Evidence for overt visual attention to hand gestures as a function of redundancy and speech disfluency

Amelia Yeo (ayeo@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Martha W. Alibali (martha.alibali@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

We investigated the effect of gesture redundancy and speech disfluency on listeners' fixations to gestures. Participants watched a speaker producing a redundant or non-redundant gesture, while producing fluent or disfluent speech. Eye movements were recorded. Participants spent little time on a speaker's gestures regardless of condition. Gesture redundancy and speech disfluency did not affect listeners' percentage dwell time to a speaker's gestures. However, listeners were more likely to fixate to a speaker's gestures when they expected the gesture to be non-redundant. Listeners were also more likely to fixate to a speaker's gestures when the speaker was disfluent. Thus, listeners allocate overt visual attention based on the expected usefulness of a speaker's gestures, although evidence does not suggest that they spend more time fixating on these gestures. Furthermore, listeners are sensitive to disfluency in a speaker's utterance and change how they attend to gestures based on qualities of the speech.

Keywords: gesture; eye tracking; communication; multimodal information processing; spatial features

Introduction

Speakers usually move their hands when conveying a message. It seems intuitive to suggest that speakers gesture to communicate information to their audience. Indeed, at times speakers appear to produce gesture specifically for the purpose of communicating with the listener (Alibali, Heath & Myers, 2011).

During the process of comprehension, listeners integrate speech and gesture (Willems, Özyürek, & Hagoort, 2007). Since co-speech gestures can influence listeners' comprehension of messages, how then do listeners allocate visual attention resources to speakers' gestures? Some researchers have argued that the content of gestures could be perceived peripherally (Gullberg & Holmqvist, 1999). If true, this would negate the need for listeners to fixate to gestures during comprehension. However, gestures have also been shown to convey additional semantic content not found in speech (e.g., McNeill, 1992; Alibali, Evans, Hostetter, Ryan & Mainela-Arnold, 2009; Hostetter, 2011). Fixating to these gestures could help comprehension. Communicating in everyday life is often a multimodal process that involves auditory input from speech and visual input from the speaker's face and body (i.e., MacDonald & McGurk, 1978; Ekman, 2004). Hence, understanding how listeners allocate visual attention during the process of face-to-face comprehension is important for understanding the mechanisms involved in the online process of interpersonal communication.

There is some evidence that listeners extend overt visual attention to a speaker's gestures (Nobe, Hayamizu, Hasegawa, & Takahashi, 1997; 2000). In these studies, the authors presented participants with animations of a speaker uttering short phrases while making hand gestures, and recorded eye movements of the participants during the animations. Participants in the study were found to fixate to gestures consistently on most of the videos presented, preferring to fixate to gestures that occurred more slowly. In the follow-up study (Nobe et al., 2000), participants were found to be able to complete gesture reproduction and comprehension tasks without necessarily fixating to the specific gesture, suggesting that listeners indeed can encode aspects of speakers' gesture without gaze fixations. However, this raises the question of why listeners would consistently fixate to speakers' gestures if comprehension can occur without fixation.

In contrast, other studies of visual attention to gesture have found that listeners rarely fixated to a speaker's gestures, even when those gestures were essential for comprehension (i.e., listeners seldom fixated to gestures that offered information absent from, and thus, non-redundant, with speech). Listeners fixated overwhelmingly on the speaker's face (Gullberg & Holmqvist, 1999, 2006; Gullberg & Kita, 2009; Beattie, Webster, Ross, 2010), contrary to the findings by Nobe and colleagues (1997; 2000).

A possible explanation for the differences found in overt visual attention to gestures in these studies is that the speech content of the speakers in the previous experiments was vastly different. Some studies used stimuli that contained a narrative element, while other studies used shorter utterances without a story element, such as "let's count fingers". The difference in speech content could have caused listeners to attend more to the face of the speaker due to the expectation or existence of emotion cues in the speaker's face. Therefore, listeners might spend more time fixating to non-redundant gestures for speakers if the speech content does not contain a narrative element.

In cases where listeners were found to fixate to a speaker's gestures, numerous factors have been cited as potentially driving the fixations. These factors include whether the speaker fixated upon the gesture, the duration of the post-stroke hold (i.e., an aspect of the "form" of the gesture) and the location of the gesture in the speaker's gesture space (e.g., Gullberg & Holmqvist, 2006; Gullberg & Kita, 2009). The focus in the literature has thus been on particular physical features of gestures, with little research into the role of listener expectation on overt attention to gestures. Expectations, or predictions, that listeners hold about the usefulness of a speaker's gestures could influence how they attend to the speaker's gestures. In this study, we examine a higher-level feature of gesture, expected redundancy. Keeping all other physical features constant, we test whether the expected redundancy of a speaker's gestures will affect how listeners attend to those gestures. If listeners do allocate attention differently to gestures depending on whether they expect the gesture to be useful for comprehension, then we should see listeners spend more time fixating to gestures and also be more likely to fixate to a speaker's gesture when the gesture offers disambiguating information absent in speech.

As mentioned above, previous studies that examined visual attention to gestures have focused on how physical qualities of a gesture influenced listeners' fixations. In multimodal communication, however, elements of speech can also influence how listeners attend to a speaker's gestures based on existing expectations. To date, no study to our knowledge has examined the role of speech disfluencies on listeners' fixations to a speaker's gestures. Disfluencies such as filled pauses cause a break in speech content and can occur at several points in speech (Ferreira & Bailey, 2004). A filled pause (i.e., *um*) that occurs in the middle of a clause has been linked to the need for the speaker to select an option for production from among several competing choices (Clark & Fox Tree, 2002). Listeners were more likely to remember a word when it was preceded by a filled pause (Corley, MacGregor & Donaldson, 2007), suggesting that a filled pause could give rise to expectation in listeners that what is to follow is important, signaling listeners to allocate more attentional resources to encode what follows from it. When listeners hear an "um" from a speaker, they might also be more likely to fixate to the speaker's gesture space when the disambiguating information might be produced in gesture as compared to a situation where there is no need for disambiguation.

In this study, we examine the effect of gesture redundancy (i.e., whether a gesture is useful for disambiguating between two options) and speech disfluency on listeners' visual attention to gesture. To do this, we conducted a 2 by 2 fully within-subjects experiment, manipulating gesture redundancy and speech disfluency. We recorded the gaze fixation data (i.e., how long each participant fixated and how many fixations) of each participant as they watched a video of a speaker on each trial. The speaker produced either redundant or non-redundant gestures for a following task and spoke with either disfluency or without disfluency. We hypothesize that listeners will be more likely to fixate to gestures that are non-redundant with speech. Participants are predicted to fixate at least once to gestures more often for trials with non-redundant gestures than for trials with redundant gestures. Participants are also predicted to spend more time fixating on these gestures. In addition, we also hypothesize that listeners will be more likely to fixate to gestures that accompany disfluent speech than to gestures that accompany fluent speech. This experiment will also allow us to examine whether spatial speech free from narrative content provides a context in which listeners attend less to the speaker's face. However, if the narrative nature of the stimuli used in previous studies was not the reason for the little time listeners spent gazing at gestures, then we expect participants in this study will display similarly low durations of fixations to gesture.

Method

Participants

Participants were 30 undergraduate students, all of whom reported being native English speakers. They were recruited from an Introductory Psychology course in exchange for extra credit.

Materials

There were two sets of stimuli: shape arrays and speaker videos. We created four pairs of shape arrangements using Microsoft PowerPoint, giving eight arrays in total. Each of these eight arrays was repeated twice in the experiment, once paired with a speech-fluent video and once paired with a speech-disfluent video. Thus, there were sixteen target trials in total.

Each pair of shape arrays was identical in every aspect except for a single shape. In the arrays used for the gesture redundant condition, only one triangle was present. In the arrays used for the gesture non-redundant condition, two triangles were present. Thus, to create the arrays for the gesture non-redundant condition, one of the non-triangle shapes in the arrays for the gesture redundant condition was replaced with a triangle (Fig. 1).

Figure 1. An example of a shape array in the gesture redundant condition (left) and in the gesture non-redundant condition (right).

Next, we created eight videos, four featuring fluent speech and four featuring disfluent speech. Each video lasted approximately six seconds and showed a speaker describing a triangle according to a script, while facing the camera (Fig. 2). In the videos with fluent speech, the speaker produced an utterance, such as "the triangle changed color and turned green". In videos with disfluent speech, the speaker produced an utterance with the "um"

disfluency, for example, “the, um, triangle changed color and turned green”. In the other of these eight videos, the actor produced exactly the same utterance except with a different color (e.g., orange/red/yellow instead of green). We created the videos such that there were fluent and disfluent pairs containing the same utterance that differed only in the inclusion, or exclusion, of the disfluency “um”. In addition, the speaker produced four types of gestures that were paired with corresponding shape arrays. These gestures referred to the triangle that was undergoing the color change. Thus, in the gesture non-redundant condition, the gesture functioned to disambiguate the target triangle from the other triangle in the array. In each video, the speaker’s gesture depicted either the pointed tip of the triangle (pointing up or down), or the relative placement of the triangle in the shape array (located high in the array or located above a line). Each gesture was scripted such that the actor began forming the gesture just before the word “triangle” in the utterance and held the gesture for approximately 2 seconds before dropping her hands. In each video, the actor produced only one gesture and gazed at the camera for the duration of the video.

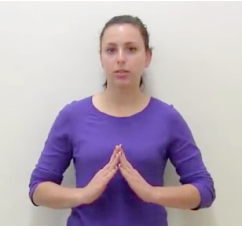

Figure 2. Screen capture of the speaker producing a gesture of an upward-pointing triangle.

We also created shape arrays and speaker videos for filler trials. The purpose of the filler trials was to present the participant with variation in the speaker videos so as to reduce the chances of the participant inferring the purpose of the study. These filler trials contained an assortment of videos where the speaker did not gesture, or gestured while producing a slightly different utterance, such as “the orange triangle changed color and turned green”. There were ten filler trials in total. The eight target trials and the filler trials all contained the same actor wearing the same clothing.

Procedure

Participants were tested individually. Each participant was seated in front of a computer screen and a desk-mounted Eyelink 1000 eye tracker camera. The eye tracker recorded real-time fixations of each participant throughout the entire experiment and was calibrated for each participant before the trials began.

Before the experiment, participants were told that the speaker would always describe a color change of a triangle in the array. Thus, participants began the experiment knowing that it would always be a triangle that changed color. They were not told that the speaker would gesture; participants were not informed in any way that the study was about gesture or speech disfluency.

During the experiment, each participant viewed 26 trials presented in random order using Experiment Builder from SR Research (Canada). Each trial contained a shape array that was presented onscreen for 5 seconds, followed by a video of the speaker describing the color change occurring to a triangle in the array. The video was programmed to start automatically. After the video, participants were presented with four options of shape arrays and were instructed to say aloud the option that fit the description of the speaker in the video. For example, if a trial presented the array in the gesture non-redundant, speech fluent condition (e.g., the array on the right in Fig. 1) followed by a video of the speaker producing an upward-pointing gesture (Fig. 2) while saying, “the triangle changed color and turned green”, the correct option (in Fig. 3) to select would be option C.

The trials in the gesture non-redundant, speech disfluent condition were identical except that the speaker produced a filled pause, for instance, “the um, triangle changed color and turned green”. Thus, gesture redundancy was manipulated by having either one or two triangles in the shape array.

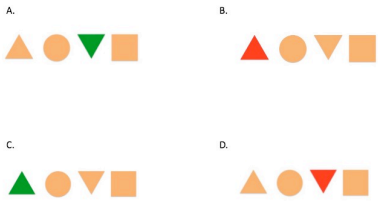

Figure 3. Example of four response options in the gesture non-redundant conditions.

An example of a trial in the gesture redundant, speech fluent condition would be the left array in Figure 1 followed by a video of the speaker producing an upward-pointing gesture (Fig. 2) while saying, “the triangle changed color and turned green”. The four response options would then contain the same shapes as in the original array but with the single triangle colored in four different colors. The trials in the gesture redundant, speech disfluent condition were identical except that the speaker produced a filled pause, for instance, “the um, triangle changed color and turned green”. Thus, gesture redundancy and speech disfluency were perfectly orthogonal.

Each participant’s verbal response for each trial was recorded with a microphone that was clipped on to his or her clothing. The verbal responses were recorded to audio files in the computer. At the end of the experiment, participants were debriefed and asked if they could guess the purpose of the study. None of the participants correctly stated the hypothesis about gesture redundancy or speech disfluency on listeners’ fixation to a speakers’ gestures. Throughout the whole procedure, an experimenter sat in a corner in the room unobtrusively and had no interaction with the participant. The whole experiment lasted for about 20 minutes.

Coding

Each video was divided into interest areas for eye tracking analysis. The speaker's face was a separate interest area from her gesture space. The fixations of interest for this study were those that occurred to the speaker's gestures from the start to the end of her utterance, since her gestures always occurred as she was speaking. Fixation data that included each dwell time on each area and number of fixations was then exported from Data Viewer (SR Research) for analysis. For each trial, we thus obtained data regarding how long a participant fixated to the speaker's face, how long a participant fixated to the speaker's gesture space, how many fixations a participant made to the speaker's face and how many fixations a participant made to the speaker's gesture space.

Results

Averaging across all conditions, participants spent the majority of the time fixated on the speaker's face, spending only 9.3% of the time fixating on the speaker's gestures.

Table 1 displays the average percentage dwell time spent by participants on the listeners' gestures across conditions. We conducted two-way within-subjects analysis of variance on the average percentage dwell time spent fixating on the speaker's gestures as a function of gesture redundancy and speech disfluency. There was no significant main effect of gesture redundancy, F (1, 112) = 1.34, p = 0.25, nor was there a significant main effect of speech disfluency, F < 1, p = .66.

There was also no significant interaction between gesture redundancy and speech disfluency on participants' dwell time to speaker's gestures, F (1, 112) = 2.30, p = .13. Even though participants on average spent a higher percentage of dwell time on non-redundant gestures, this difference was not significant.

Table 1. Average dwell time % to the speaker's gestures as a function of gesture redundancy and speech disfluency.

	Gesture	
Speech	Redundant	Non-redundant
Disfluent	7.33	10.1
Fluent	9.02	10.8

Since participants overwhelmingly fixated to the speaker's face in this experiment, we wanted to examine whether gesture redundancy and speech disfluency affected the likelihood of participants fixating at least once to the speaker's gestures. To test whether participants were more likely to fixate to a speaker's gestures as a function of gesture redundancy or speech disfluency, we classified whether each participant fixated on the video speaker's gesture space at least once while the speaker was talking. Thus, the outcome variable for this analysis was dichotomous, i.e., whether or not the participant fixated at least once to the speaker's gesture in each trial.

We analyzed these data using a binomial multilevel model with gesture redundancy and speech disfluency as fixed effects and participant as a random effect. The dependent variable was whether the participant had fixated to the speaker's gesture space (yes/no). The mean proportion of trials on which participants fixated at least once to the speaker's gesture space is displayed as a function of gesture redundancy (Fig. 4) and speech disfluency (Fig. 5).

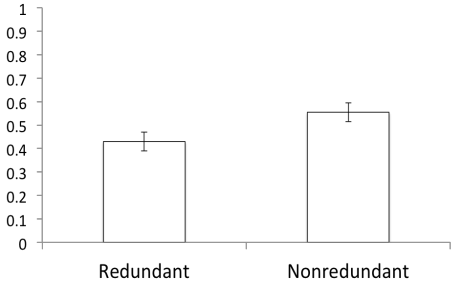

Figure 4. Proportion of trials on which participants had at least one fixation to the speaker's gesture space as a function of gesture redundancy. Error bars are ±SE.

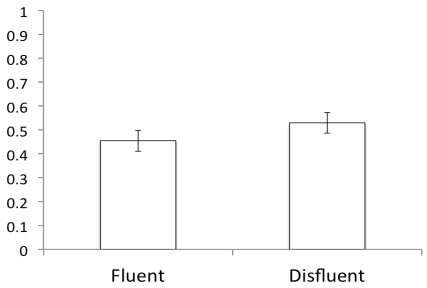

Figure 5. Proportion of trials on which participants had at least one fixation to the speaker's gesture space as a function of speech disfluency. Error bars are ±SE.

Listeners were significantly more likely to fixate to the speaker's gesture in the gesture non-redundant condition than in the gesture redundant condition, Wald's z = 3.06, p < .01, odds ratio = 2.41. Additionally, listeners were significantly more likely to fixate to the speaker's gesture in the disfluent speech condition than in the fluent speech condition, Wald's z = 2.21, p = .027, odds ratio = 1.88. There was no significant interaction between gesture redundancy and speech disfluency on the likelihood of participants fixating to a speaker's gesture, Wald's z = 1.35, p = .18. In sum, participants were more likely to fixate at least once to non-redundant gestures, and they were also more likely to fixate at least once to the speaker's gestures when the speaker was disfluent.

Discussion

The finding that participants spend little time fixating to a speaker's gestures reflects the results from some past studies. For example, Gullberg and Kita (2009) reported that listeners fixated on gestures only 8% of the time, even

though these gestures were first fixated by the speaker, showing that gesture fixation duration was low even when there was social impetus (i.e., directed gaze) to fixate at a gesture. Our findings align with this value. Listeners fixated to gestures on average about only 10% of the time, even for gestures that contained information not present in the speaker's utterance. These findings do not support the hypothesis that the previously reported low fixation durations on gestures were due to the narrative element in speech. Instead, listeners fixate overwhelmingly on the speaker's face even when the narrative element in speech is absent or greatly reduced.

However, we do not yet know if listeners direct so little overt visual attention to gesture because of the communicative context. Past studies, including this one, have featured speakers passively describing objects or actions. Although strengths of this paradigm are its simplicity and ease of experimental control, a limitation is that it tells us little about how people attend to each other's gestures when they are engaging in dialogue. During dialogue, speakers gesture differently depending on the feedback they receive from the listener (Holler & Wilkin, 2011). This finding reflects observations of research involving instructional gestures. In the classroom, teachers have been found to gesture more when students lack understanding of the lesson (Alibali et al., 2013). Further research could explore how listeners attend to gestures in an instructional setting or in dialogue, using a wearable eye tracker.

As predicted, participants were more likely to fixate to non-redundant gestures than to redundant gestures. This finding implies that listeners preferentially direct overt visual attention to gestures that they expect to be useful for comprehension. Listeners direct overt attention to a speaker's gestures more often when the gesture conveys relevant information not present in speech, implying that listeners generate expectations about the perceived importance of the speaker's gestures and direct attention accordingly. However, we did not find support for the hypothesis that listeners would spend *more time* fixating to a speaker's gestures. While listeners were more likely to gaze at least once to the speaker's non-redundant gestures, they did not spend more time dwelling on those gestures, implying that the additional fixations to non-redundant gestures occurred very quickly. A potential explanation for this behavior is that visual information from fixated gestures is gleaned very quickly, making it unsurprising that fixation durations across conditions did not differ significantly.

On the surface, it might be unsurprising that listeners are less likely to fixate to gestures that are redundant. This study demonstrates that listeners are less likely to fixate to gestures that are redundant even when those gestures are holds (i.e., the form of the gesture is held in a pause) and occur in the center of the speaker's body, qualities that were reported to best attract listeners' fixations (e.g., Gullberg & Holmqvist, 1999; 2006) Since we controlled for these features across the gesture redundant and gesture non-redundant conditions, our findings imply that top-down factors such a redundancy can influence listeners' visual attention to gestures beyond the physical characteristics of those gestures. Few studies to date have explored the role of higher-level cognitive factors, such as expectations, on how listeners process gestures. For example, individuals could hold expectations about the usefulness of gesture based on an individual's communicative fluency, or individual's communicative style. A further direction would be to examine how these factors influence how listeners attend to gestures.

In this study, we also found support for the hypothesis that speech disfluency causes listeners to be more likely to attend to gestures during communication. These findings support the idea that disfluencies in speech can function as a signal to listeners on how to direct their cognitive resources during comprehension. However, we did not find support for the hypothesis that listeners spent more time fixating to gestures that co-occurred with disfluent speech as compared to gestures that occurred with fluent speech. Once again, it is possible that listeners quickly obtained information from gestures. If filled pauses in speech do indeed work as a signal for cross-modal attention shifts, future work could examine if how other forms of speech disfluencies (e.g., false starts) influence visual attention to a speaker's gestures.

As with any investigation, there are some limitations to this experiment. Due to convenience sampling, our sample was comprised of college undergraduates. Undergraduates could offer little variation in terms of cognitive skills as compared to the population at large. While little published research to date exists examining the role of individual differences in cognitive skills on attention to gestures, there is evidence suggesting that people produce gestures differently due to individual differences in spatial abilities (Hostetter & Alibali, 2007; 2011). It may be the case that listeners with vastly different spatial skills could process a speaker's gestures differently. One way to address this would be to administer measures of verbal and spatial skills to undergraduate participants in future studies. Another way to address this limitation would be to recruit participants outside of the undergraduate pool.

Our participants were English speakers in the Midwestern USA, thus the results might not generalize to speakers of a different language or culture. Past studies on visual attention to gestures have sampled from English-speaking students in the United Kingdom (Beattie, Webster & Ross, 2010), Dutch-speaking students (Gullberg & Kita, 2009) and native Swedish speakers (Gullberg & Holmqvist, 2006). Consistently low fixation durations to gesture across these samples appears to suggest that the effect is generalizable. However, Nobe and colleagues (1997; 2000) sampled from Japanese speakers, raising the question of whether the difference in attention to gestures of a speaker is partly due to cultural norms.

For instance, Graham and Argyle (1975) found that Italian speakers were better able to decode shapes being

described by the speaker when gesture was produced, in contrast to English speakers. If speakers' gestures possess different utility value to listeners depending on the language, we might expect listeners to attend to gestures differently too. Further research should test the assumption that listeners' processing of speakers' gestures is universal. There are undoubtedly common processes involved in multimodal communication across humans, but cultural norms in communication or in the use of hand gestures could also influence how listeners process these gestures.

Another limitation of this study involves the nature of scripted disfluencies. When disfluencies are produced naturally, they could be accompanied by changes in speech rate, tone of voice, or changes in facial expression. Having an actor utter a statement with a scripted disfluency across multiple trials is unnatural. While this choice was made to reduce stimuli variability, further research could use videos of speakers conversing naturally and examine the gaze of listeners when disfluency occurs naturally.

In conclusion, these findings provide another perspective on the question of how listeners process gestures. We show that listeners are more likely to fixate to a speaker's gestures when those gestures are non-redundant, after controlling for physical properties of gesture that have been reported to capture the attention of listeners. We also demonstrate that speech disfluencies can act as signals for listeners to shift attention multimodally. These findings highlight the causal role of expectations in how listeners attend to speakers' gesture.

Acknowledgments

We thank Maia Ledesma and Kayla Diffee for assistance with production of stimuli, Youn Ku Choi for assistance with data collection, and Mitchell Nathan and Virginia Clinton for assistance with the eye tracker. We also thank Marianne Gullberg for helpful comments during presentation of a version of this work at the conference of the International Society for Gesture Studies in Paris.

References

Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language, 44(2)*, 169-188.

Alibali, M. W., Evans, J. L., Hostetter, A. B., Ryan, K., & Mainela-Arnold, E. (2009). Gesture–speech integration in narrative: Are children less redundant than adults?. *Gesture, 9*(3), 290-311.

Alibali, M. W., Nathan, M. J., Church, R. B., Wolfgram, M. S., Kim, S., & Knuth, E. J. (2013). Teachers' gestures and speech in mathematics lessons: Forging common ground by resolving trouble spots. *ZDM, 45*(3), 425-440.

Beattie, G., Webster, K., & Ross, J. (2010). The fixation and processing of the iconic gestures that accompany talk. *Journal of Language and Social Psychology, 29*(2), 194-213.

Clark, H. H. & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition, 84*, 73-111.

Corley, M., MacGregor, L. J. & Donaldson, D. I. (2007). It's the way you, er, say it: Hesitations in speech affect language comprehension. *Cognition, 105*, 658-668.

Ekman, P. (2004). Emotional and conversational nonverbal signals. In *Language, knowledge, and representation* (pp. 39-50). Springer Netherlands.

Ferreira, F. & Bailey, K. G. D. (2004). Disfluencies and human language comprehension. *Trends in cognitive sciences, 8*, 231-237.

Graham, J. A., & Argyle, M. (1975). A cross‐cultural study of the communication of extra‐verbal meaning by gestures. *International Journal of Psychology, 10*(1), 57-67.

Gullberg, M., & Holmqvist, K. (1999). Keeping an eye on gestures: Visual perception of gestures in face-to-face communication. *Pragmatics & Cognition, 7*(1), 35-63.

Gullberg, M., & Holmqvist, K. (2006). What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition, 14*(1), 53-82.

Gullberg, M., & Kita, S. (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of nonverbal behavior, 33*(4), 251-277.

Holler, J., & Wilkin, K. (2011). An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. *Journal of Pragmatics, 43*(14), 3522-3536.

Hostetter, A. B., & Alibali, M. W. (2007). Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture, 7*(1), 73-95.

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological bulletin, 137*(2), 297.

Hostetter, A. B., & Alibali, M. W. (2011). Cognitive skills and gesture–speech redundancy: Formulation difficulty or communicative strategy?. *Gesture, 11*(1), 40-60.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Attention, Perception, & Psychophysics, 24*(3), 253-257.

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

Nobe, S., Hayamizu, S., Hasegawa, O., & Takahashi, H. (1997, September). Are listeners paying attention to the hand gestures of an anthropomorphic agent? An evaluation using a gaze tracking method. In *International Gesture Workshop* (pp. 49-59). Springer Berlin Heidelberg.

Nobe, S., Hayamizu, S., Hasegawa, O., & Takahashi, H. (2000). Hand gestures of an anthropomorphic agent: Listeners' eye fixation and comprehension. *Cognitive Studies, 7*(1), 86-92.

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex, 17*(10), 2322-2333.