JOURNAL OF COMPUTER-MEDIATED COMMUNICATION

Seeing and Being Seen: The Effects on Gesture Production

Lisette Mol
Emiel Krahmer
Alfons Maes
Marc Swerts

Department of Communication and Information Sciences, Tilburg University

Speakers are argued to adapt their language production to their addressee's needs. For instance, speakers produce fewer and smaller hand gestures when interlocutors cannot see each other. Yet is this because speakers know their addressee cannot see them, or because they themselves do not see their addressee? By means of computer-mediated communication we manipulated these factors independently. We found that speakers took into account what their addressee saw. They produced more and larger gestures when they knew the addressee could see them. Seeing the addressee increased gesture production only if speakers could readily interpret the addressee's eye gaze, which is not usually the case in mediated interaction. Adding this affordance resulted in gesturing being similar in mediated and unmediated communication.

Key words: gesture; perspective taking; mediated communication; gaze; theory of mind; eye-catcher.

doi:10.1111/j.1083-6101.2011.01558.x

Language use sometimes requires taking into account what another person can or cannot see. For example, when watching a documentary on Venice with a friend, you might ask your friend "have you ever been there?" where *there* refers to Venice. However, if your friend was in the same room, but working on her computer "have you ever been to Venice?" may be more appropriate. Because you know your friend is not watching the documentary, you may choose a more explicit reference. On the other hand, if you were asked by your friend, "have you ever been there?" while working on your computer, your knowledge of her watching a documentary on Venice may help in arriving at the correct interpretation. Yet would you do so correctly if you happened to be browsing a website on Berlin?

Language production and interpretation are argued to be adapted to our knowledge of another interlocutor, including what the other person knows about and sees (e.g. Grice, 1989). In this paper we focus on language production. Gesture and speech production can both be considered part of a speaker's language production (Kendon, 2004; McNeill, 1992). Therefore, one way of measuring to what extent speakers adapt their language production to their addressee is by looking at the hand gestures people spontaneously produce along with speech. It is well established that speakers produce fewer and different cospeech gestures when interlocutors cannot see each other (Cohen & Harison, 1973), such as on the phone (Bavelas, Gerwing, Sutton, & Prevost, 2008) or on both ends of an opaque

screen (Alibali, Heath, & Myers, 2001). Thus, speakers seem to take into account what their addressee cannot see and thus cannot know about. Yet several empirical studies suggest that interlocutors tend to base their language production and interpretation on *their own* visual perspective, rather than that of their conversation partner (e.g. Keysar, Lin, & Barr, 2003; Wardow Lane, Groisman, & Ferreira, 2006). So do speakers truly take into account what their addressee sees?

Traditionally, in studies of hand gesture production, visibility has been manipulated symmetrically, such as by a screen that keeps the addressee from seeing the speaker, but at the same time keeps the speaker from seeing the addressee. We know that speakers tend to gesture less when visibility is obstructed in this way. Yet is this due to the addressee not seeing the speaker, or because of the speaker not seeing the addressee? Can we correctly apply our knowledge of what the other sees to our language production, or do we tend to base it on our own observations?

To answer these questions, we need to somehow manipulate visibility asymmetrically, such that the speaker's perspective differs from the addressee's. Computer-mediated communication offers this possibility. With communication through web cams, we can separate the factors of seeing the addressee and being seen by the addressee, thus gaining important insights into what knowledge speakers apply to their language production.

To support the validity of this method, we need to draw the comparison between mediated and unmediated communication. Can we generalize from the results found in mediated communication to unmediated communication? What exactly is needed to make these two forms of communication optimally similar? Is it enough for interlocutors to have a live audiovisual presentation of each other? One possibly important difference between face-to-face interaction and communication through web cams is how readily interlocutors can interpret each other's eye gaze. We know that gaze and mutual gaze serve a variety of functions in unmediated interaction (Argyle & Cook, 1976; Kendon, 1967). In the second part of this paper we examine whether being able to interpret each other's eye gazing patterns influences speakers' language production in mediated communication.

Before describing how we used computer-mediated communication to study perspective taking, and presenting our findings on the importance of (mutual) gaze, we provide a brief overview of the literature on the communicative use of cospeech hand gestures, on when interlocutors have trouble taking into account each other's visual perspective, on the comparison of mediated and unmediated communication, and on the role of eye-gaze in interaction.

Background

Communicative Cospeech Gestures

When talking, many people move their hands and arms around, without the objective of directly manipulating their environment: They gesture. Several functions of such hand gestures have been identified, such as facilitating speech production (e.g. Hadar, 1989; Krauss, 1998), supporting learning (e.g. Goldin-Meadow, 2010), and aiding cognition (e.g. Chu & Kita, 2008; Melinger & Kita, 2007). In this paper, we focus on the communicative import of cospeech hand gestures. Numerous studies have addressed the communicative role of cospeech gestures. Some studies have shown that addressees gain information from gesture (e.g. Beattie & Shovelton, 1999; Cassell, McNeill, & McCullough, 1998; Chawla & Krauss, 1994; Cutica & Bucciarelli, 2008; Goldin-Meadow & Sandhofer, 1999; Mol, Krahmer, Maes, & Swerts, 2009). Additionally, other studies have rendered converging evidence that hand gestures are part of a speaker's communicative effort (Kendon, 2004). We will address this viewpoint below.

Kendon (1988) recognizes a continuum of how conventional and language-like hand gestures are. On the one end of this continuum are sign languages, in which signs have a conventional meaning and can be interpreted in the absence of speech. On the other end of the continuum is *gesticulation.* This is the production along with speech, of gestures that are not embedded in the grammatical structure of speech. For example, while saying "he went away," one could move one's arms back and forth along one's upper body, illustrating the manner of the action described. Alternatively, quickly wiggling one's down pointing fingers while moving the hand forward could illustrate the same event. Gestures at this end of Kendon's continuum are the most idiosyncratic gestures and their interpretation is highly dependent on the accompanying speech (Feyereisen, Van de Wiele, & Dubois, 1988).

These cospeech gestures are generally divided into several categories (e.g. McNeill, 1992). One broad distinction can be made based on whether a gesture depicts some of the content of the message a speaker is trying to convey, or whether it rather structures the conversation (e.g. Bavelas, Chovil, Lawrie, & Wade, 1992), or emphasizes certain parts of speech (e.g. Effron, 1941; Ekman & Friesen, 1969; Krahmer & Swerts, 2007). In this paper we focus on gestures that express some of the content a speaker is conveying, which are known as *illustrators* (Ekman & Friesen, 1969), or *representational gestures* (McNeill, 1992). Especially these gestures have been found to be produced differently by speakers in different communicative settings (e.g. Alibali, et al., 2001).

Effects of the Communicative Setting on Representational Gestures

When gestures have communicative potential, that is, when they can be seen by an addressee, they have been found to be larger (Bangerter & Chevalley, 2007; Bavelas, et al., 2008) and more frequently produced (Alibali, et al., 2001; Bavelas, et al., 2008; Cohen, 1977). Yet does this imply that speakers take into account the addressee's visual perspective? In one study, Bavelas et al. (2008) asked participants to describe a picture of an elaborate dress. When participants interacted face-to-face, their gestures about the dress were full sized, as of an actual dress one could wear, whereas when interaction took place over the telephone, gestures were only the size of the dress in the picture. Although this clearly illustrates speakers' sensitivity to the communicative context, we can not be sure that speakers were adapting to what their addressee saw, since gesturing based on what they themselves saw would result in the same behavior.

Alibali et al. (2001) found that when speakers were asked to retell an animated cartoon to an addressee, they produced representational gestures more frequently in a face-to-face setting than when speaker and addressee were separated by an opaque screen. Again, this shows speakers' sensitivity to the environment. Yet it does not tell us whether speakers were taking into account their addressee's visual perspective, since whether or not the addressee saw the speaker corresponded with whether or not the speaker saw the addressee. Therefore, speakers may have based their gesturing on their own visual perspective. Using a similar narration task, Jacobs and Garnham (2007) found that gestures were produced more frequently when a speaker *knew* that information was new to the addressee as well as when the addressee appeared attentive and interested. This suggests that speakers do take into account listener needs, but again these needs were also readily visible to the speaker. (Speakers either retold the same cartoon twice to the same addressee or to two different addressees, and either saw an interested or a less interested addressee.) This also holds for a study by Özyürek (2002), which showed that speakers produce their gestures differently, depending on where the addressee is located relative to them. Although speakers were shown to change the orientation of their gestures based on the addressee's location, we do not know if this resulted from the change in their own visual perspective or in that of their addressee.

In an earlier study (Mol et al., 2009), we manipulated whether speakers thought to be talking to a human addressee or to an audiovisual speech recognition system. In this study, contrary to the

aforementioned studies, the environment of the speaker was exactly the same across conditions. The only difference was in the preceding instruction. The speaker was seated alone in a room in front of a camera and was either told that the audiovisual output of this camera was sent to another participant, or to an artificial system. Speakers were found to gesture more frequently and produce more large gestures when they thought to be addressing a human addressee. This time, the difference in gesturing could only be caused by a different belief about the addressee. Yet it remains unknown whether speakers apply more of their knowledge about the addressee to their language use than just whether the addressee is human.

Although the above-mentioned results all point in the direction that speakers apply their knowledge of their addressee to their hand gesture production, and thereby to their language production, these results leave open the possibility that the actual application of such knowledge is very limited and that speakers mostly use an egocentric perspective. That is, they may base their language production on what they themselves see. We therefore turn to some studies that have shown people's difficulty in applying their knowledge about their addressee's visual perspective to their language use.

Taking Into Account What an Interlocutor Sees

Keysar, Lin, and Barr (2003) have shown that people tend to make 'mistakes' in their interpretation of speech, when arriving at the correct interpretation requires taking into account what a speaker can and cannot see. By studying participants' eye movements, they found that addressees mistakenly considered objects that they knew a speaker could not see as possible referents of speakers' referring expressions. Wardlow Lane, Groisman, and Ferreira (2006) found similar results for reference production. In their study, a speaker had private visual access to an object that only differed from the target object in size. Even though the addressee could not see this competing object, speakers often included a contrasting adjective, such as 'small,' in their reference to the target object, despite this not being informative to the addressee. Surprisingly, they did so even more when instructed to conceal their private information from the addressee. Note that the contrasting adjective provides a cue to the properties of the object that was hidden from the addressee. Therefore, it seems that speakers have difficulty in applying their knowledge of what their addressee sees to their speech production. Similar difficulties may affect speakers' hand gesture production.

Computer-Mediated vs. Unmediated Interaction

Mediated communication can help us resolve the issue of whether speakers employ an egocentric perspective or not, when it comes to adapting their cospeech gestures to a communicative setting. Yet can mediated communication be representative of face-to-face interaction? Social presence theory (Short, Williams, & Christie, 1976) proposes that the extent to which social presence is experienced in mediated communication depends on the affordances offered. The more affordances available, the more warmth and affection interlocutors will experience and express. Social information processing theory (Walther, 1992) adds that interlocutors can also adjust both their motives and their communicative efforts to a medium, such that mediated communication does not necessarily fall short of face-to-face interaction when it comes to experienced presence. For example, Walther, Slovacek, and Tidwell (2001) have shown that seeing a picture of the addressee promotes affection and social attraction in short-term mediated interaction, but that this is not true for long-term mediated interaction. Given ample time, the highest levels of affinity were established through a text-based medium.

Consistent with this approach, Brennan and Lockridge (2006) use the grounding framework to describe how communication is affected by mediation: ''The grounding framework conceptualizes mediated communication as a coordinated activity constrained by costs and affordances (Clark &

Brennan, 1991)'' (p. 1). From this perspective, the more the costs and affordances of mediated communication resemble the costs and affordances of face-to-face interaction, the more similar the two will be. For example, Brennan and Ohaeri (1999) argue that mediated written conversation can be less polite compared to spoken interaction, because the production costs of politeness are higher when typing than when speaking. This in turn could lead to interlocutors perceiving each other differently, rather than these different perceptions resulting from mediation directly. From these frameworks, we can infer that both being able to see the addressee and being seen by the addressee will result in mediated communication being more similar to face-to-face interaction.

Communication through desktop videoconferencing, such as with Skype, offers many of the affordances available in face-to-face communication. The use of web cams and microphones allows speakers to see and hear each other almost real time, even though they are in different locations. Would this result in interlocutors behaving the same way as in face-to-face interaction? Isaacs and Tang (1994) observed interactions between technical experts that took place over the phone, through desktop videoconferencing, or face-to-face. They found that the experts indeed used the visual modality in videoconferencing much like they did in face-to-face communication. ''Specifically, participants used the visual channel to: express understanding or agreement, forecast responses, enhance verbal descriptions, give purely nonverbal information, express attitudes through posture and facial expression, and manage extended pauses'', p. 65. However, Isaacs and Tang also listed some differences between videoconferencing and face-to-face communication, for example, managing turn-taking, having side conversations, and pointing to objects in each other's space were more difficult in videoconferencing.

One difference in affordances between video-conferencing and face-to-face interaction is the availability and interpretability of information from gaze. For example, when interlocutors are not copresent and the physical environment is not shared, the direction of each other's gaze cannot readily be interpreted. When using a web cam, it can even be misleading. Since the image of the conversation partner and the web cam are not in the same location, looking at the web cam means not looking at the other interlocutor. Yet when someone looks into the camera, their image misleadingly appears as though they are looking at the person watching the image. To what extent do the availability of an interlocutor's gaze and the ease with which it can be interpreted influence language production? Can the difference in the availability of information from gaze account for some of the differences found between communicating face-to-face and by means of videoconferencing?

Using Information from Others' Gaze

We know that gaze and mutual gaze serve many functions in unmediated interaction (Argyle & Cook, 1976; Kendon, 1967). Among other functions, gaze can be used to infer if the other person is attending and whether a message is understood, as well as to solicit such signals from the conversation partner. It has also been found that when speakers gaze at their own gestures, addressees are more likely to fixate on these gestures as well, and are also more likely to retain information from these gestures (Gullberg & Kita, 2009). Thus, speakers may use gaze as a way to direct their addressee's attention to their gestures.

Hanna and Brennan (2007) found that addressees use speakers' eye gaze to disambiguate referring expressions. Addressees could do so both when a speaker's visual perspective matched their own perspective and when it was a mirror image, showing that they could map the speaker's visual perspective onto their own. Brennan, Xin, Dickinson, Neider, and Zelinsky (2008) found that participants were able to benefit from seeing another participant's gaze indicated on their screen, when performing a simple collaborative search task. Seeing each other's gaze represented by a cursor on the display was shown to allow for a more optimal division of labor than did talking to each other. This shows that participants were able to adapt their behavior, based on their knowledge of where their partner was looking.

These studies show that people can sometimes benefit from observing other people's gaze when communicating or cooperating, both in unmediated and mediated settings. They also show that the interpretation and production of gaze can be adapted to a task or a medium. How important is this factor when it comes to the difference between mediated and unmediated communication? Do interlocutors behave differently dependent on whether gaze is easily interpreted or not, or do they interpret gaze correctly independent of the effort involved, resulting in similar communicative behavior (including gesture production)?

Present Study

Our present study consists of three experiments. First, we investigate whether the fact that speakers produce fewer hand gestures when interlocutors cannot see each other is due to the speaker not seeing the addressee, to the addressee not seeing the speaker, or to both of these factors. We do so by asking participants to perform a narration task in one of four settings, in which we independently manipulate visibility of the speaker and addressee, by means of communication through web cams. Second, we investigate how important it is for speakers to be able to readily interpret their addressee's gaze. For this we make use of a newer videoconferencing technique: the Eye-Catcher (GreenEyes, 2007). This device enables interlocutors to interpret each other's gazing behavior more readily than when using web cams. We measure how this affects gesture production. Third, we test what the differences in speakers' production behavior due to this difference in the interpretability of gaze, mean to naïve observers.

Study 1: Seeing and Being Seen Through Web Cams

In order to manipulate visibility asymmetrically, we make use of videoconferencing through web cams. Our communication task is chosen such that the differences found by Isaacs and Tang (1994) between videoconferencing and face-to-face interaction are minimized. There are only two interlocutors, so there is no possibility of having side conversations. We use a task in which a speaker retells an animated cartoon to an addressee, who is instructed not to interrupt (after Alibali, et al., 2001). Therefore, there is little need for coordinating turns. Also, this task does not relate to the physical environment of either the speaker or the addressee, so there is no need to point at real objects in the environment. Therefore, for this task, the costs and affordances of videoconferencing are a close match to face-to-face interaction. Hence, we expect that manipulating mutual visibility will have similar effects in our mediated settings as it does in unmediated settings.

The use of gesture production as a dependent variable enables us to measure how participants' communicative behavior is influenced by the communicative setting, rather than how they subjectively experience it. We can look at both the frequency and the quality of the gestures produced. Both these factors have been related to communicative effort, and speakers are known to adjust these aspects of their communicative behavior to whether or not there is mutual visibility. Therefore, gesture is a suitable measure for determining if speakers tend to base their language use on their own visual perspective or if they correctly apply their knowledge of the addressee's visual perspective.

Seeing the addressee could influence gesture production for several reasons. One reason is that speakers may base their gesturing on their own visual perspective, and will therefore gesture as though there is mutual visibility when in fact they can only see the addressee. This would mean that speakers gesture more when they can see the addressee, regardless of whether the addressee can see them. If speakers solely use their own visual perspective, this would also mean that when speakers cannot see their addressee, they will produce an equal number of gestures, irrespective of their knowledge of whether the addressee can see them.

Another reason why seeing the addressee may influence gesture production is because of the signals the speaker receives from the addressee. In this case, seeing the addressee may elicit more gestures if it motivates speakers to put in more communicative effort, for example because they experience a higher degree of social presence (Short, et al., 1976) or affinity (Walther, et al., 2001), or because the addressee seems more interested (Jacobs & Garnham, 2007). Yet receiving cues from the addressee may also reduce gesture production, especially when the addressee's gaze is hard to interpret, which may cause the addressee to appear inattentive. Regardless of its direction, this effect would be independent of the effect of being seen by the addressee.

Being seen by the addressee can only influence gesture production directly if speakers correctly apply their knowledge of the addressee's visual perspective, thereby possibly overriding or replacing an egocentric perspective. If speakers indeed base their language production on their knowledge of the addressee's perspective rather than their own visual perspective, this would mean they gesture more when they can be seen by the addressee, rather than when they see their addressee.

Method

Design. We used a 2 × 2 between participants design in which we manipulated whether or not the speaker could be seen by the addressee and whether or not the speaker could see the addressee, see Figure 1. In all conditions speaker and addressee could hear each other.

Participants. Thirty-eight (21 female) native Dutch speakers, all students from Tilburg University, participated in this study as part of their first-year curriculum. Two participants were excluded from our analysis (see Analysis). The remaining 36 participants (20 female) had a mean age of 22, range 18–33. The addressee was a female confederate, who also was a student from Tilburg University.

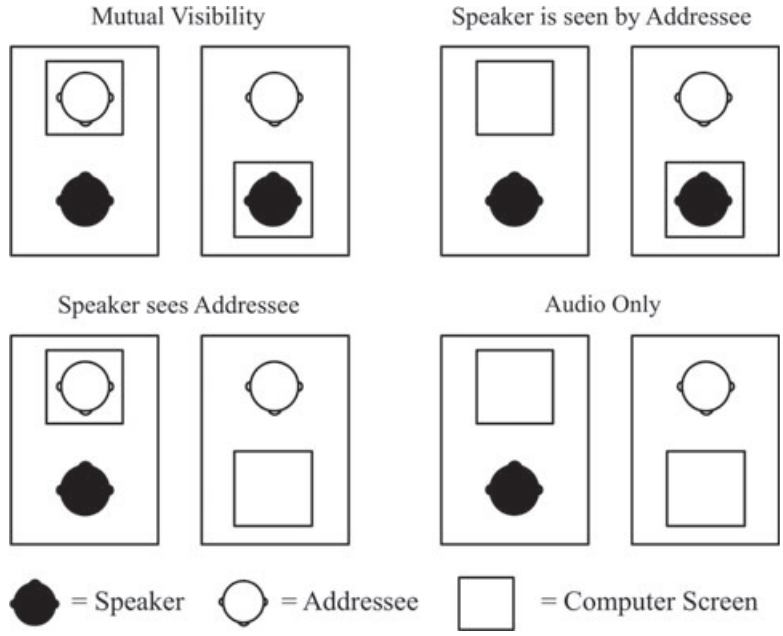

Figure 1 Schematic overview of the settings. Speaker and addressee are seated in separate rooms and can sometimes see the other interlocutor on a monitor

Procedure. The experimenter received the participant and the confederate in the lab and assigned the role of speaker to the participant and the role of addressee to the confederate. Narrators were asked to retell the story of an animated cartoon (*Canary Row* by Warner Bros.) to the addressee. After reading the instructions, participants could ask any remaining questions. The confederate always posed a clarification question. The narrator's instructions stated that the addressee had to summarize the narration afterward and explained that the narrator was videotaped to enable comparison of the summary and the narration.

When all was clear the narrator was seated behind a table with a computer screen on it, which in some settings showed a live video-image of the addressee (full screen), and in the remaining settings showed just the interface of the videoconferencing application we used (Skype). If the addressee was shown, the entire upper body of the addressee was visible, rather than just the face. The computer screen was connected to a computer, which also had a web cam connected to it. Behind the table stood a tripod, which held the web cam and a digital video camera. The position of the web cam was such that the entire upper body of the speaker was captured. On the wall behind the video camera were eight stills from the animated cartoon, one from each episode, as a memory aid for the narrator and to elicit more structured and hence more comparable narrations.

The experimenter took the addressee to another room with a similar setup (but without the stills) and established a connection between the two computers over the internet, using Skype. Sound and video were both captured by the web cams and sound was played back through speakers. To familiarize the participants with the setting, sound was tested by the narrator and addressee talking to each other and if applicable, the video image was tested as well. The connection was then suspended temporarily, while the narrator was left alone to watch the animated cartoon on a different computer. When the cartoon had finished the experimenter re-established the connection, and seated the narrator behind the camera. In conditions where the addressee could see the narrator, narrators were briefly shown what the addressee saw. In the remaining conditions, the experimenter repeated that the addressee could not see the narrator. The experimenter then started the video recording and left the room.

When the narrator was done telling the story, participants completed a questionnaire, which included questions on how the communicative setting had been experienced, how interested the addressee had appeared, and whether any deception was suspected. Meanwhile, the addressee ostensibly wrote a summary on yet another computer in the lab room. None of the participants had suspected any deception. After filling out the questionnaire, they were fully debriefed. All of the participants gave their informed consent for the use of their data, and if applicable for publishing their photographs.

During the narration, the confederate refrained from interrupting, laughing, etc. When necessary, minimal feedback was provided verbally. She always gazed somewhere near the web cam capturing her, independent of whether she could see the speaker.

Transcribing and Coding. We transcribed each speaker's narration from the videotape. Filled pauses, such as 'uh' were included in the transcription. A Perl script was used to compute the number of unique words in each narration.

We coded all hand gestures produced by speakers. Based on the gesture's shape and the accompanying speech, we coded whether a gesture depicted some of the content of the animated cartoon, or whether it was about the current conversation, e.g. placing emphasis. In our analysis we focus on the former category, which we refer to as *representational gestures*. Figure 2 depicts two examples of our coding. In the scene on the left, the speaker imitates a hitting motion while talking about someone hitting. In the scene on the right, the speaker refers to the main character and briefly moves his fingers up and down.

We also coded the size of each gesture. Gestures that were produced using only the fingers received a score of 1. If the wrist was moved significantly the gesture received a score of 2. Gestures that also

Figure 2 Left: example of a representational gesture (depicting hitting), Right: example of a nonrepresentational gesture (placing emphasis while referring to a character)

involved significant movement of the elbow or lower arm received a score of 3, and gestures in which the upper arm was also used in a meaningful way, or that involved movement of the shoulder received a score of 4. This way, an average gesture size was computed for each participant.

Analysis. Analyses were done using a 2 × 2 ANOVA, with factors *speaker is seen by addressee* (levels: yes, no) and *speaker sees addressee* (levels: yes, no). The significance threshold was .05 and we report partial eta squared as a measure of effect size. As dependent variables for participants' gestures, we use the number of representational gestures produced per minute (the gesture rate) and the average size of representational gestures. We use the mean gesture rate rather than the mean total number of gestures produced, to control for any differences in the duration of the narrations between participants. Two participants were excluded from the analysis, because they deviated more than 2 standard deviations from the mean gesture rate in their condition. As a result, there were 9 participants in each condition. Inclusion of these two participants did not affect the significant effects found, but did reduce the significance of the overall model.

Although our focus is on participants' hand gestures, we also report some general measures of participants' speech. This addresses the question of whether different behavior in gesturing follows from the verbal narrations being much different across settings, rather than it being a direct result of the communicative setting. As global measures of the content of the narrations, we report the total number of words produced and the ratio of the number of unique words divided by the total number of words (*type-token ratio*). In addition, we report the number of words per second (*speech rate*) and the number of filled pauses per 100 words, as measures of fluency. To exclude a possible confounded effect of speech rate, we used the speech rate as a covariate in all our ANOVAs on gesture data in this and the following studies. Throughout the entire paper we report all means before this correction.

Results

Gesture Rate. Analyses of the number of representational gestures per minute, shown in Figure 3, revealed a main effect of the speaker being seen by the addressee, such that speakers gestured more frequently when they were seen ($M = 5.66$, $SD = 5.82$) than when they were not ($M = 2.58$, $SD = 2.03$), $F(1, 32) = 4.25$, $p < .05$, $\eta^2_p = .13$. The main effect of the speaker seeing the addressee showed a trend toward significance, such that speakers gestured less frequently when they saw their addressee ($M = 2.75$, $SD = 3.35$) than when they did not ($M = 5.49$, $SD = 5.27$), $F(1, 32) = 3.74$, $p = .06$, $\eta^2_p = .11$. The two factors did not interact, $F(1, 32) < 1$.

Gesture Size. Analyses of the average size of representational gestures, shown in Figure 4, revealed a trend toward significance for the main effect of the speaker being seen by the addressee, such that speakers' gestures were larger when speakers were seen ($M = 3.03$, $SD = .73$) than when they were not ($M = 2.53$, $SD = .78$), $F(1, 29) = 2.93$, $p = .10$, $\eta^2_p = .09$. When including nonrepresentational gestures, being seen exerted a main effect on gesture size, $F(1,31) = 4.59$, $p < .05$, $\eta^2_p = .13$. The speaker seeing the addressee did not exert a main effect on the gesture size, $F < 1$, and the two factors did not interact, $F < 1$.

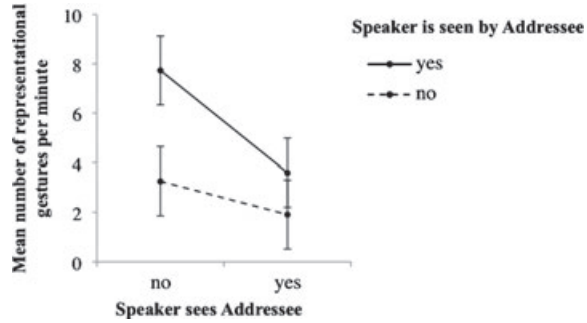

Figure 3 Study 1: Web Cam. Mean gesture rate depending on whether the speaker could be seen by the addressee (separate lines) and whether the speaker could see the addressee (x-axis). Bars represent standard errors

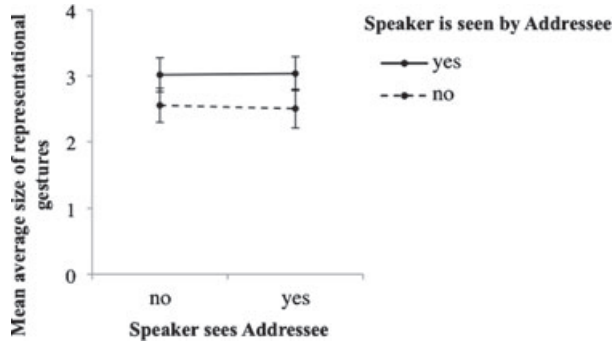

Figure 4 Study 1: Web Cam. Mean average size of representational gestures, depending on whether the speaker could be seen by the addressee (separate lines) and whether the speaker could see the addressee (x-axis). Bars represent standard errors

Speech. Neither the speaker being seen by the addressee, nor the speaker seeing the addressee exerted a main effect on the total number of words used, the type-token ratio, or the number of filled pauses per 100 words. The speaker being seen by the addressee exerted a main effect on the number of words per second, such that speakers spoke faster when they were seen ($M = 2.99$, $SD = .24$) than when they were not ($M = 2.73$, $SD = .45$) , $F(1, 32) = 4.61$, $p < .05$, $\eta^2_p = .13$. The speaker seeing the addressee did not exert a main effect on the speech rate, $F < 1$. The two factors did not interact, $F(1, 32) = 1.05$, $p = .31$.

Perceived Interest. Analyses of the extent to which speakers perceived the addressee as disinterested, shown in Figure 5, revealed an interaction between the factors *being seen by the addressee* and *seeing the addressee*, $F(1, 31) = 11.87$, $p < .01$, $\eta^2_p = .28$: Speakers agreed to the statement that the addressee was disinterested more when they could see the addressee, but the addressee could not see them. The main effect of the speaker seeing the addressee showed a trend toward significance, such that speakers agreed to the statement more when speakers could see the addressee ($M = 3.53$, $SD = 1.42$) than when they could not ($M = 2.89$, $SD = 1.28$), $F(1, 31) = 3.09$, $p = .09$, $\eta^2_p = .09$.

Other. Neither the speaker being seen by the addressee, nor the speaker seeing the addressee exerted a main effect on the duration of the narration. There was no significant correlation between the speech rate and the gesture rate, $p = .13$.

Discussion

Representational gestures were produced more frequently when speakers knew their addressee could see them. This was true both when speakers saw the addressee and when not, but especially the difference when speakers could not see their addressee is striking. This clearly shows that it was speakers' belief of being seen by the addressee that increased gesture production. In addition, gestures tended to be larger when speakers knew their addressee could see them, independent of whether speakers saw their addressee or not. These results clearly show that speakers applied their knowledge of the addressee's visual perspective to their gesture production, rather than solely using their own visual perspective.

Other than in previous work (Alibali, Kita, & Young, 2000; Bavelas, et al., 2008; Cohen, 1977; Jacobs & Garnham, 2007; Özyürek, 2002), our results cannot be explained by observable changes in the environment of the speaker. Our study therefore supports the interpretations of these earlier studies

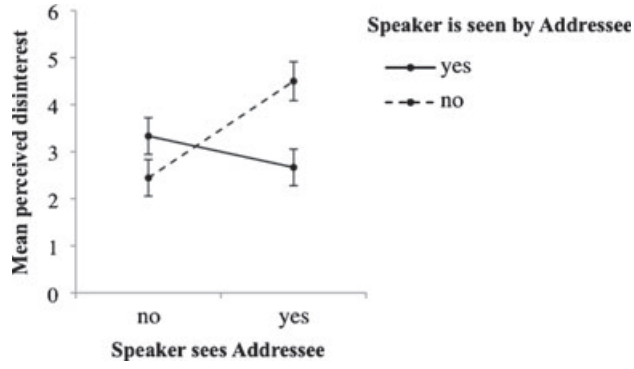

Figure 5 Study 1: Web Cam. Means of speakers' answer to the statement "The addressee was disinterested" on a 7-point scale, 1 = completely disagree, 7 = strongly agree. Bars represent standard errors

in terms of audience design. Speakers are able to adapt their gesturing to their (knowledge of their) addressee.

Participants' verbal narrations in each setting were similar, as can be seen from their length in words, the variation in vocabulary, their duration in time, and the frequency of filled pauses. Interestingly, participants spoke faster when they could be seen. It therefore seems that speakers also applied their knowledge of the addressee's perspective to their speech production. Perhaps speakers have an intuition that their speech can be understood more easily when they can be seen. It has been shown that visual information can indeed aid speech interpretation (e.g. Sumby & Pollack, 1954).

When speakers could see their addressee, they tended to produce fewer gestures than when they could not. At first sight it may seem surprising that speakers did not gesture more frequently when they saw their addressee, but the video-image of the addressee may have been confusing. The confederate addressee always gazed somewhere near the web cam capturing her, regardless of whether she could see the speaker or not. This somewhat unnatural gazing behavior may have been interpreted as the addressee being less interested. The results of our offline measure, a questionnaire, support this. After the narration task, the addressee was rated as less interested when the speaker could see the addressee but the addressee could not see the speaker. Speakers are known to gesture less when the addressee appears uninterested or inattentive (Jacobs & Garnham, 2007). Interestingly, this did not affect gesture size. Once a gesture was produced, it was produced larger when it was visible to the addressee, independent of how interested the addressee appeared to be.

Because of the effects of *being seen by the addressee* and *seeing the addressee* acting in opposite directions, the gesture rate in the setting in which speaker and addressee could see each other is surprisingly low compared to the setting in which only the speaker could be seen, and perhaps not quite representative of face-to-face interaction. Our second experiment addresses this issue.

Study 2: Eye-Catcher

One important difference remains between our setting with mutual visibility through web cams and a face-to-face setting: the usability of gaze. In study 2, we investigate the effect of the possibility of mutual gaze and the cost of interpreting gaze on language production. We do so by using a new technology for mediated communication: the Eye-Catcher, which enables interlocutors to look at each other's video-image *and* straight into the camera at the same time. This way, interlocutors can both look at each other simultaneously and appear to be looking at each other as well. This is not possible in communication through web cams, where giving the impression of looking at the other interlocutor requires looking into the camera, while seeing the other interlocutor requires looking at the screen.

Note that for the task we use, there is not much difference between using a web cam and using Eye-Catchers when using one-way visibility. In this case, either the addressee has nothing of interest to look at and thus it is of little use to the speaker to be able to interpret the addressee's gaze, or the speaker does not see the addressee and thus cannot make use of the addressee's gaze. Therefore, we use the Eye-Catcher in a setting with mutual visibility only. In such a setting, the addressee's gaze can inform the speaker of what the addressee is looking at. By comparing the Eye-Catcher setting to the Web Cam setting with mutual visibility, we can see how the availability of information from gaze affects language production.

If the availability of mutual gaze affects language production, we can re-examine the effects of being seen by the addressee and seeing the addressee, replacing the data from the Web Cam setting with mutual visibility with the data of the Eye-Catcher setting. This may provide a closer match with unmediated communication. Also, we can compare the effect of mutual visibility in mediated and unmediated communication, to see if the results obtained with mediated communication are likely

to generalize to unmediated communication. To draw this comparison, we make use of the data of an earlier study (Mol, et al., 2009), in which we used the same paradigm of retelling a cartoon in two unmediated settings. In the Face-to-Face setting, speaker and addressee were seated in the same room facing each other, such that visibility was unimpaired. In the Screen setting, interlocutors sat in the same room but on either side of an opaque screen, such that they could not see each other.

Method

The method for our second study was the same as for our first study, except that this time we used Eye-Catcher technology instead of communication through web cams. The Eye Catcher consists of a one-way mirror, in which a screen is reflected. Behind the one-way mirror is a camera. This way, the person in front of the Eye-Catcher can be captured while watching the reflected screen. We used two connected Eye-Catchers, such that the image captured by one Eye-Catcher's camera was shown on the screen of the other Eye-Catcher. This way, it appears as though interlocutors can look each other in the eyes, see Figure 6.

Figure 6 The Eye-Catcher seen from the front, with a video-image of the addressee

Through the Eye-Catchers, the narrator and the (confederate) addressee were able to see and hear each other and there was a possibility for mutual gaze. They were each seated in front of a table with an Eye-Catcher on it, at the same angle and distance, such that the setting was maximally symmetrical, enhancing the interpretability of gaze. The confederate's gazing was natural and her other back-channeling behavior was restricted in the same way as before.

Participants. Nine (six female) native Dutch speakers, all students from Tilburg University, participated in this study as part of their first year curriculum. They had a mean age of 21, range 19–26. In the earlier study we used for comparative analyses, 19 native Dutch students from Tilburg University participated. In the Face-to-Face setting, there were 10 (8 female) participants, with a mean age of 19, range 17–21. In the Screen setting, there were 9 (7 female) participants with a mean age of 18, range 18–19.

Transcribing, Coding & Analysis. We transcribed participants' speech and coded their hand gestures in the same way as before. First, we compare the Eye-Catcher setting to the Web Cam setting with mutual visibility in an independent samples *t*-test. This way, we can see if and how the Eye-Catcher's affordance of mutual gaze affects gesture and speech production. Second, we repeat our ANOVA with factors *speaker is seen by addressee* (levels: yes, no) and *speaker sees addressee* (levels: yes, no), with the data of the Eye-Catcher setting replacing the data of our previous setting with mutual visibility through web cams. The one-way visibility settings were not replaced in this analysis. As explained earlier, there is no relevant difference between using Eye-Catchers and using a web cam in one-way visibility settings. We also do an ANOVA with factors mutual visibility (levels: yes, no) and mediation (levels: yes, no), comparing the mediated Audio Only setting of experiment 1 and the Eye-Catcher setting, to the unmediated Face-to-Face setting and the Screen setting of our earlier study (Mol, et al., 2009).

Results

Comparing the Web Cam and Eye-Catcher Settings

Gesture. Representational gestures were produced more frequently in the Eye-Catcher setting ($M = 14.25$, $SD = 8.08$) than in the Web Cam setting ($M = 3.60$, $SD = 4.26$), $t(16) = 3.50$, $p < .01$. The gesture size was comparable in both conditions, $t(16) = .26$, $p = .80$.

Speech. The total number of words produced, the duration of the narration, the type-token ratio, and the number of filled pauses per 100 words did not differ significantly across the two settings. The speech rate was also comparable in the Web Cam ($M = 2.92$, $SD = .29$) and Eye-Catcher setting ($M = 3.01$, $SD = .45$), $t(16) = -.46$, $p = .74$.

Analysis with Factors: Speaker is Seen by Addressee, Speaker Sees Addressee

Since the gesture rate was much different in the Web Cam and Eye-Catcher setting, we repeat our previous analysis of experiment 1, with the data of the Eye-Catcher setting replacing the data of the Web Cam setting with mutual visibility.

Gesture Rate. Analyses of the number of representational gestures per minute, shown in Figure 7, revealed a main effect of the speaker being seen by the addressee, such that speakers gestured more frequently when they were seen ($M = 11.0$, $SD = 7.92$) than when not ($M = 2.58$, $SD = 2.03$), $F(1, 31) = 15.34$, $p < .001$, $\eta^2_p = .33$. The speaker seeing the addressee did not exert a main effect on the gesture rate, $F(1, 31) = 1.41$, $p = .24$. The two factors interacted, $F(1, 31) = 5.34$, $p < .05$,

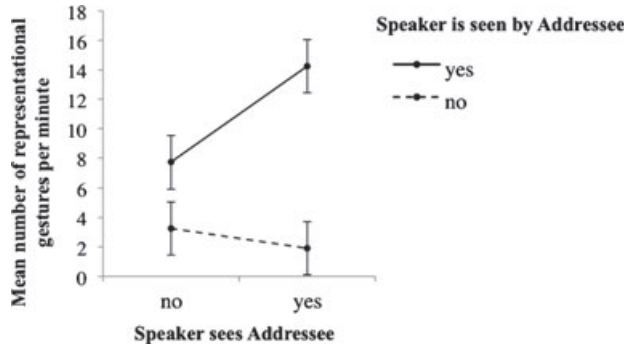

Figure 7 Study 2: Eye-Catcher. Mean gesture rate depending on whether the speaker could be seen by the addressee (separate lines) and whether the speaker could see the addressee (x-axis). Bars represent standard errors

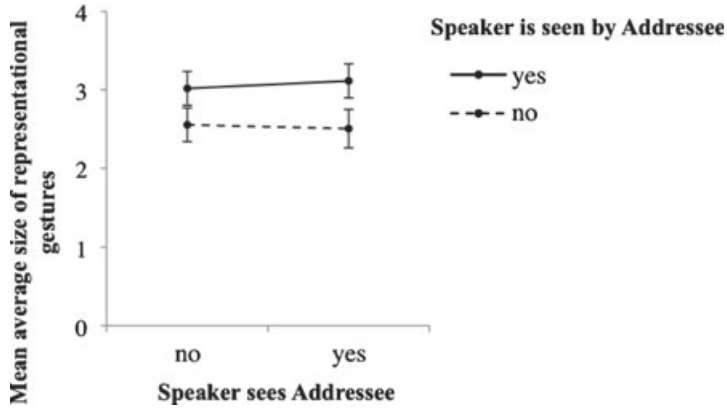

Figure 8 Study 2: Eye-Catcher. Mean average size of representational gestures, depending on whether the speaker could be seen by the addressee (separate lines) and whether the speaker could see the addressee (x-axis). Bars represent standard errors

$\eta^2_p = .15$: Seeing the addressee only increased gesture production if the addressee could also see the speaker.

Gesture Size. Analyses of the average size of representational gestures, shown in Figure 8, revealed a main effect of the speaker being seen by the addressee, such that speakers produced larger gestures when they were seen ($M = 3.07$, $SD = .46$) than when they were not ($M = 2.53$, $SD = .78$), $F(1, 29) = 4.60$, $p < .05$, $\eta^2_p = .14$. The speaker seeing the addressee did not exert a main effect on gesture size, $F < 1$. The two factors did not interact, $F < 1$.

Speech. Neither the speaker being seen by the addressee, nor the speaker seeing the addressee exerted a main effect on the total number of words used, the number of filled pauses per 100 words, or the type-token ratio. The speaker being seen by the addressee exerted a main effect on the speech rate, such that speakers spoke faster when they were seen ($M = 3.03$, $SD = .32$) than when they were not

($M = 2.73$ $SD = .45$), $F(1, 32) = 5.07$, $p < .05$, $\eta^2_p = .14$. The speaker seeing the addressee did not exert a main effect on the speech rate, $F < 1$. The two factors did not interact, $F < 1$.

Other. Neither the speaker being seen by the addressee, nor the speaker seeing the addressee exerted a main effect on the duration of the narrations in seconds. There was a significant correlation between participants' speech rate and their gesture rate, $r = .34$, $p < .05$.

Analysis with Factors: Mutual Visibility and Mediation

In this analysis we assess the effects of the speaker and addressee being able to see each other and communication being mediated on language production.

Gesture Rate. Analyses of the number of representational gestures per minute, shown in Figure 9, revealed a main effect of mutual visibility, such that speakers gestured more frequently when interlocutors could see each other ($M = 11.89$, $SD = 7.28$) than when they could not ($M = 3.92$, $SD = 2.30$), $F(1,32) = 19.75$, $p < .001$, $\eta^2_p = .38$. The main effect of mediation showed a trend toward significance, such that speakers gestured more frequently when communication was mediated ($M = 8.75$, $SD = 8.02$) than when it was not ($M = 7.31$, $SD = 5.36$), $F(1, 32) = 3.97$, $p = .06$, $\eta^2_p = .11$. The latter effect was not present without speech rate as a covariate ($F < 1$). The two factors did not interact, $F(1, 32) = 1.76$, $p = .19$.

Gesture Size. Analyses of the average size of representational gestures, shown in Figure 10, revealed a main effect of mutual visibility, such that speakers produced larger gestures when interlocutors could see each other ($M = 3.04$, $SD = .41$) than when they could not ($M = 2.62$, $SD = .52$), $F(1, 31) = 7.64$, $p < .01$, $\eta^2_p = .20$. Mediation did not exert a main effect on gesture size, $F < 1$, and the two factors did not interact, $F < 1$.

Speech. Neither mutual visibility nor mediation exerted a main effect on the total number of words used, or the type-token ratio. Mutual visibility exerted a main effect on the number of filled pauses per 100 words, such that speakers used filled pauses less frequently when interlocutors could see each other ($M = 5.70$, $SD = 3.02$) than when they could not ($M = 7.46$, $SD = 1.93$), $F(1, 33) = 4.18$, $p < .05$,

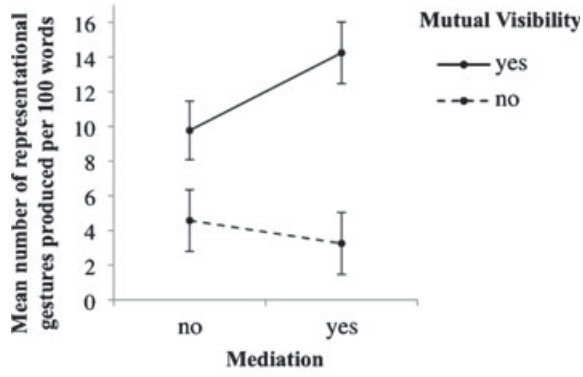

Figure 9 Study 2: Comparison With Unmediated Settings. Mean gesture rate depending on whether communication was mediated (x-axis) and whether speaker and addressee could see each other (separate lines). Bars represent standard errors

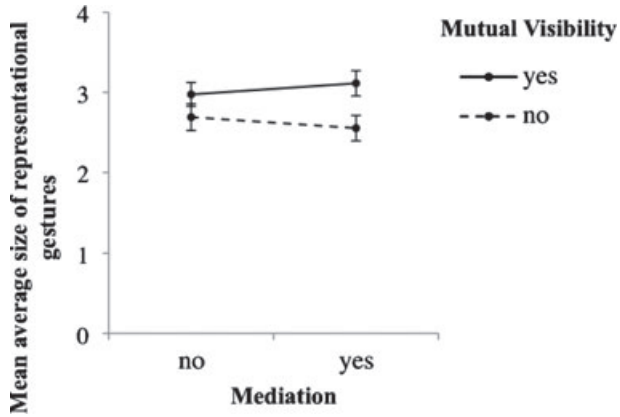

Figure 10 Study 2: Comparison With Unmediated Settings. Mean average size of representational gestures, depending on whether communication was mediated (x-axis) and whether speaker and addressee could see each other (separate lines). Bars represent standard errors

$\eta_p^2 = .11$. Mediation did not exert a main effect on the rate of filled pauses, $F < 1$, and the two factors did not interact, $F < 1$. Mediation exerted a main effect on the number of words per second, such that speakers spoke slower when communication was mediated ($M = 2.84$, $SD = .50$) than when it was not ($M = 3.27$, $SD = .51$), $F(1, 33) = 6.64$, $p < .02$, $\eta^2{}_p = .17$. Mutual visibility did not exert a main effect on the speech rate, $F(1, 33) = 1.14$, $p = .29$ and the two factors did not interact, $F < 1$.

Other. Neither mutual visibility nor mediation exerted a main effect on the duration of the narrations in seconds. There was a significant correlation between participants' speech rate and their gesture rate, $r = .38$, $p < .05$.

Discussion

Comparison of the Eye-Catcher setting to the setting with mutual visibility through web cams revealed that the extra affordance offered by the Eye-Catcher affected gesture production. Gestures were produced far more frequently when information from gaze could readily be interpreted and speakers could look at their addressee's eyes and see their addressee look at their eyes simultaneously. Therefore, it seems that the decrease in gesture production that we found in our first study when speakers could see their addressee, indeed resulted from the somewhat unnatural gazing behavior of the addressee, which may have caused her to appear inattentive or uninterested. Being able to use gaze may also have affected gesture production more directly, because speakers can use gaze to direct their addressee's attention to their gestures (Gullberg & Kita, 2009).

Provided that interlocutors can see each other (mutual visibility), the Eye-Catcher seems to allow for natural gazing behavior. Therefore, the Eye-Catcher setting is more suitable for addressing how being seen by the addressee and seeing the addressee affect language production. With the Eye-Catcher setting replacing our previous setting with mutual visibility through web cams, the effects of the speaker being visible to the addressee and vice versa become easier to interpret. What we see is that whether the speaker can be seen influences both the frequency and the size of representational gestures, showing that speakers adapt their gesture production to their knowledge of whether the addressee can see them. Seeing the addressee also causes speakers to put more communicative effort into their gesture

production, but only if the addressee can see them, and only if the addressee's gaze can readily be interpreted.

Comparison of our mediated settings with mutual or no visibility between interlocutors to similar unmediated settings showed that mutual visibility affected language production similarly, independent of whether communication was mediated or not. This suggests that the Eye-Catcher setting is a close match to face-to-face communication, and that our finding that speakers do take into account their addressee's visual perspective are likely to generalize to unmediated communication. We also found that speakers did not gesture less when communication was mediated, provided that the costs and affordances were a close match to unmediated communication. This supports the grounding framework by Brennan and Lockridge (2006) and is consistent with social presence theory (Short, et al., 1976) as well as social information processing theory (Walther, 1992). Our results do not reveal whether the effect of the lack of interpretability of gaze in web cam settings can be overcome with time, as may be predicted by social information processing theory.

Our global measures related to the content of the narrations showed that the narrations were very similar in all settings. Therefore, it does not seem that the differences in gesture production resulted from differences in speech content. In some of our analyses we again found that being seen by the addressee caused speakers to speak faster, as we found in the analysis of our first experiment. Additionally, one of our analyses showed that speakers produced fewer filled pauses when the addressee could see them. When visual information is unavailable, it may be more necessary to use filled pauses communicatively, indicating that one is still thinking (Clark & Fox Tree, 2002). We also found that speech was faster in unmediated settings, compared to mediated ones. This may indicate a limited trust in the signal quality in mediated communication. Although we found some correlation between participants' speech rate and their gesture rate, it does not seem that the increased gesture production in some settings resulted from a need to speak faster, because also in unmediated settings speech was faster, without there being an increased gesture production. We think it more likely that the differences in the communicative settings affected both speech and gesture, with expectedly, manipulations of visibility affecting gesture production more dramatically than speech production.

The differences we found in speakers' gesture production are informative of whether speakers apply their knowledge of the addressee's visual perspective to their language production. Yet do these differences in language production matter to the addressee? Our third study examines whether naïve observers are sensitive to some of the differences we found in speakers' language production. It thereby also examines in yet another way how similar mediated and unmediated communication were in our study.

Study 3: Perception Study

In this study, we ask participants to rate movie clips from speakers who communicate either through web cams, with the Eye-Catcher, or face-to-face, all with mutual visibility between speaker and addressee. As a measure of how well speakers are perceived to perform the narration task, we ask participants to rate the speakers for their expressivity. If whether communication is mediated influences speakers' behavior most, then we would expect speakers in the Face-to-Face setting to be rated differently from speakers in the two mediated settings. On the other hand, if not mediation but the interpretability of gaze affects language production most, we would expect speakers in the Face-to-Face and Eye-Catcher setting to be rated differently from speakers in the Web Cam setting. If both of these factors play a role, then speakers from each setting may be rated differently. Another possible outcome is that although differences can be found in a formal analysis of gesture, these differences do not matter for how speakers are perceived in terms of their expressivity.

Method

Participants. Twenty (17 female) native Dutch first year students from Tilburg University took part in this study. Their mean age was 21, range 18–25.

Stimuli. We created 27 trials, using 9 movie clips from speakers in each setting with mutual visibility: the Face-to-Face, Eye-Catcher, and Web Cam setting. In addition, we created two practice trials, using data from speakers in an unrelated experiment in which we used a similar cartoon narration task. From each speaker, we chose a fragment of 10 seconds, starting as soon as the speaker started to talk about the third episode of the cartoon. In this episode Sylvester tries to climb up to Tweety's window through an adjacent drainpipe, but gets stopped by a bowling ball, which was thrown into the drainpipe by Tweety. Speakers were visible from the knees up. Each movie clip was preceded by a short beep and an order number that corresponded to a line on the answering form, which was displayed for 2 seconds. After each clip, 4 seconds of blank video were inserted allowing participants time to fill out their answer. After the last movie clip a text was displayed, which indicated that the experiment had ended. The 27 actual clips were presented in a random order. We created two versions, the second one showing the clips in reversed order.

Procedure. Participants came to the lab and were asked to rate video-fragments of speakers for how expressive the speaker was. They indicated their answer for each speaker on an answering form, by circling a number on a 1 to 5 scale, '1' meaning 'very little expressive' and '5' meaning 'very expressive'. Participants first saw two practice trials. After the practice trials they were allowed to ask any clarification question on the task, which were answered by the experimenter (without her ever mentioning gesture). The participant then watched the actual fragments, filling out the answering form after each fragment. Half the participants saw the fragments in a certain order and the other half in reversed order. After this task participants filled out a brief questionnaire, which asked for participants' age, native language and what they had based their ratings on.

Analysis. We used a Repeated Measures analysis with the setting that the movie clips were taken from as a factor (levels: face-to-face, Eye-Catcher, web cam). For each setting we first computed each participant's mean rating of the nine movie clips from that setting, which we used as the dependent variable. Pairwise comparisons were done using the LSD method with a significance threshold of .05.

Results

Analyses of participants' ratings of speakers' expressivity revealed a main effect of the setting that the speaker was in, $F(2, 38) = 26.\ 10$, $p < .001$, $\eta^2_p = .58$. Posthoc analysis showed that speakers from the Web Cam setting were rated as less expressive ($M = 2.42$, $SD = .53$) than speakers from the Face-to-Face ($M = 3.07$, $SD = .78$) and Eye-Catcher setting ($M = 3.24$, $SD = .54$). The ratings for speakers from these latter two settings did not differ significantly. The order of presentation of the clips did not exert a main effect on participants' rating, $F < 1$.

In answer to an open question of what participants had based their judgment on, 14 out of 20 participants (70%) spontaneously mentioned that they had partially based their judgments on speakers' hand movements. Participants also mentioned that they had paid attention to the speakers' facial expressions (55%), posture (35%), body language (15%), gaze (10%), eye-brow movements (5%), intonation (40%), use of voice (30%), loudness (10%), clarity of voice (5%), and laughing (15%).

Discussion

Participants were sensitive to the differences in how speakers narrated between the setting with communication through web cams on the one hand, and face-to-face communication and communication

through Eye-Catchers on the other. Speakers from the Web Cam setting were rated as less expressive. Most participants took a speaker's hand gestures into account when judging the speaker's expressivity. It therefore seems that producing hand gestures more frequently (as speakers in the Eye-Catcher and face-to-face settings did), is associated with greater expressivity. In addition, there was no perceived difference in expressivity between speakers from the Face-to-Face and from the Eye-Catcher setting, again suggesting that these settings were a closer match than face-to-face interaction and communicating through web cams. Thus, whether communication is mediated or not seems of lesser influence on speakers' language production than whether gaze in mediated settings resembles gaze in unmediated settings.

General Discussion and Conclusion

When it comes to their gesturing, speakers apply their knowledge of their addressee's visual perspective to their language production. Speakers produced more and larger gestures when they knew their addressee could see them. This suggests that gesturing is at least partly intended communicatively. Other than in previous work, our results cannot be explained by observable changes in the environment of the speaker. Our study therefore supports the interpretations in terms of audience design of earlier studies (Alibali, et al., 2000; Bavelas, et al., 2008; Cohen, 1977; Jacobs & Garnham, 2007; Özyürek, 2002). Speakers are able to adapt their gesturing to their knowledge of their addressee, rather than solely using their own perspective.

This does not prove that speakers never make mistakes when they need to take into account the other interlocutors' visual perspective, as has been found for speech production and interpretation (Keysar, et al., 2003; Wardow Lane, et al., 2006). In our studies, gesture production was used as a global measure of language production. This is different from the use of eye-tracking data as in the study by Keysar et al., which may be able to capture any mistake in language interpretation. It differs similarly from the use of reference production in the study by Wardow Lane et al., which reveals any overspecification. Gesture production thus shows a general trend, rather than capturing every individual mistake.

We found some evidence that gesture frequency is reduced when the addressee seems less interested, as has also been found by Jacobs and Garnham (2007). In addition, we found that gesture size does not seem to be affected by this factor. Thus, although speakers produce fewer gestures when the addressee appears less interested, the gestures they do produce may compare well to gestures directed at addressees appearing more interested. In reality, there was no difference in how interested the addressee was between our conditions, since the addressee always was a confederate. Rather, the addressee was perceived as less interested in one condition, as a result of the costs and affordances offered by the mediated setting. When speakers could see the addressee, but the addressee could not see the speaker, speakers perceived the addressee as less interested. In this case it was not possible for the addressee to show natural gazing behavior. This effect is predicted by the grounding framework (Brennan & Lockridge, 2006), which states that it is not mediation as such that causes interlocutors to perceive each other differently, but rather the fact that differences in the costs and affordances associated with mediation affect interlocutors' behavior, which in turn leads to different perceptions of each other.

The results we found are consistent with the idea that the more the costs and affordances of mediated and unmediated communication are alike, the more similar language use will be (Brennan & Lockridge, 2006). In our first study, in which we made use of communication through web cams, we saw that seeing the addressee led to a decrease in gesture production rather than an increase. In this case the affordance of seeing the addressee was more similar to face-to-face interaction, but the costs associated with interpreting the addressee's gaze were not. Interpreting the others' gaze is harder when communicating through web cams than when interacting face-to-face. Our second study showed that when this was

compensated for by using the Eye-Catcher, seeing the addressee did increase gesture production, such that the gesture rate and size in mediated communication were now comparable to those in face-to-face interaction. The gesture rate was much higher when speaker and addressee could see each other, independent of whether they were in the same room. Therefore, it indeed seems that mediation as such does not have a large effect on language production, as predicted by the grounding framework.

We further showed that observers were sensitive to the differences in speakers' communicative behavior. When participants were asked to judge speakers' expressivity, speakers from a setting with communication through web cams were rated as less expressive than speakers from a face-to-face setting, as well as speakers from a setting in which communication was mediated by Eye-Catchers. This shows that speakers are more expressive when interlocutors can readily interpret each other's gaze. It also confirms that communication through Eye-Catchers resembles face-to-face communication, more so than communication through web cams.

Since narrations were equally long in each setting we used, both in time and in the number of words, and the variation in vocabulary did not differ, it is likely that the communicative setting affected gesture production, rather than the differences in gesture resulting from the narrations being much different. Our studies suggest that speakers may also adjust their speech rate to whether or not their addressee can see them, similar to adjusting their gesture rate and size. In this case too, speakers' knowledge of the addressee seeing them was more important than whether or not they saw their addressee. Speakers spoke faster when they knew they could be seen, which may indicate they had an intuition that visual information aids speech interpretation. Such an intuition would be consistent with actual findings (e.g. Sumby & Pollack, 1954).

Despite dialogue being possible in the settings we used, the task we used in our studies was mostly a monologue task and the addressee never interrupted the speaker. This has the upside of settings being very similar to each speaker within a certain setting, such that we could get a clear picture of the factors of interest. It also strengthens our case that speakers can use their knowledge of their addressee's perspective, rather than their own direct observation (including the addressee's back-channeling behavior). In future work, however, it will be necessary to look at other factors that come into play when scaling up from monologue to dialogue. Does mediated communication with the Eye-Catcher still pass the test when more turn-taking is involved? And what if there are more than two speakers? Our study implies that it is important for interlocutors in such situations to be able to make sense of each other's gaze. Moreover, our studies suggest that when using video-conferencing, it may be important to choose the image such that the hands are visible, since speakers adapt their gesturing to their addressee and thus seem to intend it communicatively.

Acknowledgements

We thank the anonymous reviewers for their helpful comments on earlier versions of this paper. We also thank all our participants. We thank Nelianne van den Berg, Hanneke Schoormans, Vera Nijveld, and Madelène Munnik for their help in collecting, coding, and transcribing the data. We thank Bernd Hellema for providing a Perl script and Lennard van der Laar for his technical assistance.

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*, 169–188.

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes, 15*(6), 593–613.

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

Bangerter, A., & Chevalley, E. (2007). Pointing and describing in referential communication: When are pointing gestures used to communicate? In I. Van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds.), *Proceedings of the Workshop on Multimodal Output Generation* (pp. 17–28). Enschede: Universiteit Twente.

Bavelas, J., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes, 15*, 469–489.

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language, 58*, 495–520.

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

Brennan, S. E., & Lockridge, C. B. (2006). Computer-mediated communication: A cognitive science approach. In K. Brown (Ed.), *ELL2, Encyclopedia of Language and Linguistics, 2nd Edition* (pp. 775–780). Oxford, UK: Elsevier Ltd.

Brennan, S. E., & Ohaeri, J. O. (1999). Why do electronic conversations seem less polite? The costs and benefits of hedging. In D. Georgakopoulos, W. Prinz & A. L. Wolf (Eds.), *Proceedings of the International Joint Conference on Work Activities, Coordination, and Collaboration (WACC '99)* (pp. 227–235). San Francisco, CA: ACM.

Brennan, S. E., Xin, C., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition, 106*, 1465–1477.

Cassell, J., McNeill, D., & McCullough, K.-E. (1998). Speech-gesture mismatches: Evidence for one underlying representation of linguistic & nonlinguistic information. *Pragmatics & Cognition, 6*(2), 1–33.

Chawla, P., & Krauss, R. M. (1994). Gesture and speech in spontaneous and rehearsed narratives. *Journal of Experimental Social Psychology, 30*, 580–601.

Chu, M., & Kita, S. (2008). Spontaneous gestures during mental rotation tasks: Insights into the microdevelopment of the motor strategy. *Journal of Experimental Psychology: General, 137*(4), 706–723.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. Levine & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: APA.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition, 84*(1), 73–111.

Cohen, A. A. (1977). The communicative functions of hand illustrators. *Journal of Communication, 27*(4), 54–63.

Cohen, A. A., & Harison, R. P. (1973). Intentionality in the use of hand illustrators in face-to-face communication situations. *Journal of Personality and Social Psychology, 28*, 276–279.

Cutica, I., & Bucciarelli, M. (2008). The deep versus the shallow: Effects of co-speech gestures in learning from discourse. *Cognitive Science, 32*(5), 921–935.

Effron, D. (1941). *Gesture and environment.* Morningside Heights, NY: King's Crown Presss.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*, 49–98.

Feyereisen, P., Van de Wiele, M., & Dubois, F. (1988). The meaning of gestures: What can be understood without speech? *Cahiers de Psychologie Cognitive, 8*, 3–25.

Goldin-Meadow, S. (2010). When gesture does and does not promote learning. *Language and Cognition, 2*(1), 1–19.
Goldin-Meadow, S., & Sandhofer, C. M. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science, 2*(1), 67–74.
GreenEyes. (2007). http://www.greeniii.com/index.php
Grice, P. (1989). *Studies in the way of words.* Cambridge MA: Harvard University Press.
Gullberg, M., & Kita, S. (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behavior, 33*(4), 251–277.
Hadar, U. (1989). Two types of gesture and their role in speech production. *Journal of Language and Social Psychology, 8*(3–4), 221–228.
Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language, 57*(4), 596–515.
Isaacs, E. A., & Tang, J. C. (1994). What video can and cannot do for collaboration: A case study. *Multimedia Systems, 2*(2), 63–73.
Jacobs, N., & Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language, 56*(2), 291–303.
Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22–63.
Kendon, A. (1988). How gestures can become like words. In F. Potyatos (Ed.), *Crosscultural perspectives in nonverbal communication* (pp. 131–141). Toronto, Canada: Hogrefe.
Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge: Cambridge University Press.
Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*(1), 25–41.
Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57*(3), 396–414.
Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science, 7*, 54–60.
McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* Chicago and London: The University of Chicago Press.
Melinger, A., & Kita, S. (2007). Conceptualisation load triggers gesture production. *Language and Cognitive Processes, 22*(4), 473–500.
Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2009). The communicative import of gestures: Evidence from a comparative analysis of human-human and human-computer interactions. *Gesture, 9*(1), 97–126.
Özyürek, A. (2002). Do speakers design their cospeech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language, 46*(4), 688–704.
Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications.* London: Wiley.
Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.
Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research, 19*(1), 52–90.
Walther, J. B., Slovacek, C. L., & Tidwell, L. C. (2001). Is a picture worth a thousand words? Photographic images in long-term and short-term computer-mediated communication. *Communication Research, 28*(1), 105–134.
Wardow Lane, L., Groisman, M., & Ferreira, V. S. (2006). Don't talk about pink elephants: Speakers' control over leaking private information during language production. *Psychological Science, 17*(4), 273–277.

About the Authors

Lisette Mol (L.Mol@uvt.nl) is an Assistant Professor in the department of Communication and Information Sciences at Tilburg University, The Netherlands. Her research focuses on the cognitive underpinnings of gesture production.

Emiel Krahmer (E.J.Krahmer@uvt.nl) is a Full Professor in the department of Communication and Information Sciences at Tilburg University, The Netherlands. In his research he combines computational models and experimental studies to gain a better understanding of human speech production, which in turn may help improving the way computers present information to and communicate with human users.

Alfons Maes (Maes@uvt.nl) is head of department and Full Professor in the Department of Communication and Information Sciences at Tilburg University, The Netherlands. His research topics include the way people adapt their referential behavior to specific communicative conditions as well as the design of multimodal documents for use in different communicative domains, such as advertising, health communication, and instruction.

Marc Swerts (M.G.J.Swerts@uvt.nl) is a Full Professor in the department of Communication and Information Sciences at Tilburg University, The Netherlands. He is co-editor-in-chief of the journal *Speech Communication*. His research focuses on the functional analysis of audiovisual prosody.

Address: Tilburg center for Cognition and Communication (TiCC), Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands.