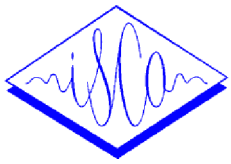
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
2007 (AVSP2007)
Hilvarenbeek, The Netherlands
August 31 - September 3, 2007

The communicative import of gestures: Evidence from a comparative analysis of human-human and human-machine interactions

Lisette Mol, Emiel Krahmer, Alfons Maes, Marc Swerts

Faculty of Humanities, Communication and Cognition, Tilburg University
Warandelaan 2, 5037 AB, Tilburg, The Netherlands
l.mol@uvt.nl

Abstract

Is communication the primary functional role of gesticulation? We conducted a study in which participants narrated to a presumed computer system, a presumed addressee in another room (via web cam), or an addressee in the same room, who could either see them or not. Participants produced significantly fewer gestures when they thought to be talking to a computer system. Our results show that people narrate differently to a computer system than to a human addressee. Considering the difference in gesticulation, it seems plausible that, in a narrative task, most gestures are produced with a communicative intent.
Index Terms: gesture, human-machine-interaction, narration

1. Introduction

1.1. Gesture types

A Dutch association for traffic safety has proposed a change in gesturing for drivers. Instead of giving the finger to aggressive drivers, people should simply raise their hand. Their hypothesis is that this will reduce aggression in driving and thus lead to safer traffic. Clearly, some gestures are thought to have an effect on their addressee. But is this effect the main reason why people gesture? Or do speakers gesture mostly for themselves? And would people also raise their finger to an aggressive artificial opponent in a racing video game?

This traffic example nicely illustrates that gestures differ. In fact, various types of hand gestures can be discriminated. Giving the finger is an *emblem*. Emblems are gestures with a fixed meaning, that can be culturally specific. They can be used to replace words and therefore they are closer to sign language than other gestures that accompany speech. Kendon [10] has proposed a continuum of gestures with sign language on one end, and gesticulation (idiosyncratic spontaneous movements of the hands and arms during speech) on the other. In between are emblems, pantomime, and language-like gestures. Moving from sign languages to gesticulation, the presence of speech becomes more obligatory, gestures get less linguistic properties (such as compositionality and grammar), and socially regulated signs are replaced by idiosyncratic gestures.

The present study is on narration. McNeill [15] defines four types of gestures in the context of narrative speech. All of these are in Kendons category of gesticulation. *Iconic gestures* depict some of the content of a scene. This content need not be redundant with the content of the accompanying speech, although the two are closely related. An example is a hand that is moved in an arc to depict the path of a motion that a speaker is talking about. *Metaphoric gestures* are similar to iconic gestures, but present an image of an abstract concept, such as knowledge, language itself, the genre of a narrative. An example presented by McNeill is that of a cupped hand representing a question. Iconic and metaphoric gestures are together called *imagistic gestures*.

Deictic gestures are defined as all pointing movements. They focus attention to a referent (e.g. Bangerter and Chevalley [3]). They need not point to concrete entities. Iconic, metaphoric, and deictic gestures together form the category of *representational gestures*. In the present study, the focus is on representational gestures.

The fourth gesture type described by McNeill are beat gestures. *Beats* are biphasic, small, low energy, rapid movements of the hands or fingers that emphasize the part of speech they accompany. There is no relationship between the form of the gesture and the semantic content of the accompanying speech. (Also see [12].)

Bavelas et al. [4] describe an additional type of gestures. *Interactive gestures* are defined as gestures that are designed to signal conversational information to the conversational partner. They suggest that these gestures include some representational gestures and subsume the category of beats.

1.2. The functional role of gestures

Many studies have been conducted to investigate the primary functional role of hand gestures. Initially, there were two main hypotheses. The first is that gestures facilitate speech production. More specifically, the Lexical Retrieval Hypothesis states that producing gestures facilitates the retrieval of phonological word forms from the mental lexicon while speaking. Hadar [8] and Krauss [13], among others, have found support for this view.

The second hypothesis is that gestures primarily serve a communicative purpose and are produced to aid comprehension by the addressee [10]. Support for this hypothesis has been found, among others, by Özyürek [18], Jacobs and Garnham [9], and Bangerter and Chevalley [3].

Alibali et al. [2] tried to reconcile various seemingly contradictive experimental results by associating different types of gestures to different functional roles. They conducted a study in which narrators told a story to an addressee either face-to-face, or with a wooden screen in between speaker and addressee. They found that speakers produced more representational gestures in the face-to-face condition, in which gestures had communicative potential. But narrators still produced this type of gestures if the addressee could not see them. Beat gestures were produced at comparable rates under both conditions. Alibali et al. therefore concluded that both types of gestures serve

both speaker-internal and communicative functions. They suggest examining 'how different speakers use gestures in different types of contexts for both speaker-internal and communicative purposes' instead of trying to find a single primary role of gesture production.

1.3. Factors influencing gesticulation

More recently, the focus has shifted to investigating the factors influencing the different gestures types, and the implications for the functional role(s) of gesticulation. The study by Alibali et al. investigated the influence of mutual visibility on representational gestures and beats. Bangerter and Chevalley [3] investigated the effect of mutual visibility on pointing gestures, in a referential communication task. They found that pointing movements that do not involve raising the arm, were produced at equal rates whether conversational partners could see each other or not, suggesting that they are automatic in production. However, pointing movements that did involve raising the arm were used more when interlocutors could see each other, suggesting that they are intended to communicate. Thus, the size of gestures seems to be relevant to their functional role. This was also found by Bavelas et al. in [5]. In addition, they found that speakers gesticulated more and differently when being in a dialogue, which also points to the communicative role of gesticulation.

Besides mutual visibility, Jacobs and Garnham [9] point out that gesture production may depend on the behavior and needs of the addressee, and on the type of task that the speaker is performing. They found that narrators produced fewer gestures when (they knew that) their addressee already knew (part of) the content of the story they were telling. They also found that speakers produced more gestures when the addressee appeared attentive, than when the addressee appeared inattentive. They therefore concluded that *during narrative tasks*, gestures are produced primarily for the benefit of the addressee.

Melinger and Levelt [16] looked at the type of information being represented. They found that speakers who used iconic gestures representing spatial information, omitted more necessary spatial information from their verbal descriptions than speakers that did not gesture. They showed that some speakers divided information between the gesture and speech modality. This shows that iconic co-speech gestures expressing spatial information can be used communicatively.

De Ruiter [6] introduced the Mutually Adaptive Modalities hypothesis. This hypothesis states that speakers are more likely to use the modality that is most suitable to convey the type of information to be expressed, and is most effective given the environmental conditions. Thus, speakers may gesture more in a noisy environment and may rather use gestures to express spatial information than to express information about color.

1.4. Purpose of this study

In this study, we use a new paradigm to investigate to what extent gesture production is automatic and serves speaker internal purposes, and to what extent it is meant communicatively. If the primary role of gesticulation were to facilitate speech, it should be relatively independent of the addressee, and even of the human nature of the addressee. So if the addressee were artificial, for example a speech recognition system, speakers are still expected to gesture a lot. However, if the primary role is communicative, it seems a reasonable hypothesis that people will gesture far less when interacting with a computer.

When considering various communicative conditions, computer mediation is an important factor. Reeves and Nass [19] state that 'people's responses to media are fundamentally social and natural'. This is the so called *media equation* and it applies to everyone. Reeves and Nass state that the confusion of mediated life and real life is not rare and inconsequential, and that it cannot be corrected with age, education, or thought. This suggests that even if gestures are used to communicate, people would also gesture at computers and other media.

Although people show social responses to media and artificial agents, the question remains whether they do so to the same extent as to human interlocutors. Aharoni and Fridlund, [1] conducted a study in which participants smiled more and used more silence fillers to a purported human interviewer than to a computer interviewer. In both cases a prerecorded stimulus was used. They found that labeling the stimulus as 'human' caused people to be more communicative. Maes et al. [14] showed that the assumption of a human addressee can induce more referential effort than the assumption of a computer addressee.

The main goal of the present study is to investigate the effect of the addressee being artificial or human on gesture production. In addition, we are interested in the effects of two other factors: speaker visibility and physical co-presence. We think that physical co-presence may also influence gesticulation, and that this may not be due to visibility alone. It has been found that communication via high quality video phone is perceived to be more similar to communication via phone, than to face-to-face communication, and that a video phone is used in the same way, and for the same tasks as a regular phone [7]. Whittaker [20] explains this by suggesting that the physical environment being visible may be more important than the addressee being visible. But in addition, it could be that simply being in the same room adds something to the communication.

2. Method

2.1. Design

We have used a between subjects design with four conditions (see Figure 1). In condition 1 narrators had to retell the story of an animated cartoon to a purported computer system. In condition 2 they told the story to a presumed other person via camera. The only difference between these conditions was whether the addressee was human or artificial.

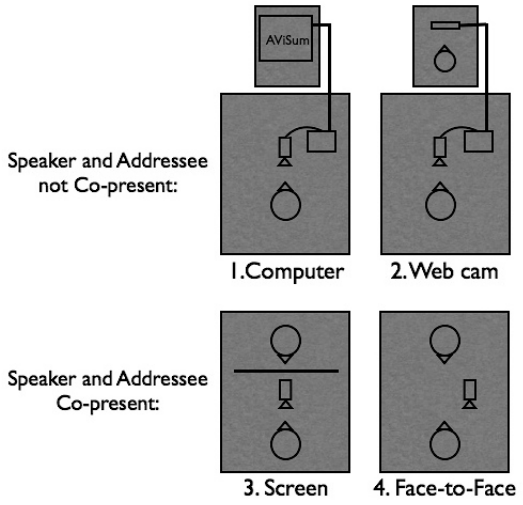
In condition 3 participants told the story to a participant who was in the same room, but behind an opaque screen. So other than in condition 2, speakers did not expect to be seen by the addressee. Condition 4 was face-to-face communication. This was the only condition in which the participant could see the addressee.

Mutual visibility is the only difference between condition 3 and 4. In condition 1 and 2, speaker and addressee are not physically co-present, that is to say, they are not in the same room. In condition 3 and 4 they are.

2.2. Participants

43 participants volunteered as narrators for this study. We excluded two participants, because they were suspicious about the experimental setup (see section 2.3). The remaining 41 participants (10 male, 31 female) were between the age of 17 and 48 (mean 23, median 19). They were all native speakers of Dutch. None of the participants objected to being recorded, and all of them consented to their data being used for research and educational purposes. Part of the participants who had already taken part as a narrator served as addressees for the two conditions

Figure 1: A design with four conditions: 1. Talking to an audiovisual speech recognition system, 2. Talking to a person via web cam, 3. Talking to a person behind a screen, 4. Talking face-to-face. And an additional level: whether speaker and addressee are in the same room (true in 3 and 4).



that required so.

2.3. Procedure

We randomly assigned participants to one of four conditions. Narrators first read the instructions and could ask any questions they had on the procedure. The experimenter checked whether they were aware of the addressee. They then watched a seven minutes animated cartoon called "Canary Row". We chose this particular cartoon because it has proven to elicit gestures in several other studies, such as McNeill [15] and Alibali et al. [2].

In condition 1, the written instructions said that the signal of the camera was sent to a beta version of an audiovisual speech recognition system (AViSum) that was located in another building on campus, and which would produce a summary of their narration. It was emphasized that the system could process both auditory and visual information. A fake phone call was made by the experimenter to check whether the signal came through all right, and whether the system was ready for use. In reality, there was no such computer system.

In condition 2, the instructions said that the camera was used as a web cam, and that another participant was watching them in another campus building, with the purpose of summarizing their narration afterwards. The experimenter pretended to set up a one-way videoconference with the presumed experimenter in the other building, and then made a fake phone call to check whether the image and sound were received well and whether they were ready to begin. There was no other participant watching.

In condition 3, narrators retold the story to an addressee that we introduced to them as another participant. Addressees were confederates. After the participant had watched the animated cartoon, narrator and addressee were allowed to pose any questions they had about the task. A wooden screen separated them, such that they could not see each other during the story telling. The narrators' instructions stated that the addressee had to summarize the story afterwards, and that they were videotaped with the purpose of comparing the addressees' summary to their narration. We instructed addressees not to interrupt the narrator. There was a little auditory feedback (laughs, uh-huh's).

Condition 4 was similar to condition 3, except that narrators retold the story in a face-to-face situation, thus without the screen in between narrator and addressee.

In each condition, participants were videotaped using a digital video camera. They were seated in front of the camera. The camera position was such that the entire upper part of the body was visible, including the upper legs.

In all conditions, the narrator could look at snapshots of each of the episodes of the cartoon, that hung either on the wall or on the screen in front of them. This was in order to aid memory, and to facilitate more structured, and hence more comparable stories.

After retelling the cartoon, in condition 3 and 4 the experimenter first took the addressee to another room. Narrators then completed a questionnaire, which included questions on how they had experienced the conversation and whether they had believed the experimental setup. We fully debriefed them and asked their consent to use the recordings. The experimenter also asked whether the participants had believed the experimental setup and whether they had suspected any deceit. Two participants in the computer condition had found the setup suspicious and were therefore excluded from the study. All other participants believed the instructions and the experimenter and had no doubts about the experimental setup.

2.4. Transcribing and Coding

We transcribed each narration from the videotape. Repairs, repeated words, false starts, and filled pauses were included.

In the annotation of gestures, we only took movements of the hands into account. We first discriminated between gestures and other movements such as self adjustment. We then coded gestures largely according to McNeill [15], p78–82, and Jacobs and Garnham [9], p3. Gestures were coded as iconic, metaphoric, deictic, beat, interactive, or emblem. Iconic, metaphoric, and deictic gestures were counted as representational gestures (even though they may also have an interactive function). All other gestures were counted as non representational gestures.

2.5. Statistical Analysis

All tests for significance were performed using univariate analysis of variance (ANOVA), with a significance treshold of .05. For pairwise comparisons, the Tukey HSD method was used.

3. Results

3.1. Gesticulation

We first performed an ANOVA with mean gestures per 100 words as the dependent variable and condition as the fixed factor (with levels: computer, web cam, screen and face-to-face). The results are shown in Figure 2. Condition had a significant effect on the number of gestures ($F(3,37) = 5.624$, $p < 0.01$, $\eta_p^2 = 0.313$). Fewer gestures were produced in the computer condition than in the other three conditions. This difference was reliable for the screen and face-to-face condition.

We also found a significant effect of condition on representational gestures ($F(3, 37) = 4.469$, $p < 0.01$, $\eta_p^2 = 0.266$), see Figure 3. Representational gestures were produced at a reliably

lower rate in the computer condition than in the face-to-face condition.

For non representational gestures, we found a significant effect of condition as well (F(3, 37) = 5.478, p $<$ 0.01, η_p^2 = 0.308), see Figure 4. Non representational gestures were produced at a significantly lower rate in the computer condition than in the screen and face-to-face condition.

No significant differences between the mean rates of gestures were found between the web cam, screen, and face-to-face condition.

We also did a univariate analysis of variance with another fixed factor: whether the addressee was in the same room or not (physical co-presence, see Figure 1). For gestures per 100 words we found a significant effect of co-presence (F(1, 37) = 9.851, p $<$ 0.01, η_p^2= 0.210). More gestures were produced when interlocutors were co-present. This also held for representational gestures (F(1, 37) = 7.466, p $<$ 0.01, η^2= 0.168) and for non representational gestures (F(1, 37) = 10.705, p $<$ 0.01, η_p^2= 0.224).

Four of the eleven participants in the computer condition did not produce any gestures. In the other conditions there were no participants that did not gesticulate at all.

Figure 2: Mean gesture rate over conditions.

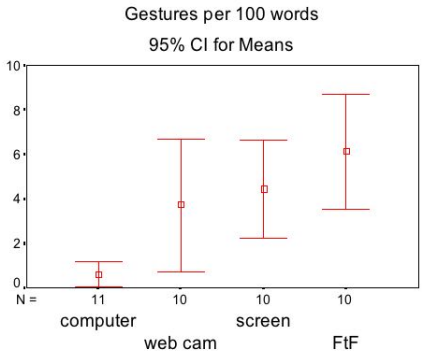


Figure 3: Mean representational gesture rate over conditions.

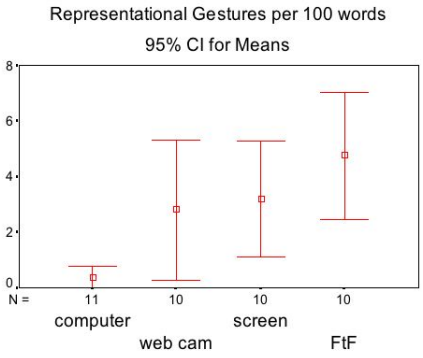
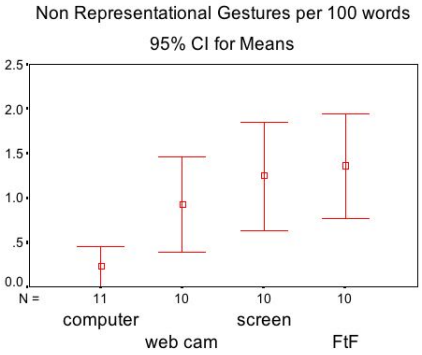


Figure 4: Mean non representational gesture rate over conditions.



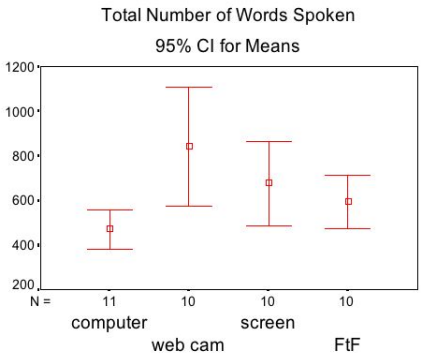
3.2. Speech

Condition had a significant effect on the total number of words used by participants(F(3, 37) = 4.021, p $<$ 0.014, η_p^2 = 0.246)), see Figure 5. In the computer condition, significantly fewer words were used than in the web cam condition.

We also found a significant effect of condition on the number of words spoken per second (F(3, 37) = 5.123, p $<$ 0.005, η_p^2 = 0.294), see Figure 6. Speech was slower in the computer condition than in the screen and face-to-face condition. Speech was faster when interlocutors were physically co-present (F(1, 37)= 11.178, p $<$ 0.01, η^2 = 0.232).

No significant difference was found for the number of filled pauses (i.e. uhs) per word (F(3,37) = 1.849, p = 0.155).

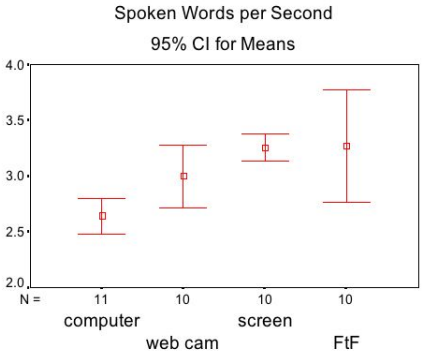
Figure 5: Mean total number of spoken words.



4. Discussion

Fewer gestures were produced by participants who thought to be talking to an audiovisual speech recognition system than by participants who were talking to a human interlocutor, or thought to be doing so. And no reliable differences in gesture rate were found between the three conditions with (presumed) human addressees. So the (presumed) nature of the addressee, either human or artificial, influenced gesticulation. Unlike the media equation predicts, people behaved very differently towards the

Figure 6: Mean words spoken per second.



artificial system than towards real people.

Fewer words were used when talking to the computer system, just as fewer gestures were used. Participants' descriptions to the computer system were less elaborate. Thus, it does not seem that information was transmitted through speech instead of gestures when talking to the computer. Rather, less information was transmitted through both modalities. This corroborates well with the idea that people put less effort in communication with computers, which is put forward in Maes et al. [14].

When measured from filled pauses, speech was just as fluent in the computer condition as in the other conditions. However, participants did speak significantly slower when they thought to be talking to a computer system. This seems to be partly because interlocutors were not co-present, thus, as a result of mediation. And partly because participants put in less effort. It does not seem that participants spoke slower because of limited trust in the computers' ability to capture their speech, since intonation in the computer condition seemed very flat. (We intend to do a phonological analysis on this data in the near future.)

We found that for both representational and non representational gestures, fewer gestures were produced in the computer condition than in the conditions with a human addressee. If the only functional role of gestures were to facilitate speech, this difference would not be expected. Therefore, it seems that both types of gestures can be used communicatively. This corroborates well with the results in Alibali et al. [2] and Jacobs and Garnham [9].

We did not find any significant differences between the other three conditions (web cam, screen, and FtF) for the rate of either representational or non representational gestures produced. This may be due to the large within group variation and the limited number of participants in our study. But it may also be that this difference is indeed very small. This would mean that the influence of mutual visibility on the amount of gestures produced is relatively small.

Our results are not fully in line with the mutually adaptive modalities hypothesis proposed by De Ruiter [6]. Participants still produced many gestures when they could not be seen by the addressee. Although the difference was not significant, participants in the screen condition (in which the addressee could not see them) even produced more gestures than participants in the web cam condition (in which they were told that the addressee could see them).

More gestures were produced when speaker and addressee were in the same room. Thus, it may be that physical co-presence indeed has an influence on gesticulation. Why would speakers produce more gestures when the addressee is in the same room, even when they cannot be seen by the addressee? Bavelas et al. [5] found that being in a dialogue leads to an increased gesture rate. Being physically co-present offers the possibility of a dialogue, whereas a dialogue was not physically possible in our two conditions in which speaker and addressee were not in the same room. In our conditions in which speaker and addressee were co-present, the addressee was instructed not to respond to the narrator, but participants were ignorant about this. Thus, they may well have behaved as if they were in a dialogue.

Also, being physically co-present often offers the possibility of seeing each other. Situations in which people are physically co-present, but there is no chance of being seen by the other person, such as in the screen condition, are somewhat artificial. Therefore, the cognitive process of taking this into account may not be automated (as opposed to some of the gesturing). This is in line with research by Keysar et al. [11], in which addressees consider objects that they can, but the speaker cannot see, as possible referents. As other tasks, in this case recollecting and retelling the story, place demands on speakers' processing capacity, they may not have enough resources left to adapt their gesturing to the unusual context. In other words, to take into account what the addressee can and cannot see. This is in line with the theory proposed in [17], which explains how skill acquisition may not end with automation, and how reflective cognition (such as self-monitoring) and basic tasks performance can be considered a dual-task. To further test this hypothesis, we may vary the difficulty of the basic task in future experiments. For example, investigate whether participants suppress more gestures when the basic task is easier.

In gesture coding, little has been written about the first step of deciding what counts as a gesture. We found that for some speakers, it can be hard to discriminate between self adjustment or seemingly random movements, and for example beats. Also, different categories could have been used for coding in this study. In future work, we could test whether, for example, the distinction made by Alan Cienki in referential gestures (that are subdivided in concrete and abstract reference) and discourse related gestures (that express the speaker's attitude, or have an interpersonal or discourse-structuring function), leads to different results.

Other studies, such as [3] and [5], have looked at the size of gestures, i.e. whether the elbow is raised or not. We plan to use this in our analysis in the near future. As Figure 7 illustrates, there may be differences over conditions in the size of produced gestures. In addition, we intend to take a closer look at the relation between speech and gesture, focussing on reference. We will also do a perception experiment, in which we will measure what differences participants perceive in video clips of participants in the different conditions of this study.

5. Conclusion

Our study shows that the nature of the addressee, either human or artificial, influences gesture production for both representational and non representational gestures. It therefore seems plausible that, in a narrative task, both representational and non representational gestures can be produced with a communicative intent.

Also, we found that narrators communicated their story dif-

Figure 7: Iconic gesture representing Granny hitting Sylvester with an umbrella, produced by a participant in the computer (top left), the web cam (top right), screen (bottom left), and face-to-face condition (bottom right)

ferently to a presumed audiovisual speech recognition system than to a (presumed) human addressee, both verbally and non verbally.

6. Acknowledgements

We thank Carel van Wijk for his help in the statistical analysis, Alan Cienki for his useful feedback and inspiring ideas on the coding of gestures, and Lennard van de Laar for his technical support.

7. References

[1] E. Aharoni and A. Fridlund, "Social reactions toward people vs. computers: How mere lables shape interactions," *Computers in Human Behavior*, vol. 23, pp. 2175–2189, 2007.

[2] M. W. Alibali, D. C. Heath, and H. J. Myers, "Effects of visibility between speaker and listener on gesture production : some gestures are meant to be seen," *Journal of Memory and Language*, vol. 44, pp. 169–188, 2001.

[3] A. Bangerter and E. Chevalley, "Pointing and describing in referential communication: When are pointing gestures used to commnunicate?" in *CTIT Proceedings of the Workshop on Multimodal Output Generation*, I. Van der Sluis, M. Theune, E. Reiter, and E. Krahmer, Eds., 2007, pp. 17–28.

[4] J. Bavelas, N. Chovil, D. Lawrie, and A. Wade, "Interactive gestures," *Discourse Processes*, vol. 15, pp. 469–489, 1992.

[5] J. Bavelas, J. Gerwing, C. Sutton, and D. Prevost, "Gesturing on the telephone: Independent effects of dialogue and visibility," *Journal of Memory and Language*, 2007 (in press).

[6] J.-P. de Ruiter, "Can gesticulation help aphasic people speak, or rather, communicate?" *Advances in Speech-Language Pathology*, vol. 8(2), pp. 124–127, 2006.

[7] R. Fish, R. Kraut, R. Root, and R. Rice, "Evaluating video as a technique for informal communication," in *Proceedings of the SIGCHI conference on human factors in computing systems*. New York: ACM Press., 1992, pp. 37–48.

[8] U. Hadar, "Two types of gesture and their role in speech production," *Journal of Language and Social Psychology*, vol. 8, pp. 221–228, 1989.

[9] N. Jacobs and A. Garnham, "The role of conversational hand gestures in a narrative task," *Journal of Memory and Language*, 2006 (in press).

[10] A. Kendon, "Do gestures communicate? a review," *Research on language and social interaction*, vol. 27, pp. 175–200, 1994.

[11] B. Keysar, S. Lin, and D. Barr, "Limits on theory of mind use in adults," *Cognition*, vol. 89, pp. 25–41, 2003.

[12] E. Krahmer and M. Swerts, "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception, and visual perception," *Journal of Memory and Language*, 2007, to appear.

[13] R. Krauss, "Why do we gesture when we speak?" *Current Directions in Psychological Science*, vol. 7, pp. 54–60, 1998.

[14] A. Maes, P. Marcelis, and F. Verheyen, "Referential collaboration with computers: do we treat computer addressees like humans," in *Anaphors in Text: Cognitive, formal and applied approaches to anaphoric reference.*, M. Schwarz-Friesel, M. Consten, and M. Knees, Eds. Amsterdam: John Benjamins Publishing Company, 2007, pp. 49–68.

[15] D. McNeill, *Hand and Mind: what gestures reveal about thought*. The University of Chicago Press, Chicago and London, 1992.

[16] A. Melinger and W. J. Levelt, "Gesture and the communicative intention of the speaker," *Gesture*, vol. 4:2, pp. 119–141, 2004.

[17] L. Mol, N. Taatgen, R. Verbrugge, and P. Hendriks, "Reflective cognition as secondary task," in *Proceedings of Twenty-seventh Annual Meeting of the Cognitive Science Society*, B. Bara, L. Barsalou, and M. Bucciarelli., Eds. Mahwah (NJ): Erlbaum, 2005, pp. 1925–1930.

[18] A. Özyürek, "Do speakers design their cospeech gestures for their addressees? the effects of addressee location on representational gestures," *Journal of Memory and Language*, vol. 46, pp. 688–704, 2002.

[19] B. Reeves and C. Nass, *The Media Equation, how people treat computers, television, and new media like real people and places*. CLSI Publications, Stanford, California, 1996.

[20] S. Whittaker, "Theories and methods in mediated communication," in *The handbook of Discourse Processes*, A. Graesser, M. Gernsbacher, and S. Goldman, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 2003, pp. 243–286.