Journal of Language and Social Psychology OnlineFirst, published on March 8, 2010 as doi:10.1177/0261927X09359589

Journal of Language and Social Psychology
XX(X) 1–20

DOI: 10.1177/0261927X09359589
http://jls.sagepub.com
SAGE

The Fixation and Processing of the Iconic Gestures That Accompany Talk

Geoffrey Beattie,[1] Kate Webster,[1] and Jamie Ross[1]

Abstract

In everyday communication, semantic information is transmitted by both speech and the gestures that accompany speech. Listeners, therefore, need to monitor two quite different sources of information, more or less simultaneously. But we know little about the nature or timing of this process. This study analysed participants' attentional focus on speech–gesture combinations, differing in both span and viewpoint, using a remote eye tracker. It found that participants spent most time fixating the face with just 2.1% of the time looking at gestures, but with certain categories of gesture, up to 26.5% of the stroke phases were successfully fixated. In other words, visual attention moves unconsciously and quickly to these information-rich movements. It was also found that low-span Character-Viewpoint gestures attracted most fixations and were looked at longest. Such gestures are particularly communicative, and the way these gestures attract visual attention may well be a crucial factor.

Keywords

gestures, eye fixations, speech, visual attention

Many studies have demonstrated that the spontaneous imagistic gestures that accompany speech in talk (termed *iconic* gestures because of their mode of representation) can represent and convey meaningful information to an addressee, and that addressees can successfully process this gestural information and combine it with the information in the speech itself to form a more complete representation of an utterance (Beattie,

[1]University of Manchester, Manchester, UK

Corresponding Author:
Geoffrey Beattie, School of Psychological Sciences, The University of Manchester, Manchester M13 9PL, UK
Email: geoff.beattie@manchester.ac.uk

2003; Beattie & Shovelton, 1999a, 1999b; McNeill, 1992). This whole process requires careful monitoring by the listener of two quite different sources of information (one is verbal in form with a linear and sequential pattern, and the other is imagistic with a non-sequential pattern). The monitoring of these two sources has to be carried out more or less simultaneously, but we know little about the nature or timing of this whole process. Exactly when do listeners attend to a speaker's gestural movements and how is this pattern of listener attention and gaze fixation affected by the properties of the movements produced by the speaker?

Research in this area of communication research has clearly demonstrated that many types of gesture are highly communicative but also critically that some categories of gesture are significantly more communicative than others (Beattie & Shovelton, 2002, 2005). It seems that the particular semantic properties represented by the gesture are one crucial feature in determining the communicative power of the movement—gesture that represent the "relative position" of objects or the relative position of objects, and characters and gesture that represent the "size" of characters or objects are particularly effective (Beattie & Shovelton, 1999a, 2001). Another significant variable is the particular viewpoint from which the gesture is generated (Beattie & Shovelton, 2001, 2002). McNeill (1992) argued that gesture can be produced either from the viewpoint of the character being talked about (termed *Character-Viewpoint* or C-VPT for short) or from that of an observer of the situation (*Observer-Viewpoint* or O-VPT). For example, someone who illustrates the phrase "He ran away" with a gesture showing how both arms pump up and down (as if running) is using a C-VPT gesture—the speaker is acting as if he or she *is* the person being referred to. The same phrase can also be illustrated by an O-VPT gesture involving one hand moving from one side to another in front of the speaker, thus representing an observer's perspective of the scene in question. The speaker's gesture here represents the whole person as seen by an observer. McNeill (1992) had suggested that C-VPT gesture are more communicative than O-VPT gesture because "In a narrative, the voice of a character seems to push the communication forward more than the voice of an inside observer, and an inside observer might push it more than an outside observer" (p. 208). Beattie and Shovelton (2002) empirically investigated this claim and found that C-VPT gesture were in fact significantly more communicative than O-VPT gesture and, furthermore, that C-VPT gesture were particularly good at communicating information about the "relative position" of actors and objects as they can directly show the position of something in relation to the speaker's body. So viewpoint is one feature of a gesture that does seem to affect its communicative power. But more basic physical features of a gesture may also be important, for example, the span of the stroke phase of the gesture (the meaningful part of the gesture) may be highly significant, in that gesture with a longer stroke phase could potentially be more noticeable and therefore more readily interpretable (Beattie & Shovelton, 2005).

The identification of the factors that affect the communicative power of individual gesture is obviously a crucial issue in communication research for both theoretical and practical reasons. In terms of theory, it is crucial because McNeill (1992) maintains that human semantic communication *ordinarily* proceeds through both speech and

iconic gesture working together. But why then do only certain gesture appear to be communicative? How general is the claim that iconic gesture are indeed a central component of semantic communication? And why do some gestures appear to be much less significant than others? Is it because they are expressing essentially the same things as the speech itself and are therefore redundant? Or is because there is critical non-redundant information embedded within them, but the listener is failing to pick up on this information, perhaps, because the gestures are not being attended to? In other words, is it something to do with the fundamental semiotic organization of the two systems of speech and gesture or merely to with the psychological vagaries of how people attend, or fail to attend, to the gestures themselves. In terms of practical implications, the communicative power of individual gestures is also a crucial issue because to design effective communications (such as TV ads or scripted and choreographed political presentations!) involving this new theoretical perspective (Beattie, 2003, Beattie & Shovelton, 2005), we need to know which iconic gestures to include in our messages and which to omit, and if we are to include gestures, what specific properties should they have to maximise their effectiveness.

One of the most obvious ways to explore this issue is to use an eye-tracking methodology to investigate which gestures are attended to and how the attentional focus of the listener is reflected in the uptake of information from the gestures. This issue is relatively unexplored but it could potentially be very revealing. When we look at a scene, our eyes move around continually, locating interesting points and building up a corresponding mental image. These small, rapid movements of the eyes are known as saccades. "Between the saccades, our eyes remain relatively still during fixations for about 200-300 ms" (Rayner, 1998, p. 373). These fixations are thought to reduce image blur, allowing the visual system time to process the image' (Turano, Geruschat & Baker, 2003, p. 333). Research has shown that little or no actual visual processing occurs during the saccades themselves (Fuchs, 1971). We make saccades so frequently because the visual field is divided into three regions: foveal, parafoveal, and peripheral; only the foveal region has very good acuity because of its high concentration of colour-sensitive photoreceptor cells called cone cells. So, we move our eyes to reorient the fovea on that area of the stimulus that we want to see accurately. The less central parts of the retina are mostly made up of rod cells, which are particularly good at motion detection.

Several studies have looked at gaze patterns in relation to gestures, but few have directly investigated the relationship between fixation and the uptake of information from gestures. Gullberg and Holmqvist (1999, 2002) found that in face-to-face interaction with naturally occurring gestures addressees fixated the speaker's face for 96% of the total viewing time. Only 0.5% of the total viewing time was spent actually fixating gestures, and only 7% of all gestures were fixated. However, in another study, Nobe, Hayamizu, Hasegawa, and Takahashi (1998, 2000) presented an anthropomorphic agent (instead of a real speaker) on a computer screen and found that addressees in this situation fixated the majority of gestures (as much as 75% of the total). The authors suggested that these results may be because of the fact that using an anthropomorphic

agent removes the social constraint of focusing attention on the speaker's face (see, for example, Argyle, 1967) and allows greater fixation on other areas of the stimulus that might not be acceptable in face-to-face interaction. In an attempt to resolve these apparent contradictory results, Gullberg and Holmqvist (2006) investigated whether attention is modulated by changes in social situation (*actual partner* vs. *partner on video*, in which there are fewer social obligations about focusing almost exclusively on the face) as well as investigating whether attention is affected by the display size of the stimulus. In all conditions, the face dominated as the addressees' fixation target and only a minority of gestures drew fixations. In addition, Gullberg and Kita (2009) attempted to establish whether the location of the gesture performance had an effect on the pattern or frequency of fixation. Gullberg (2003) noted that it is often presumed that gestures performed in the speaker's peripheral gesture space attract overt visual attention. This is because the majority of a speaker's gestures are performed in the central gesture space, so if the addressee naturally fixates the speaker's face (Gullberg & Holmqvist, 1999, 2002), then these centrally performed gestures will already be in the peripheral vision of the addressee and will not require any overt head movements or eye movements to interpret them. If, however, the speaker performs a gesture in their peripheral gesture space, this gesture will only appear in the addressee's extreme peripheral vision and therefore may well attract more direct visual attention. However, Gullberg and Kita (2009) actually found that the location of gesture performance had little discernible impact on the addressees' fixation.

Although Gullberg and Holmqvist (2002) found that participants' tendency to fixate the gestures was very low, certain types of gestures were found to reliably attract higher levels of fixation. These were "holds," those momentary cessations in the movement of a gesture, and "autofixations," those gestures that were fixated by the encoder themselves. During "holds," the movement of a gesture comes to a stop, and therefore, peripheral vision is no longer sufficient for obtaining information from that gesture, thus necessitating a degree of fixation. "Autofixations," on the other hand, serve as a powerful social cue to joint attention in an interactive setting. The authors attributed the low frequencies of gesture fixation generally to the fact that peripheral vision is sufficient for detecting broad gestural information, such as location, direction and size, provided that the gestures are moving.

This research clearly fails to answer definitively the question of the relationship between the fixation or non-fixation of gestures and the amount of information received from these gestures and indeed whether gestures that attract the highest levels of fixation are the most communicative (as revealed by the standard paradigms in this area). This study aims to remedy this by directly testing the relationship between the level of fixation of gestures and the information uptake from those gestures (although Gullberg & Kita, 2009, did consider information uptake from gestures, their study was solely concerned with one semantic property, namely directional information, i.e., left or right). The gestures in the present study will be encoding core semantic features such as "size," "relative position," "shape," and "movement" (and not just directional information), and this semantic information will only be encoded in the *complementary* gestures themselves and not in the speech itself.

Method

Participants

Ten students from the University of Manchester participated in the study. All were compensated with course credits.

Equipment

An ASL Model 504 remote eye tracker was set up in a laboratory, in front of a computer monitor on which the stimulus material was to be shown. The eye tracker employs a camera surrounded by infrared emitting diodes to illuminate the eye of the participant looking at a screen. The participant's point of gaze on the screen is determined by the camera combining the position of the pupil and the corneal reflection. The remote camera in the eye tracker fed into a screen for the experimenter's observation of the positioning of camera observing the eye. From a separate computer, the experimenter was able to adjust the illumination of the infra red camera and the "Pan/Tilt" of the camera in the eye tracker to enable recognition of the pupil and corneal reflection.

Stimulus material consisted of 12 short video clips,[1] each lasting between 5 and 20 s in length, of a person narrating cartoon stories (the person was an actor, paid to recite the short scripts containing speech and cospeech gestures, the scripts being based on data collected from actual participants in a previous study; see Beattie, Webster, & Ross, in press). Every video clip contained one gesture, which was scripted to conform to one of the four gesture categories (see below for further information on these categories). There were three clips (i.e., three different gestures) for each of the four gesture categories.

Each short video clip was followed by two "Yes/No" questions relating to the information encoded only within the gesture in the preceding clip. For example, one of the video clips showed the actor saying,

> There's a guy, you can only see one of the guys but he's obviously playing beach ball with somebody or something um on the edge of a pier and he goes to hit it and is just about to fall off the end of it into the [water] um and just about catches himself and doesn't fall.

[Left hand is low at left side of body with palm facing down and fingers spread; hand moves slightly towards the right in front of body]

The gesture in this clip coincided with the word "water." The two questions following the clip, which related to the information encoded in the gesture alone, were:

"Is the man quite far away from the water?" (Y/N)
"Is the water still?" (Y/N)

Because this study aims to differentiate between the level of fixation and communicative power of different types of gestures, according to span and viewpoint, four

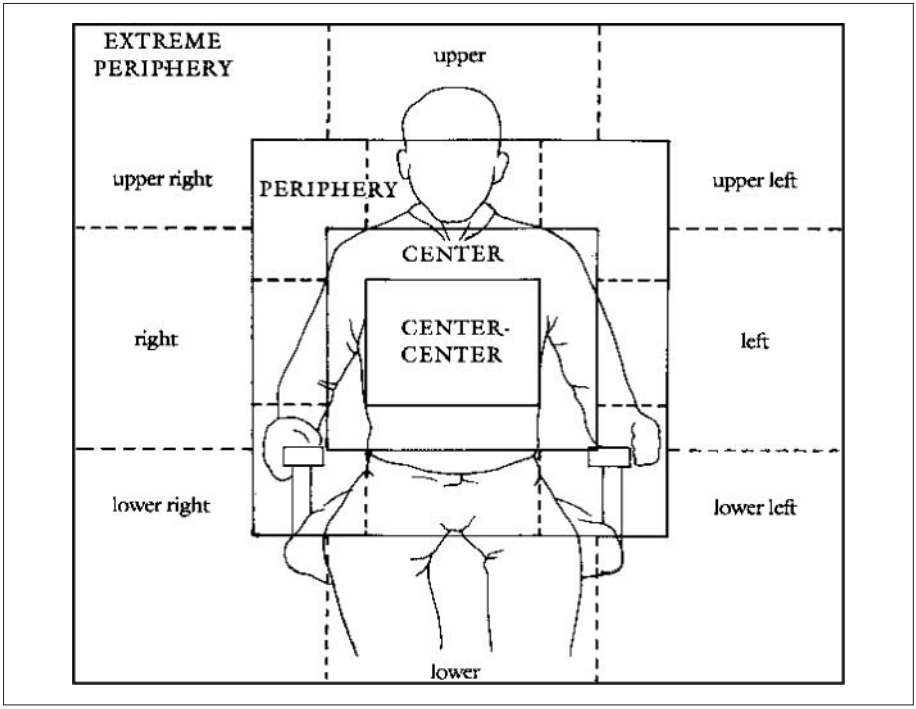


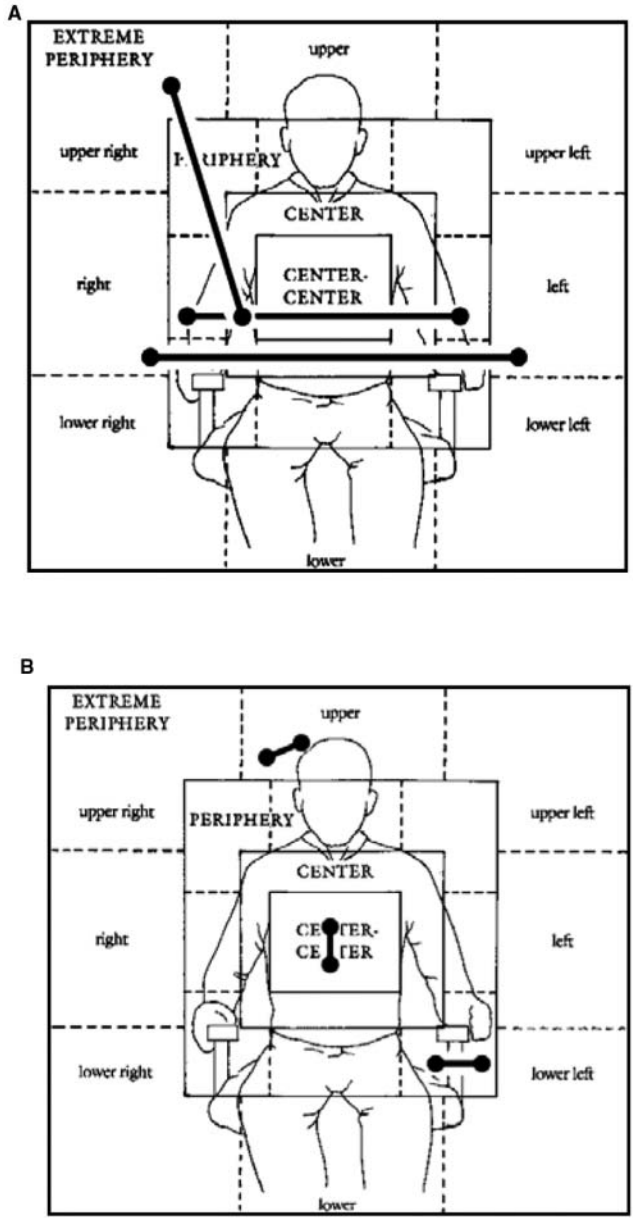
Figure 1. McNeill's gesture space diagram

categories of gesture type were established—High span/Low span, C-VPT/O-VPT. "High span" are gestures that cross at least two major boundaries on the gesture space diagram (see Figure 1) and "Low span" are gestures that cross no major boundaries.

Major boundaries are the solid line boundaries on the diagram in Figure 1, which separate the four main areas—"Center Center," "Center," "Periphery," and "Extreme Periphery." The dotted lines were not included as boundary divisions for the purpose of this exercise. Figure 2A to D shows the spans of the 12 gestures scripted for the video clips. Some gestures were produced in the central space, others in the peripheral space, and others crossed between both spaces.

Experimental Procedure

On entering the laboratory, participants were asked to take a seat in front of the computer monitor and eye tracker. After reading and signing the information and consent forms, they were told to sit comfortably in the chair, look forward at the middle of the screen and keep their heads still while the eye tracker was set up to track their right eye. The experimenter located the participant's right eye on their screen by illuminating the infrared camera and adjusting the "Pan/Tilt" of the eye tracker. By adjusting the pupil



(continued)

Figure 2. (continued)

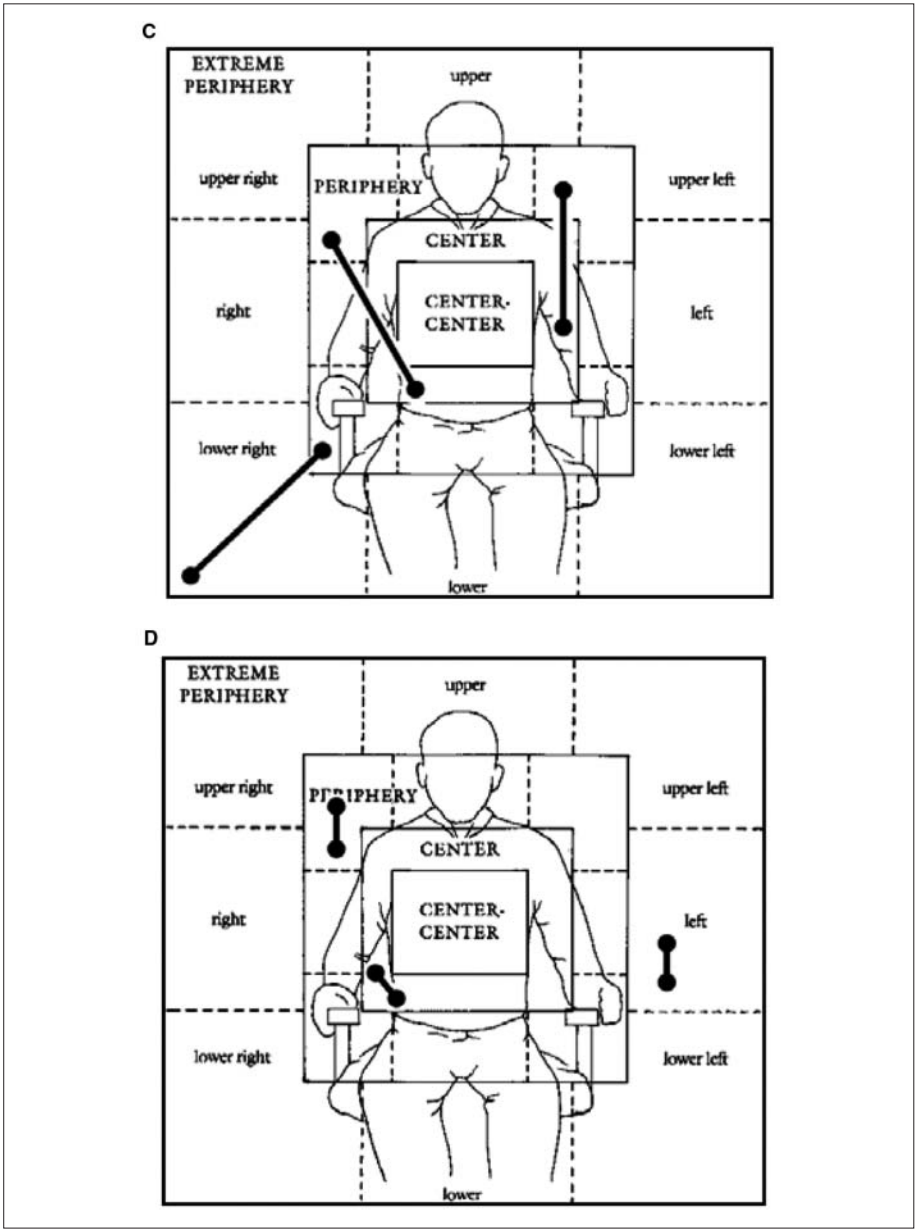


Figure 2. (A) Spans of the three high-span O-VPT gestures. (B) Spans of the three low-span O-VPT gestures. (C) Spans of the three high-span C-VPT gestures. (D) Spans of the three low-span C-VPT gestures

and corneal reflection configuration controls on the screen, the experimenter enabled the eye tracker to obtain optimum recognition of the pupil and cornea for accurate tracking of the eye. Once these positions were recognised, the experimenter switched the "Pan/Tilt Tracking" radio button to "Auto," allowing the camera to track the eye when the participant made slow head movements, keeping the eye in the middle of the experimenter's screen.

Next, the calibration chart was brought up on the monitor in front of the participant, and the experimenter instructed the participant to look at each of the nine numbers in turn on the screen while the eye was calibrated. This process was necessary, as it provided the data that allowed the eye tracker processor to account for differences between individual participants. Once the calibration was completed and checked, the recording equipment in the adjacent laboratory was started to record the output as the participant watched the video clips. A clip order was selected for the participant by the experimenter. In an attempt to retain the accuracy of the calibration throughout the experiment, the participant was reminded that, to answer questions during the experiment, they had only to press the "Y" and "N" keys on the keyboard, so if possible they should try to keep their hand resting over the keys so that they did not have to look down every time they answered a question. Participants were then asked to start the experiment in their own time by pressing the "space" bar.

Output was in the form of a video recording of the computer screen as the participant had seen it during the experiment, with a small black fixation marker overlaid, which denoted where the participant had been looking while they were attending to the video clips. The fixation marker shown in the output moved around the screen to represent the participant's point of gaze on *their* screen as they watched the video clips. Gaze fixations, blinks, saccades, and other eye movements could all be distinguished from the output for each individual. Figure 3A to D is an example of the output from four of the clips presented to the participants; the frames selected illustrate the stroke phases of one gesture from each of the four gesture categories, indicating the differences between high-span and low-span gestures. The black fixation marker can be clearly distinguished on each frame, indicating fixation of the face, the gesture and so on.

Coding Areas of Fixation

Coding was carried out independently by two experimenters (20% of the data were coded by both participants yielding a Cohen's κ of .90, indicating that the scoring was highly reliable). The full recording of each participant's eye gaze during the experiment was converted into individual frames using WinFF software. The frames could then be viewed one by one and scrolled through, using IrfanView. The conversion rendered 25 frames for each second of the recording, so each frame represented a time span of 40 ms (or 0.04 s) from the original data. The 12 gestures scripted for presentation to participants contained an assortment of different gesture properties other than span and viewpoint. Some gestures were performed in the peripheral gesture space, whereas others were performed centrally (see Figure 2A-D for the McNeillian gesture

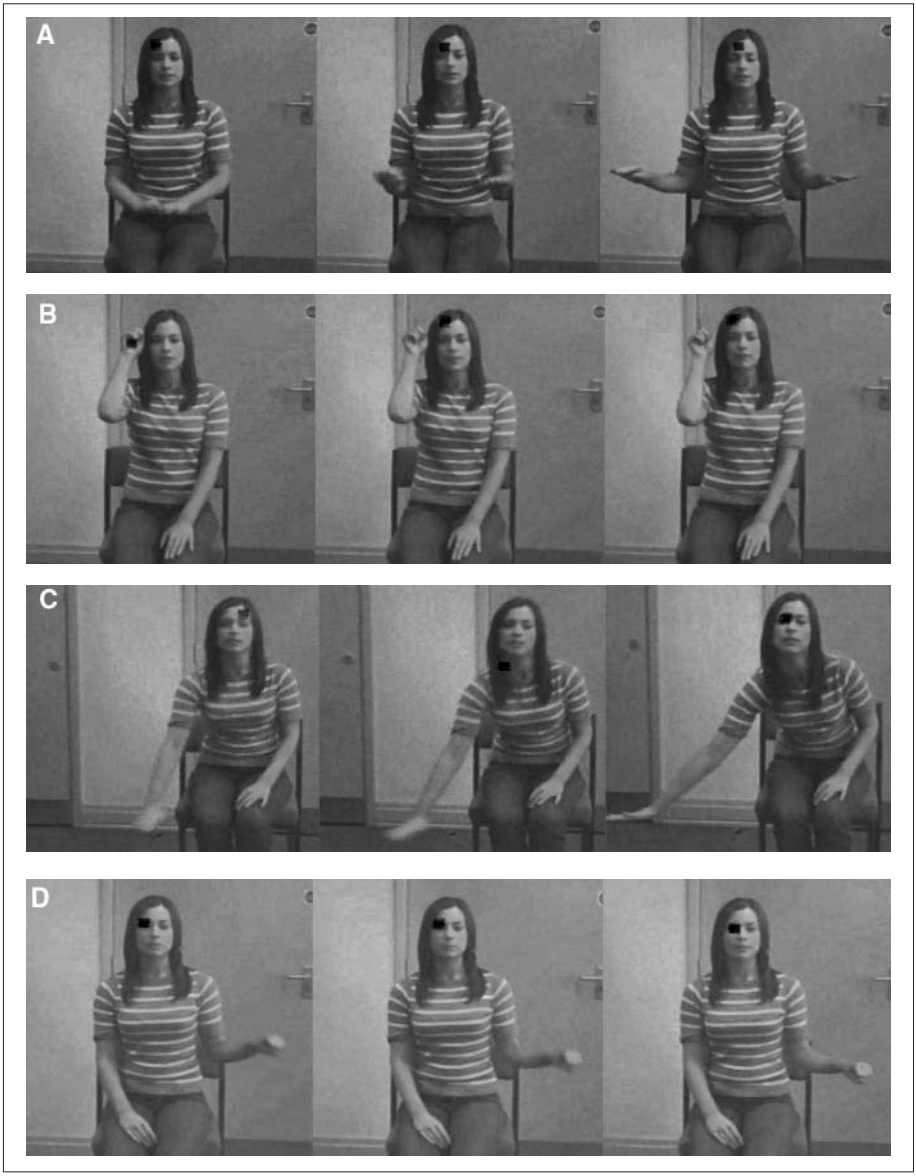


Figure 3. (A) High-span O-VPT gesture. (B) Low-span O-VPT gesture. (C) High-span C-VPT gesture. (D) Low-span C-VPT gesture

space diagrams showing the span of the different gestures). Some gestures included "holds" and "speaker fixation," both of which have been found to increase levels of decoder fixation (Gullberg & Homqvist, 2002).

Table 1. An Example of the Scoring of Fixation Areas for One Participant

Gesture Type	Start and End of Clip	Frames	Area Fixated	Gesture	Gesture Fixated?
C-H	7,153-7,376	7,165-7,168	Background	7,244-7,287	Y
		7,169-7,244	Face		
		7,247-7,269	Gesture		
		7,273-7,376	Face		

The first stage of coding involved the scoring of each different area fixated by the participants while they were watching the 12 video clips. Before coding of the fixation areas began, it was necessary to differentiate between different areas that could potentially be fixated. Six categories were devised for this purpose: face, torso, hand gesturing, other body, background, hand still (not gesturing); these were labelled 1 to 6 for ease of scoring. The frame numbers marking the beginning and end of each of the 12 clips were recorded in detail for each participant, as were the frames that marked the beginning and end of each of the 12 gestures (stroke phase only, using McNeill's classification, 1992) and the frame numbers at the beginning and end of each period of fixation on any of the six areas determined above. For example, the scoring for one gesture viewed by Participant 1 is illustrated in Table 1.

An area was counted as fixated if the fixation marker remained on that same area for at least three video frames (representing a time of 120 ms or 0.12 s), following Gullberg and Holmqvist (2002).

It is evident from the example in Table 1 that there was a slight delay between the beginning of the video clip and the participant's gaze moving to the first area of fixation (in this example, the interval between frames numbers 7,153 and 7,165). As the gaze moved between different areas of fixation, the fixation marker was in motion and so was not counted as fixating any particular area until it had settled to a clearly defined black marker. Figure 4 is an illustration of one participant's fixation behaviour during their viewing of a single video clip. Their sequence of fixations began on the face (1), moving to other areas such as the background (2) and the hand gesturing (7), and returning to the face frequently in between non-facial fixations.

Next, the frame numbers were converted to time durations, with each frame representing 40 ms of real time. This enabled the experimenters to work out the length of each video clip, the proportion of time participants spent fixating different areas in each clip, how long the stroke phase of the different types of gestures themselves lasted for, and so on. Information is only extracted from a scene during fixations. Short deviations from the area of fixation, caused by saccades or eye blinks lasting just a few frames, were not coded. Therefore, the time intervals represented general time spent fixating one particular area before the eye gaze moved to a distinctly different area, regardless of small divergences such as blinks and saccades. If the tracker was absent from the original area of fixation for longer than a few frames and if it moved in a different direction far away from the area, not just straight down and back up (as for a

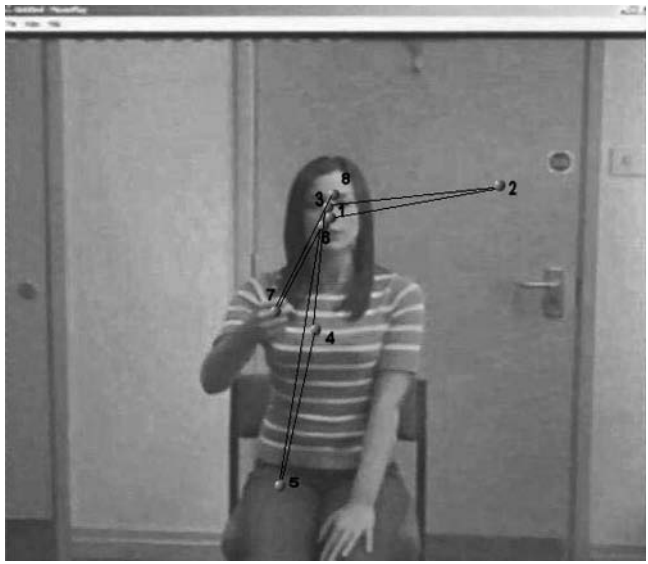


Figure 4. The fixation behaviour of a participant during the viewing of one video clip

blink), the area was coded as "Other," and the frame numbers were marked so that these times could be taken into account for working out average fixation times.

Coding Participants' Answers

Each participant answered 24 "Yes/ No" questions in total, two questions following each video clip and relating to the information encoded by the gesture presented in the clip. Answers for the four different categories of gesture (O-H, O-L, C-H, and C-L) were considered separately. The percentage of correct answers for the questions (six in total) relating to the three gestures of each category was calculated for each individual participant, rendering 10 percentage scores for each gesture category. These percentages of correct answers could then be compared with the percentage of time participants spent fixating the gestures in the different categories.

Results

Where Do Participants Look?

The results show that participants spent most time fixating the speaker's face (an average of 84.9% of participants' overall looking time), followed by other regions of the

Table 2. Chi-Square Test on How C-VPT Gestures Were Fixated Overall

	Fixated	Not Fixated
High span	9	21
Low span	17	13

Note: C-VPT = Character-Viewpoint.

body (2.7%), then gestures (2.1%), and then background (0.5%). 9.8% of the overall time was coded as participants looking at "Other." The "Other" category refers to those times in between direct fixations, times when the participants looked away from the screen, or when the calibration of the eye was too weak for the eye tracker to pick up an accurate fixation point.

Which Gestures Are Looked at Most?

The low-span C-VPT gestures attracted more fixations than any of the other categories of gestures (32.7% of fixations of all gestures), although not significantly so. High-span O-VPT and low-span O-VPT gestures attracted the same number of fixations from participants (25.0%) and high-span C-VPT attracted the least fixations (17.3%). A chi-square test (see Table 2) shows that with C-VPT gestures, the low-span gestures were significantly more likely to be fixated than the high-span gestures ($\chi^2 = 4.34$, $df = 1$, $p < .05$, two-tailed).

One possible hypothesis to explain the different number of fixations that each of the categories of gesture attracted is that the gestures from the various categories were located in different parts of the gestural space with greater or lesser proximity to the natural focus of visual attention, the gesture itself. If participants spent approximately 85% of the time focussing on the face, then any gestures occurring in regions adjacent to the face might well have a special premium in attracting participants' attention to the face. To test this hypothesis, the exact onset and offset position of each gesture was identified and the midpoint of the gesture located. The distance from the centre of the face to this midpoint was measured and converted to a standardized scale. The four categories of gesture displayed the following pattern—the high-span and low-span O-VPT were characterised by the lowest distance (and the same in each case), the low-span C-VPT was 1.235 times higher than this minimal distance, and the high-span C-VPT was the highest, with a ratio of 1.415 times the minimum. In other words, the positioning of the gestures in the gestural space relative to the face cannot easily account for the pattern of fixation observed, because although the high-span C-VPT gestures had the lowest proportion of overall fixations (17.3%) and, in addition, were the set of gestures that were furthest from the modal (facial) focus, the low-span C-VPT gestures had the highest proportion of fixation (32.7%) and yet were 1.235 times further from the facial focus than either of the two categories of O-VPT gestures. In other words, there is no strict correlation between distance from facial focus and the efficacy of different types of

Table 3. The Average Duration of Gesture Fixations for Each Gesture Category

	Observer Viewpoint	Character Viewpoint
High span	180 ms	130 ms
Low span	120 ms	310 ms

Table 4. The Average Percentage of the Stroke Phase of the Gesture that Participants Fixated

	Observer Viewpoint	Character Viewpoint
High span	12.2	9.4
Low span	16.7	26.5

gestures for attracting visual attention, although there is the possibility that if gestures occur at too great a distance from the facial focus (like the high-span C-VPT gestures) then attention might not divert to them on time.

How Long Is Spent Looking at Each Type of Gesture?

Low-span C-VPT gestures were looked at for the longest period of time and for the highest percentage of the duration of their stroke phase (see Tables 3 and 4). A two-way analysis of variance (with the variables span and viewpoint) on the average *percentage* of the stroke phase fixated reveals that there is a significant main effect for span ($F = 6.457$, $df = 1$, $p < .05$) but no significant main effect for viewpoint ($F = .620$, $df = 1$, n.s.) and no significant interaction effect between span and viewpoint ($F = 2.112$, $df = 1$, n.s.). A two-way analysis of variance on the average *duration* of fixation on the stroke phase, however, shows that there is a significant interaction effect between span and viewpoint ($F = 5.003$, $df = 1$, $p < .05$) but no significant main effect for span ($F = 1.814$, $df = 1$, n.s.) or for viewpoint ($F = 1.381$, $df = 1$, n.s.). On average, participants fixated low-span C-VPT gestures for 26.5% of the duration of their stroke phase and looked for an average of 0.31 s. Low-span O-VPT gestures received the next highest percentage-of-duration fixation, at 16.7% of the stroke phase. The gestures that were fixated for the shortest percentage of the duration of their stroke phase were the high-span C-VPT gestures, which were only looked at for 9.4%, although on average the high-span C-VPT gestures were fixated for slightly longer in terms of actual time (130 ms) than the low-span O-VPT gestures (120 ms).

How Did Participants' Fixation Styles Vary?

Of the 120 gesture presentations, 52 were fixated by the 10 participants. Naturally, there was a great deal of variation in the length of time that participants fixated each

Table 5. An Illustration of One Participant's Onset and Offset of Fixation Relative to a Selection of Gestures Stroke Phases

Gesture (2)	**on** **off**
"… his boat gets **[whooped over and turned upside down]** and he's out of the boat."	
[Right hand is in front of body, fingers curved and apart and palm facing upwards; arm moves round in a circle right and upwards, ending with elbow out at a right angle to the body and palm facing forwards with fingers splayed]	
Gesture (5)	**on** **off**
"… got a little hat on **[with a flower]** coming out of it …"	
[Right hand is at right side of head, palm facing forwards with fingers extended towards head; fingers move towards the right, away and slightly down from the head]	
Gesture (10)	**on** **off**
"The old man's holding a balloon and he's **[got a walking stick in]** the other hand …"	
[Left hand is in a fist as if holding a stick, with back of hand facing upwards; hand is out to the middle of the left side of the body and moves downwards slightly]	

gesture presented to them. Some were quick to move their visual attention to the gesture once the stroke phase began, presumably having been attracted by the preparation phase of the gesture and coordinating their fixation perfectly with the beginning of the stroke phase, although no participants were found to begin fixating before the start of the stroke phase itself. Other participants took longer to begin fixating the gesture. Some moved their gaze away from the gesture before the stroke phase had finished, whereas others fixated the gesture for its full length, even continuing to fixate the area in which the gesture had been performed after the hand or hands had started to move away. Of course, there were those who did not fixate certain gestures or, in the case of one of the participants, any gestures at all. Table 5 is a representation of Participant 5's fixation pattern, giving an indication of the onset and offset of their fixations on a selection of the gestures. Gesture 2 is a high-span O-VPT gesture, Gesture 5 is a low-span O-VPT gesture, and Gesture 12 is a low-span C-VPT gesture. Participant 5 did not fixate any of the three high-span C-VPT gestures presented. The words "on" and "off" above the speech give an indication as to the participant's onset and offset of fixation relative to the stroke phase of the gesture, which is indicated by square brackets. Participant 5 fixated five out of the 12 gestures, which themselves varied greatly in length (between 0.56 and 1.88 s). Participant 5's fixation onsets had a range of +0.28, to +0.40 s relative to the beginning of the stroke phase. The fixation offsets ranged between –0.80 and 0 s relative to the end of the stroke phases; in other words, each one of Participant 5's fixations finished either before the end of the stroke phase of the gesture or at the same time as the end of the stroke phase. The range of onset of all fixations (for all participants) was from 0 s (i.e., fixation begins at the same time as the start of the stroke phase) to +1.40 s (i.e., fixation begins 1.4 s after the onset of the stroke phase of the gesture). The range of offset of fixations was from –3.08 s (i.e., fixation ends 3.08 s before the stroke phase of the gesture ends) to +1.36 s (i.e., fixation ends 1.36 s after the stroke phase of the gesture ends).

Fixation Pattern and the Extraction of Semantic Information

One crucial question is the relationship between the specific attentional focus and the extraction of semantic information from the gestures. The approach of asking two direct probe questions has been successfully applied before (Beattie & Shovelton, 1999a), and in this experiment, it was found that low-span C-VPT gestures were the most communicative (83.3%), followed by high-span C-VPT and high-span O-VPT (both 80.0%). The lowest score was for low-span O-VPT (71.7%). Of course, low-span C-VPT gestures not only attracted the highest number of fixations (32.7% of fixations of all gestures were on this category) but was also categorised by the highest average duration of fixation (310 ms) and the highest average percentage of the stroke phase of the gesture that participants fixated (26.5%). In other words, this pattern of results suggests that there may well be a relationship between attentional focus and the uptake of information from categories of gestures and that certain types of C-VPT gestures seem to be particularly effective at drawing in the gaze of the interlocutor. This may well have significant implications for understanding why certain types of gestures emerge as being particularly communicative.

There was, however, no significant correlation between the average accuracy score and the average duration of fixation across the 10 participants for any of the four gesture types. The reason for this would seem to be attributable to the underlying distribution of the semantic probe scores and the fact that these data were clumped rather than evenly distributed. A ceiling effect was reached, in fact, in 20% of all cases (despite being on the surface relatively testing questions), but the rest of the data were not normally distributed. In addition, even in the case of the low-span C-VPT gestures, the ones characterised by the highest proportion of the stroke phase with direct visual attention, there was no significant correlation between the percentage of stroke phase fixated and the percentage accuracy (Spearman's rank correlation coefficient = .255). In other words, even with these gestures that attract the most visual attention, there is no obvious correlation between the amount of gaze on the stroke phase of the gesture and the successful uptake of information from the gestures across participants. There is clearly an underlying pattern here between general fixation of types of gestures and the successful recovery of the semantic information from the different types, but it will require additional research to demonstrate a consistent pattern on a participant by participant basis. Of course, general measures of amount of gaze at gestures or even the proportion of stroke phase fixated may not be the most appropriate metrics in this case. Perhaps, the attention of the listener is cued much more directly by a combination of complex verbal and gestural cues to just the right transitory moments in the gesture, those moments where the gesture carries the critical bit of information. Scrutiny of the patterning of listener gaze in Table 5 suggests that there might well be something in this, but any first investigation must, of course, start with the more basic metrics first. And further experimental research will be needed to systematically manipulate the salience of individual elements of a story to determine the effects of this manipulation on both the patterns of gestural production and the specific patterns of listener fixation on the gestures. Such a manipulation would produce

much stronger conclusions than merely trying to analyze patterns of information content in the message on a post hoc basis. This manipulation may well wish to consider pragmatic factors around the generation of the utterance that could influence the salience of otherwise insignificant elements of the story. The results of such an investigation could be very significant indeed.

Discussion

One extremely influential model within the area of communication (McNeill, 1992) maintains that speech and gesture together convey semantic information in everyday talk. One major issue for this model is how this information is extracted by listeners from each modality and combined (seemingly) unconsciously and effortlessly. This study attempted to provide an answer to one small aspect of this by analysing the patterns of visual attention to the iconic gestures that accompany speech. Previous research has demonstrated conclusively that the human face is the primary focus of attention in conversation (particularly the eyes and the mouth). But how and when does this visual attention move to the gestures that accompany speech? And what are the consequences of this attentional patterning for the extraction of semantic information from these dynamic gestural movements?

This study used an analogue task, tracking visual attention on a frame-by-frame basis as participants watched an actor speaking and producing a number of highly scripted and choreographed gestures of different types. The results revealed that participants spent 84.9% of the time fixating the face, followed by other regions of the body (2.7%), then the gesture (2.1%). However, if one considers the proportion of the stroke phase of the gesture (the meaningful part of the gesture) that is fixated, this varies between 9.4% for some types of gestures to 26.5% for others (notably low-span C-VPT gestures). In some individual cases, the proportion of the stoke phase of the gesture actually fixated rose to 92%. Some types of gestures do successfully divert the attention of the interlocutor and interestingly the gestures that seem to be most effective in this regard are low-span C-VPT gestures. The reason that this is potentially significant is that C-VPT gestures have been shown to be more communicative than O-VPT gestures (Beattie & Shovelton, 2002) and the fact that there is more direct fixation on these gestures may be crucial in helping us to understand this result. Low-span C-VPT were most likely to be fixated (32.7% of the fixations of all gestures were on this category), they were also characterised by the highest average duration of gesture fixation (310 ms, compared with 180 ms for high-span O-VPT gestures and 120 ms for low-span O-VPT gestures).

This study also demonstrated that the low-span C-VPT gestures were the most communicative with an average accuracy score of 83.3%, but there was no obvious correlation between this accuracy score and the average duration of this fixation or between this accuracy score and the average proportion of stroke phase fixated, across individual participants. But this is probably as much to do with the nature and underlying distribution of the accuracy scores as much as anything else.

In summary, this research has shown how the visual fixation patterns of a listener shift while they attend to a speaker. The visual attention of the listener moves, apparently unconsciously and effortlessly, to the gestural movements of the speaker while they watch him or her speak. These eye movements are crucial to the whole communication process, because there is information in these gestures that cannot be obtained from any other source. It would seem that the shorter C-VPT gestures are more effective than other types of gesture at eliciting this visual attention of the listener and at communicating the information within them most effectively (and not because these gestures are located closer in the gestural space to the natural focus of attention—the face). More than a quarter of the stroke phases of these gestures are fixated as the eyes move a distance from the face to focus on these meaningful movements. But, as usual in research, the study raises as many questions as it does answers. What governs the precise micropatterning of the participants' gaze within the gesture? Even in the case of the low-span C-VPT, 26.5% of the gaze is fixated. Why this 26.5%? Why do high- and low-span C-VPT gestures differ so dramatically in terms of fixation patterns? What are the constraints on the extraction of information in peripheral vision? These are important and significant questions.

This study in many ways just represents the first tentative step in this new area. But it does need to be said that any model of human communication as far-reaching and as radical as that of McNeill (1992), which argues that speech and gesture together really are essential components of the same basic process, does need good empirical data to show that the model has psychological plausibility and that we can demonstrate that our minds really do know when and where to look when we listen to another human being talk, to fully understand what he or she is trying to say.

Acknowledgement

We would like to thank Andrew Stewart for his expert assistance in using the eye-tracking equipment. In addition, we would like to thank Howard Giles and two anonymous referees for their valuable comments on an earlier draft of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research and/or authorship of this article:

Grant RES-000-22-1917 from the Economic and Social Research Council.

Note

1. Twenty-four different gestures were originally scripted and filmed; subsequently, the three clearest examples of each of the four gesture categories were chosen for presentation to participants. For example, those that were especially high or low in span were chosen over those that were more ambiguous in span.

References

Argyle, M. (1967). *The psychology of interpersonal behaviour*. Harmondsworth, UK: Penguin.

Beattie, G. (2003). *Visible thought: The new psychology of body language*. London: Routledge.

Beattie, G., & Shovelton, H. (1999a). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental. investigation. *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. (2001). An experimental investigation of the role of different types of iconic gesture in communication: A semantic feature approach. *Gesture, 1*, 129-149.

Beattie, G., & Shovelton, H. (2002). An experimental investigation of some properties of individual iconic gestures that affect their communicative power. *British Journal of Psychology, 93*, 179-192.

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.

Beattie, G., Webster, K., & Ross, J. (in press). Do speakers really unconsciously and imagistically gesture about what is important when they are telling a story? *Semiotica*.

Fuchs, A. F. (1971). The saccadic system. In P. Bach-y-Rita, C. C. Collins, & J. E. Hyde (Eds.), *The control of eye movements* (pp. 343-362). New York: Academic Press.

Gullberg, M. (2003). Eye movements and gestures in human face-to-face interaction. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eyes: Cognitive and applied aspects of eye movements* (pp. 685-703). Oxford, UK: Elsevier.

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. (2002). Visual attention towards gestures in face-to- face interaction vs. on screen. In I. Wachsmuth & T. Sowa (Eds.), *Gesture and sign language based human-computer interaction* (pp. 206-214). Berlin, Germany: Springer-Verlag.

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*, 53-82.

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

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

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. Frohlich (Eds.), *Gesture and sign language in human-computer interaction* (pp. 49-59). Berlin, Germany: 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.

Rayner, K. (1998). Eye movements in reading and information processing: Twenty years of research. *Psychological Bulletin, 124*, 372-422.

Turano, K.A., Geruschat, D. R.& Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research, 43*, 333-346.

Bios

Geoffrey Beattie is Head of School and Dean of Psychological Sciences at the University of Manchester. He got his PhD from the University of Cambridge and is a past recipient of the British Psychological Society's Spearman Medal (for "published psychological research of outstanding merit"). In 2005-2006 he was president of the Psychology Section of The British Association for the Advancement of Science. He is also a published novelist and the author of many books of non-fiction, which have been shortlisted for a number of major national and international literary prizes.

Kate Webster was a research assistant in the School of Psychological Sciences at the University of Manchester on this Economic and Social Research Council–funded project on gestures and speech.

Jamie Ross was a research assistant in the School of Psychological Sciences at the University of Manchester.