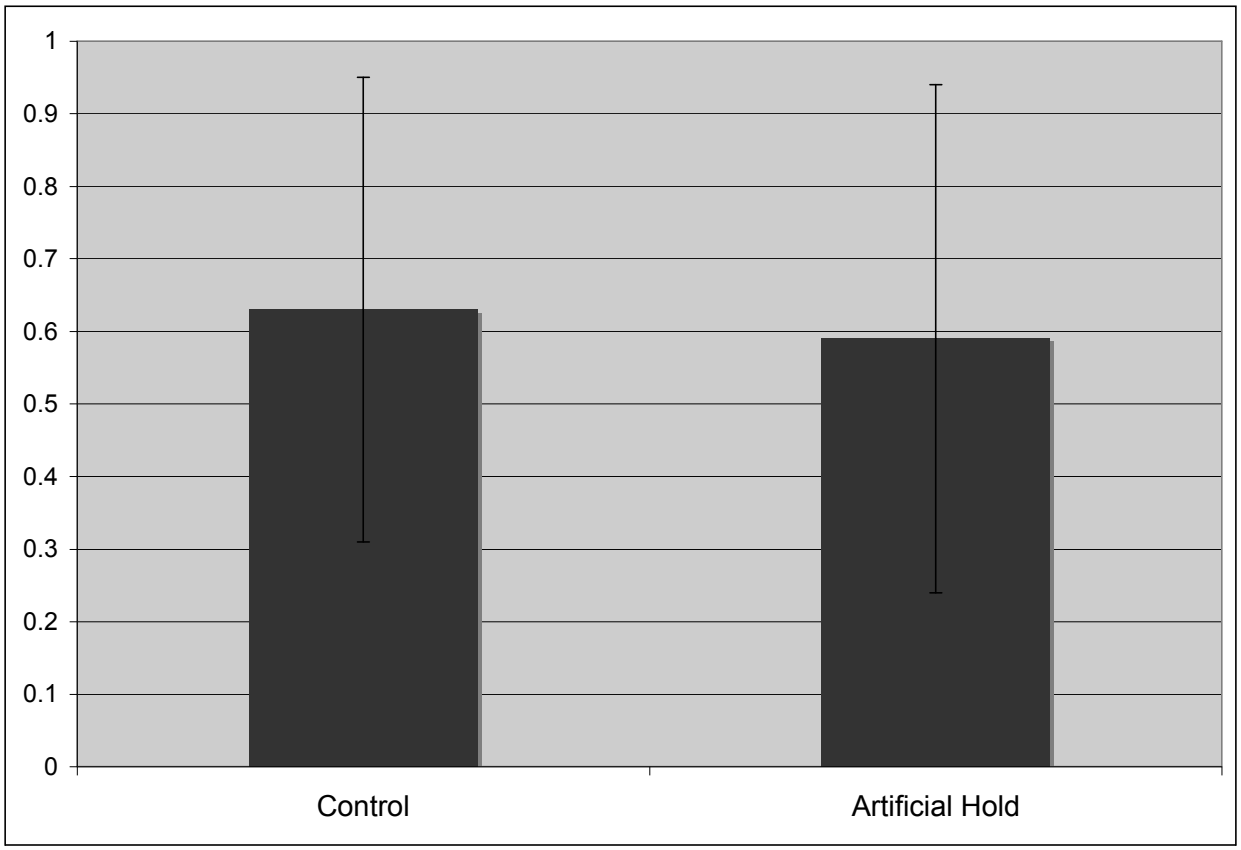


Figure 6b

Appendix 1 Scenes Described in the Stimulus Items

Scene 1: wave

A mouse in a rowing boat at sea tries to row, but a wave is preventing it from making any progress. The mouse makes two holes in the bottom of the boat and sticks its feet through these holes. It moves by walking on the bottom of the sea.

Scene 2: garbage

A cat is in one building, a bird in another. The cat looks at the bird through binoculars, then runs down out of the building, crosses the street and runs into the bird's building. The cat is thrown out of the building and lands on a pile of garbage.

Scene 3: plank

There is a large log with a plank on top of it, forming a springboard. A cat stands on one end of the plank and throws a weight onto the other end. The cat is launched upward, reaching a bird at the top of the building. The cat catches the bird, and comes down landing on the plank again. The weight shoots up, and as the cat is running away, the weight drops on his heads.

Scene 4: trashcan

A mouse is eating a banana and throws the skin away in the trashcan. The skin comes out again and lands on the mouse's face. He throws it in the trashcan again, but the same thing happens again. The mouse looks in the trashcan, throws the banana skin in again, and turns the trashcan upside down. As the mouse walks away the trashcan follows him. The feet of an elephant are sticking out from under the trashcan.

Scene 5: bowling ball

A cat is pacing outside a building spying on a bird in one of the top windows. The cat climbs up inside a drainpipe to catch to the bird. When the bird sees the cat, he throws a bowling ball into the drainpipe. The cat swallows the bowling ball, and then comes shooting out of the drainpipe, rolling onto the street.

Scene 6: pole

A mouse jumps over a bar using a long pole (pole vault). He lands on his face and his body moves up and down like a spring.

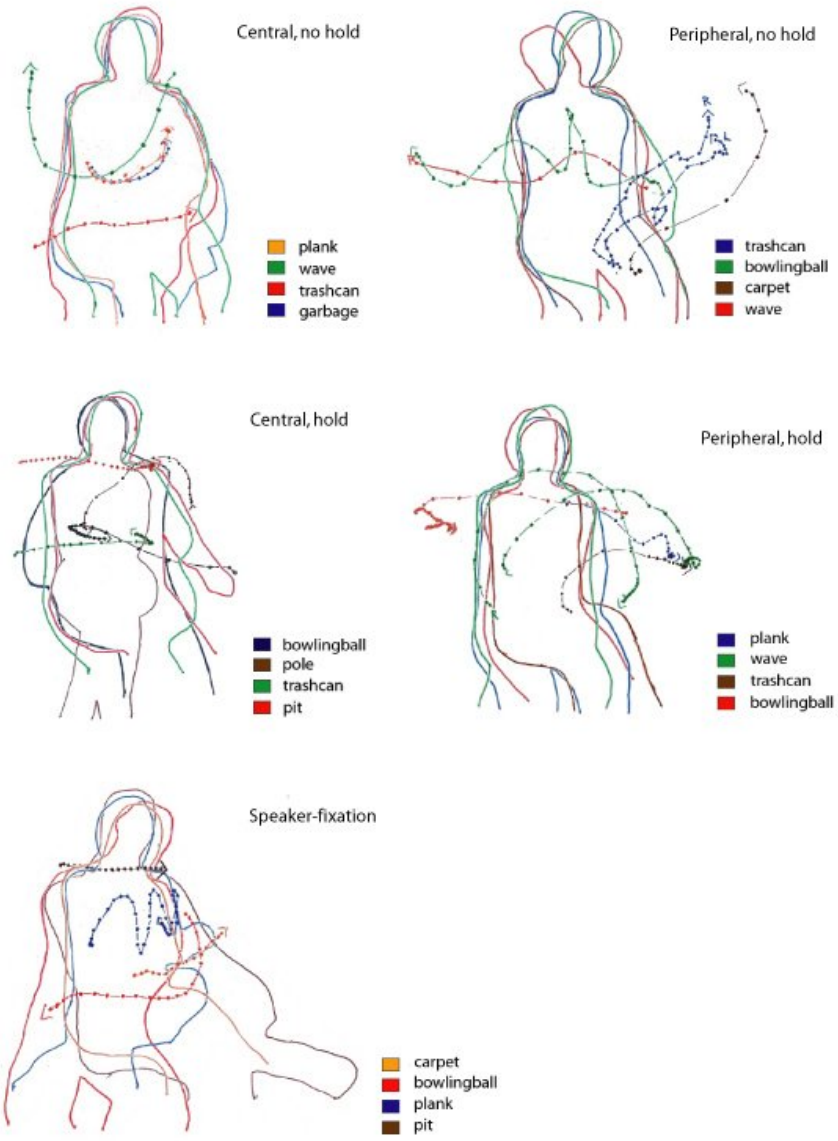
Scene 7: pit

A mouse is walking towards the edge of a large pit. He tries to jump across, but fails and falls into the pit.

Scene 8: carpet

A mouse and an elephant are walking along on a carpet. The elephant stumbles over some folds in the carpet. The mouse shows him how to flatten the folds by stamping on them but the elephant cannot do it. The mouse ‘winds up’ the elephant by turning his trunk, and then the elephant stamps on the carpet to flatten the folds.

Appendix 2 The Four Target Gestures in Each Condition (Speakers' Outlines Superimposed on Each Other)



Each dot on the spatial trajectory represents a video frame, i.e. 40 ms. The spatial distance between dots therefore also indicates speed of the gestural movement. The labels “plank”, etc., refer to the scenes described (cf. Appendix 1). Note that the gestures in the "Central, no hold" condition were also used as stimuli for the No-speaker-fixation condition in Study 1 and for both experimental and control conditions in Studies 3 and 4 (except that artificial holds were digitally introduced after the gesture stroke in the stimuli for the experimental condition in Study 4).

Appendix 3 Spoken Descriptions Co-Occurring with the Target Gestures in Each Condition

1. Speaker-fixation (Study 1)

a) carpet

[de] olifant loopt nog voor hem uit en strijkt alle [kreukels] '[the] elephant still walks ahead of him and irons out all [folds]

b) bowling ball

volgens komt Sylvester aan de onderkant 'then Sylvester comes [out] at the bottom'

c) plank

en dan loopt die gewoon verder 'and then he just walks on'

d) pit

komt die muis aanlopen 'the mouse comes walking along'

2. No-speaker-fixation (Study 1), Central, No Hold (Study 2) and both experimental and control conditions (Studies 3 and 4)

a) plank

en dan rent die weg 'and then he runs away'

b) wave

[een] hele hoge golve '[a] very high wave'

c) trashcan

en een beetje zo vooruit loopt 'and walks a bit ahead like that'

d) garbage

en sprint die het gebouw in 'and he runs into the building'

3. Central, plus Hold (Study 2)

a) bowling ball

[Sylvester die wordt] uitgeschoten die terras beneden '[Sylvester is] shot out down onto the terrace'

b) pole

[en dan] valt die vlak voorover '[and then] he falls down straight ahead'

c) trashcan

en dan loopt die mand achter hem aan 'and then the trashcan follows him'

d) pit

en hij springt ‘and he jumps’

4. Peripheral, no Hold (Study 2)

a) trashcan

die loopt natuurlijk met hem mee ‘it walks with him of course’

b) bowling ball

en dan rolt die maar door ‘and then he just keeps rolling’

c) carpet

en dan loopt die voor de muis ‘and then he walks ahead of the mouse’

d) wave

en dan loopt die zo verder ‘and then he walks on like that’

5. Peripheral, plus Hold (Study 2)

a) plank

loopt die weg ‘he walks away’

b) wave

en dan [krijgt hij het dat water over zich heen] ‘and then [he gets the the water all over himself]’

c) trashcan

en dan loopt die weg ‘and then he walks away’

d) bowling ball

wordt die richting een bowling centrum [gestuurt] ‘he is [sent] in the direction of a bowling center’

Appendix 4 Drawing task questions for each clip in each condition

The original Dutch question is followed by a translation into English in italics.

1. Speaker-fixation (Study 1)

a) *Wat doet de olifant nadat de muis hem heeft 'opgepompt'?* What does the elephant do after the mouse 'pumped the it up'?

b) *Wat gebeurt er met de kat nadat hij de bowlingbal heeft ingeslikt?* What happens to the cat after he swallows the bowling ball?

c) *Wat is de muis aan het doen voordat hij in de kuil valt?* What is the mouse doing before it falls into the pit?

d) *De kat lanceert zichzelf omhoog met een springplank. Hij landt weer op de plank met de vogel in zijn hand. Wat gebeurt er voordat hij geraakt wordt door het gewicht?* The cat launches itself using a springboard. It lands on the board with the bird in its hand. What happens before the cat is hit by the weight?

2. No-speaker fixation (Study 1), Central, No Hold (Study 2) and both experimental and control conditions (Studies 3 and 4)

a) *De muis heeft moeite met roeien. Waarom?* The mouse has trouble rowing. Why?

b) *De kat ziet de vogel in het andere gebouw door zijn verrekijker. Wat doet hij daarna?* The cat sees the bird in the other building through his binoculars. What does the cat do next?

c) *De kat lanceert zichzelf omhoog met een springplank. Hij landt weer op de plank met de vogel in zijn hand. Wat gebeurt er voordat hij geraakt wordt door het gewicht?* The cat launches itself using a springboard. It lands on the board with the bird in its hand. What happens before the cat is hit by the weight?

d) *De muis gooit een bananenschil in de prullenmand, zet 'm op z'n kop en loopt weg. Wat gebeurt er dan met de prullenmand?* The mouse throws a banana skin in the trashcan, turns it up side down and walks away. What happens next to the trashcan?

3. Central, plus Hold (Study 2)

a) *De muis gebruikte een lange stok om mee te springen. Hoe sprong hij?* The mouse used a long pole to jump with. How did it jump?

b) *De muis zit op de bodem van een ravijn. Hoe kwam hij daar terecht?* The mouse sits at the bottom of the pit. How did it get there?

c) *De kat slikt een bowlingbal in en hij valt naar beneden in de regenpijp. Wat gebeurt er met hem wanneer hij eruit komt?* The cat swallows a bowling ball and falls down inside the drainpipe. What happens to the cat after it comes out?

d) *De muis gooit een bananenschil in de prullenmand, zet 'm op z'n kop en loopt weg. Wat gebeurt er dan met de prullenmand?* The mouse throws a banana skin in the trashcan, turns it up side down and walks away. What happens next to the trashcan?

4. Peripheral, no Hold (Study 2)

a) *De muis heeft moeite met roeien. Hoe komt hij toch vooruit?* The mouse has trouble rowing. How does it make progress?

b) *De muis gooit een bananenschil in de prullenmand, zet 'm op z'n kop en loopt weg. Wat gebeurt er dan met de prullenmand?* The mouse throws a banana skin in the trashcan, turns it up side down and walks away. What happens next to the trashcan?

c) *De kat slikt een bowlingbal in en rolt de regenpijp uit. Wat gebeurt er daarna met hem op de straat?* The cat swallows a bowling ball and rolls out of the drainpipe. What happens to the cat next on the street?

d) *De muis kan niet zo goed over het tapijt lopen. Wat gebeurt er telkens met hem?* The mouse has some trouble walking over the carpet. What keeps happening to it?

5. Peripheral, plus Hold (Study 2)

a) *De kat lanceert zichzelf omhoog met een springplank. Hij landt weer op de plank met de vogel in zijn hand. Wat gebeurt er voordat de kat geraakt wordt?* The cat launches itself using a springboard. It lands on the board with the bird in its hand. What happens before the cat is hit?

b) *De kat slikt een bowlingbal in en rolt de regenpijp uit. Wat gebeurt er daarna met hem?* The cat swallows a bowling ball and rolls out of the rain pipe. What happens to the cat next?

c) *De muis gooit een bananenschil in de prullenmand, zet 'm op z'n kop en loopt weg. Wat gebeurt er dan met de prullenmand?* The mouse throws a banana skin in the trashcan, turns it up side down and walks away. What happens next to the trashcan?

d) *De muis heeft moeite met roeien. Wat deed de golf met zijn bootje?* The mouse has trouble rowing.

What did the wave do to its boat?

Figure 1
Click here to download high resolution image

Figure 2a

Click here to download high resolution image

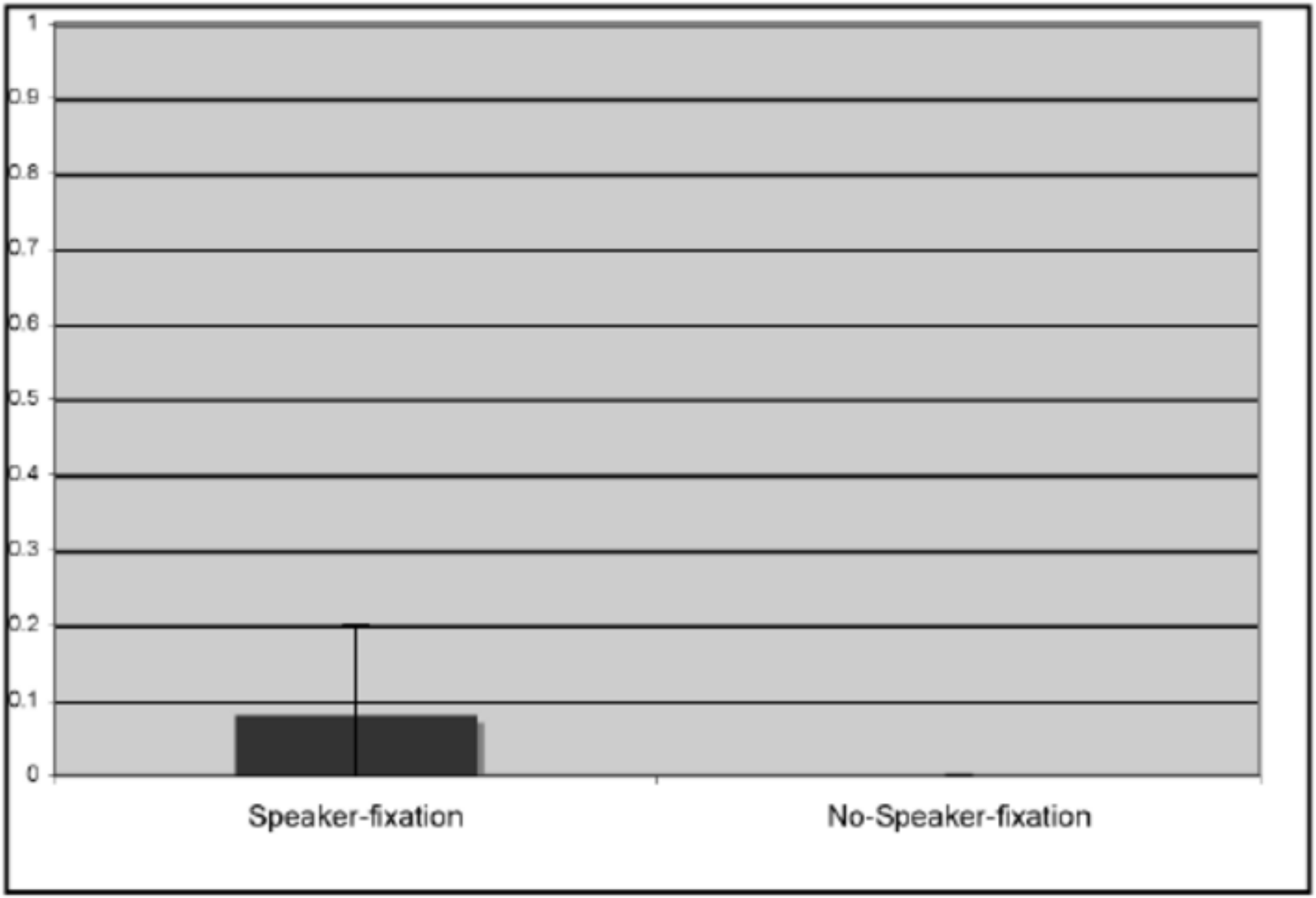


Figure 2b

Click here to download high resolution image

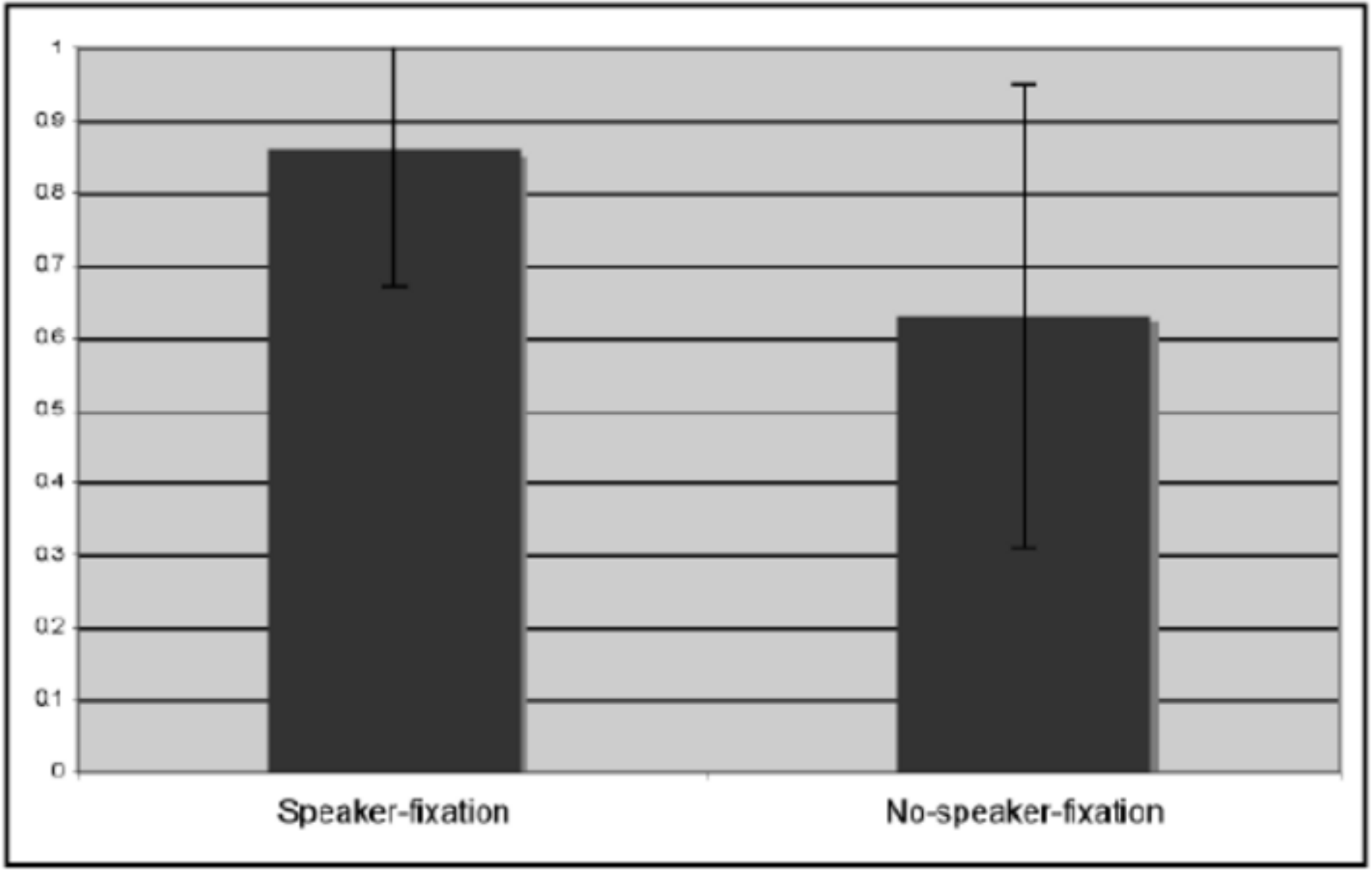


Figure 3a
Click here to download high resolution image

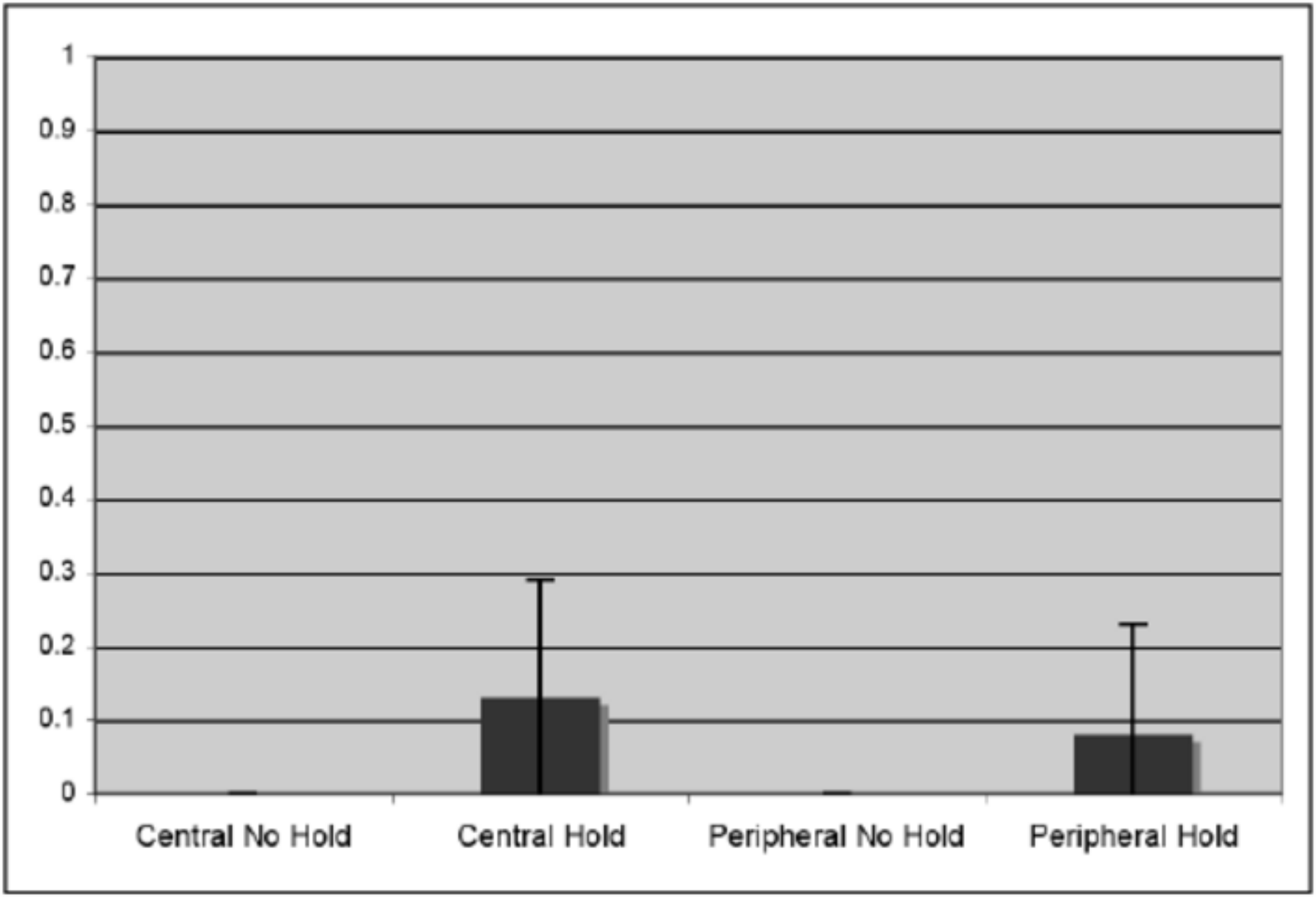


Figure 3b

Click here to download high resolution image

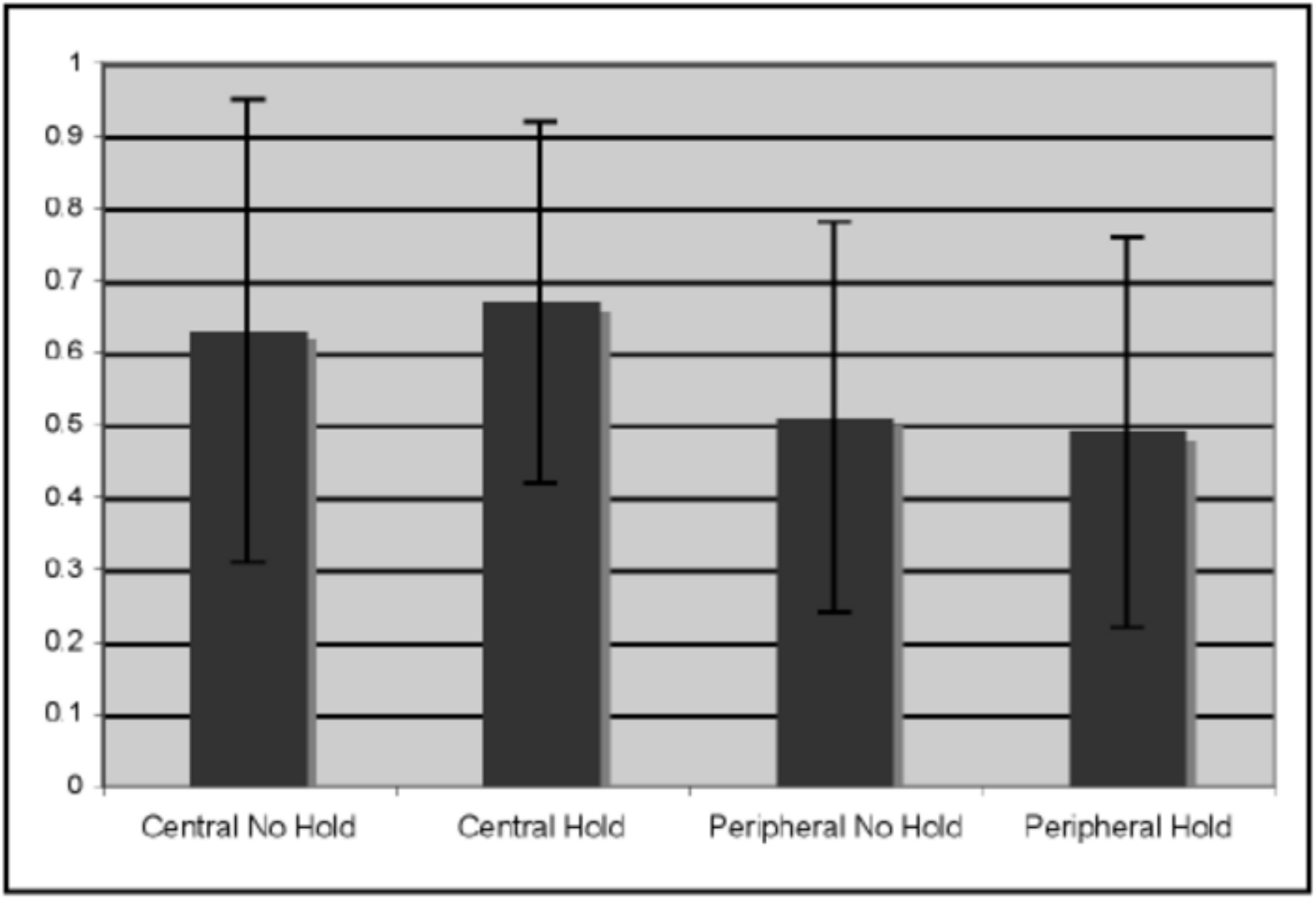


Figure 4
Click here to download high resolution image

Figure 5a
Click here to download high resolution image

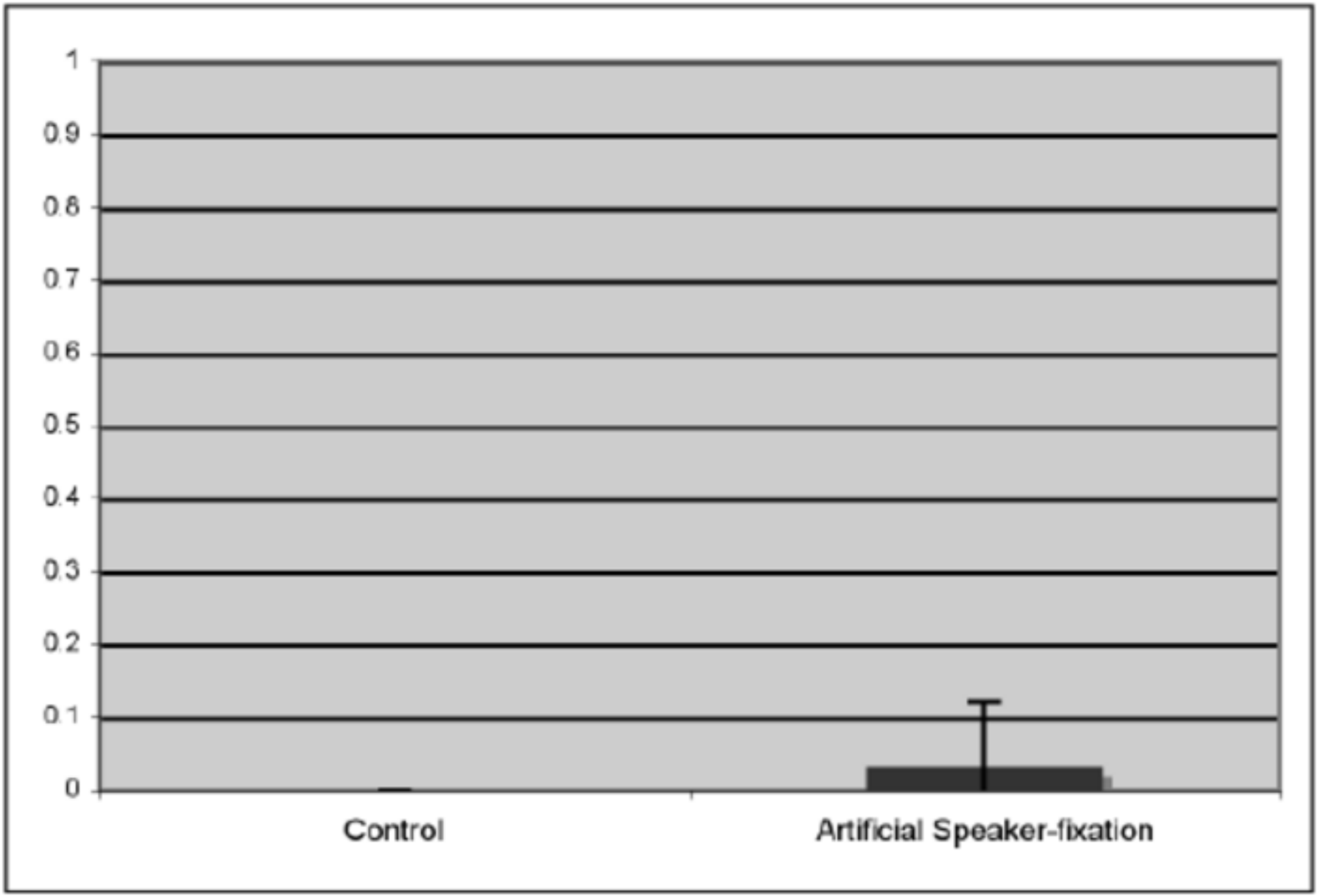


Figure 5b
Click here to download high resolution image

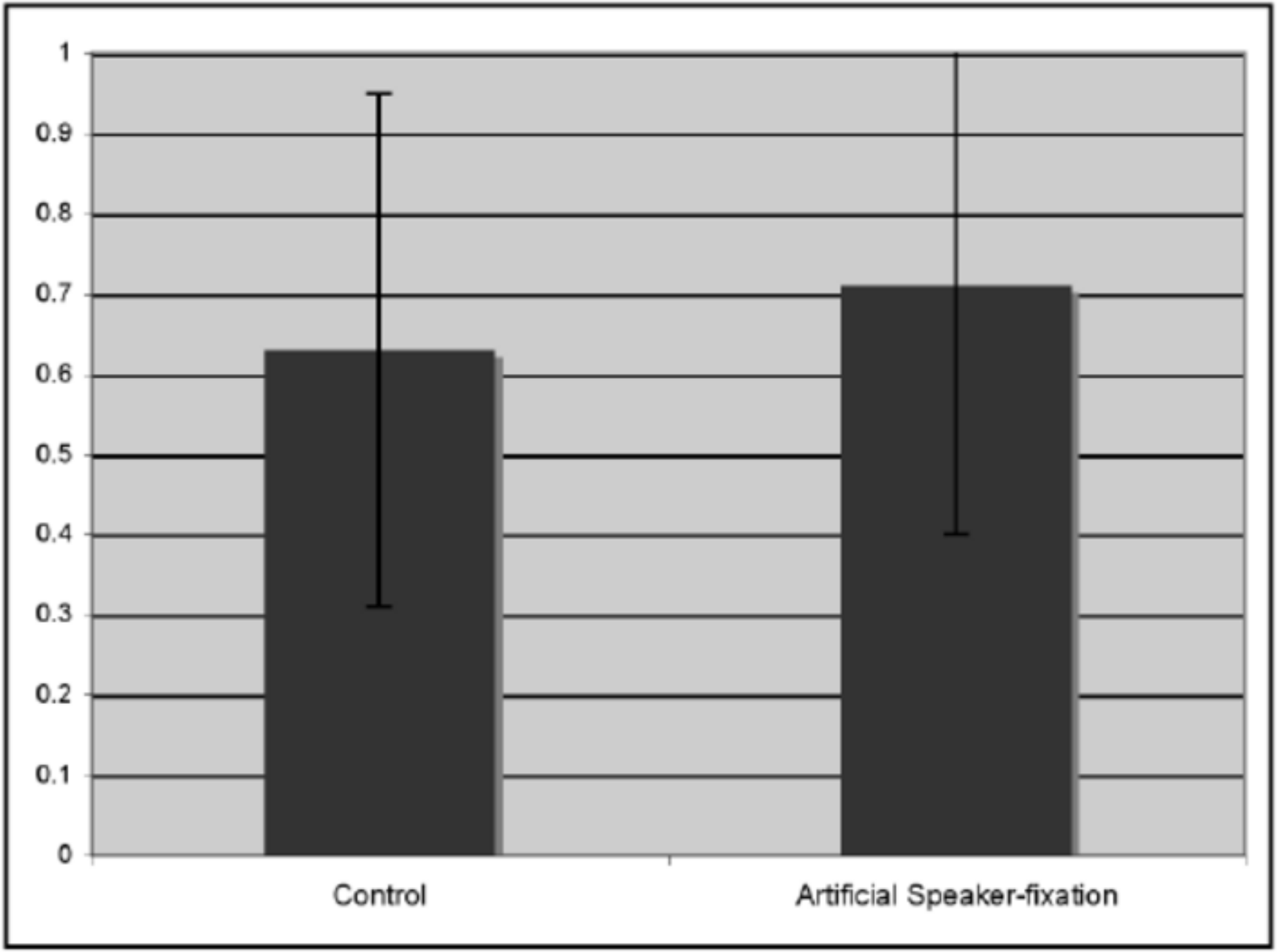


Figure 6a

Click here to download high resolution image

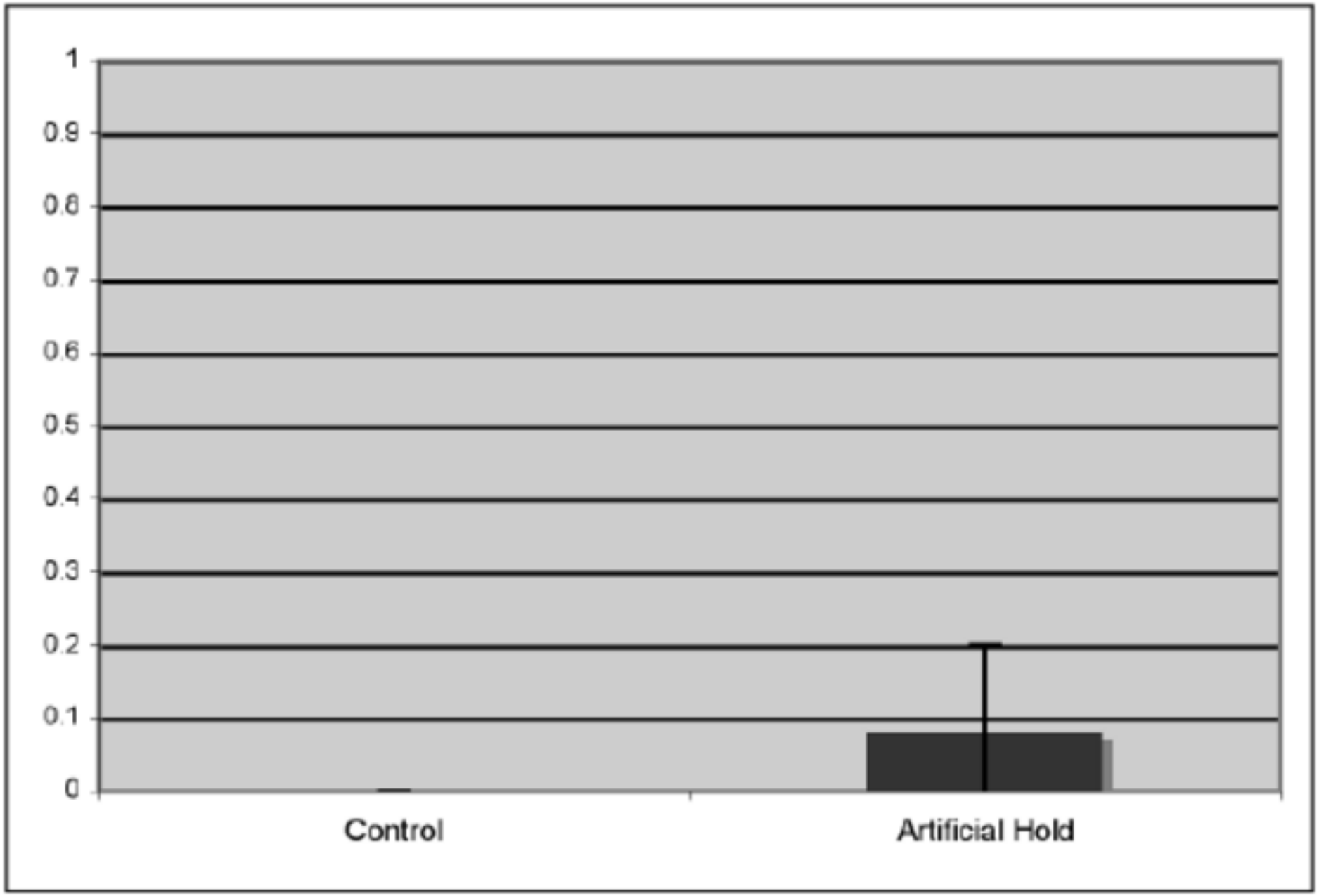
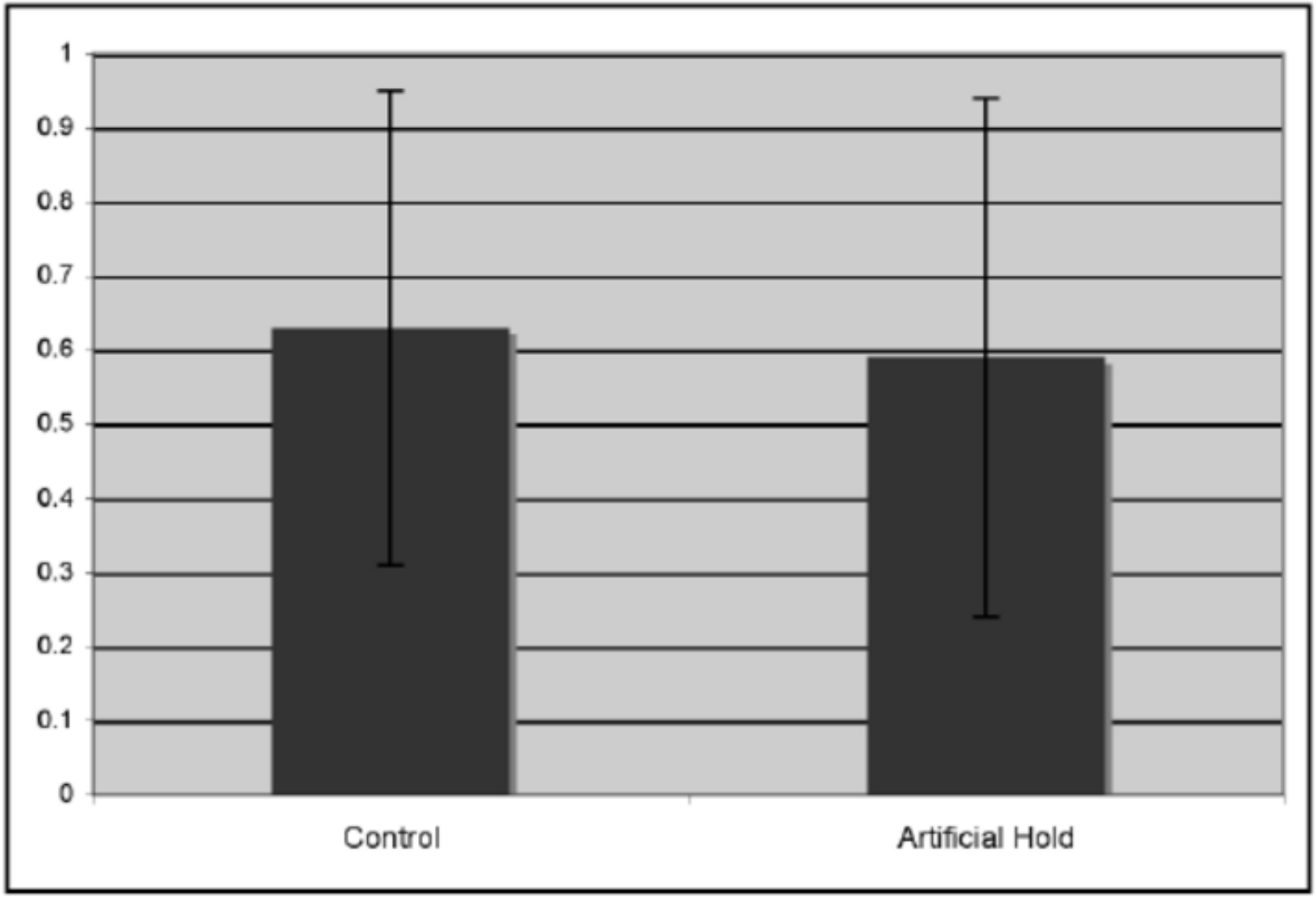
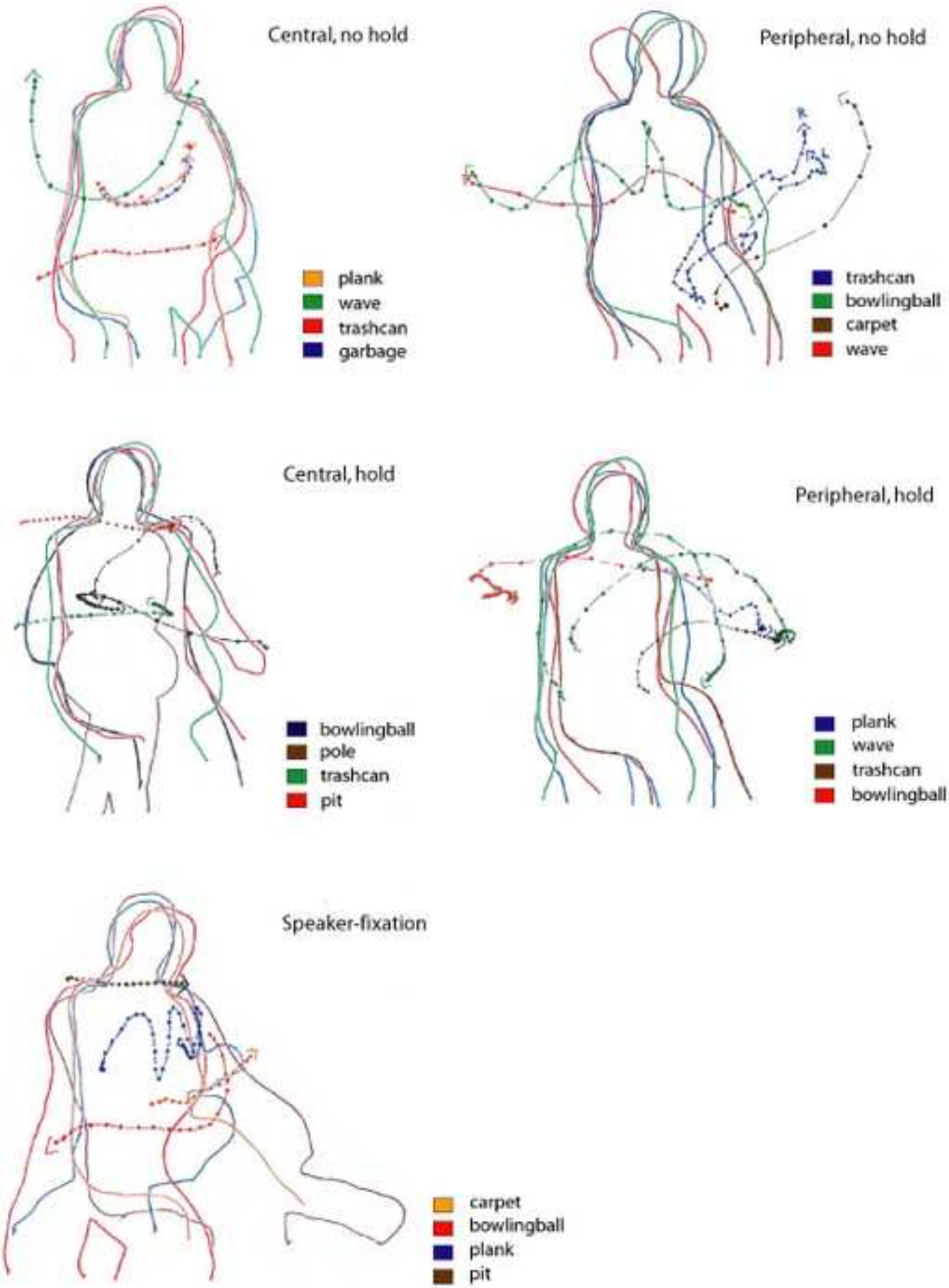


Figure 6b
Click here to download high resolution image



Appendix 2 figure

Click here to download high resolution image