LOOK WHO'S TALKING TO WHOM

Mediating Joint Attention in Multiparty Communication & Collaboration

Roel Vertegaal

Cognitive Ergonomics Department
Twente University, Enschede, The Netherlands 1998

Design: Roel Vertegaal and Mirjam Netten

Printing and Binding: Febodruk BV, Enschede, The Netherlands

Pictures: page 12: by courtesy of Bill Buxton; page 21: by courtesy of A. Yarbus; pages 39 and 48: by courtesy of Robert Slagter; page 94: by courtesy of Harro Vons

Samenstelling promotiecommissie:

Voorzitter, secretaris:	prof.dr. P.M.G. Apers
Promotoren:	prof.dr. B.M. Velichkovsky (TU Dresden, Germany)
	prof.dr.ir. A. Nijholt
Assistent-Promotor:	dr. G.C. van der Veer (Vrije Universiteit Amsterdam)
Leden:	prof.ir. D. Bosman (emeritus)
	prof. W.A.S. Buxton (University of Toronto, Canada)
	prof.dr. S.J. Mullender

Published by:

Cognitive Ergonomics Department
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands
e-mail: roel@acm.org

ISBN 90-3651-1747

LOOK WHO'S TALKING TO WHOM

Mediating Joint Attention in Multiparty Communication & Collaboration

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof.dr. F.A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 4 september 1998 te 16:45 uur.

door
Roeland Petrus Hubertus Vertegaal
geboren op 13 juli 1968
te Hazerswoude

Dit proefschrift is goedgekeurd door depromotoren:	prof.dr. B.M. Velichkovsky
	prof.dr.ir. A. Nijholt
en assistent-promotor:	dr. G.C. van der Veer

to Mirjam, Tamas, Gerrit and Ted
for having a little faith in me

Contents

Preface

During one of the many moments in the past four years that I had gotten out of my car to take a photograph of another stunningly beautiful natural landscape, I had just framed what felt like the perfect shot. It was so splendid, that I hesitated. What if I would take just one step further to the left? Might the picture be even better? Of course, I took that step. It was worse. I took another step. Worse still. Three days later, after having walked a full circle around the mountain I was trying to photograph, I returned to my starting position. And it *was* the best picture I had seen in the whole of the full circle. Although I had answered 360 questions, one for each degree in the circle, one question remained. What had I learned?

I would like to thank all those who helped me complete the full circle presented in this book. First and foremost, I would like to thank my wife, Mirjam, who is even more beautiful from the inside than she is from the outside. Without her great support, I could never have poured so much energy into this project. Four years ago, at the start of this project, I stood by the deathbed of my former promotor, the late Ted White. I promised him then I would complete this book. It is dedicated to the loving memory of this wonderful man. Throughout those four years, Gerrit van der Veer has done a great job supervising my work as a highly intelligent coach, and as a walking encyclopedia of scientific methodology. Politics stood, as usual, in the way of Gerrit becoming my promotor. I therefore thank Boris Velichkovsky (a great inspirer) and Anton Nijholt for carrying out this task. I also thank Nancy and Dixon Cleveland, and the whole LC Technologies team, for their kind and exceptional support throughout this project. Without them, it would have failed miserably. Of course, I am greatly indebted to my graduate students, Harro Vons and Robert Slagter. They were instrumental, not just in building the GAZE Groupware System, but also as ideal assistants during the experiments presented in this book. I thank them for all the energy and intelligence they put into what I always considered *our* project. Other persons and institutions I would like to thank are: Tamas Ungvary (for being my friend), Bert Lenting (for being around), Ronald Leenes, Herman Adèr, Pieter v/d Berg, Marian van Blanken, Jolijn Hendriks, Dave Kasik of Boeing Commercial Airplanes, Luuk Lagerwerf, F. Luteijn, Axel Mulder, Robert Rathbun of Cyberian Outpost, Bart Schermer, Arjen de Vries, Armanda Zandberg, members of the Ergonomics Department, committee members, TU Delft WITlab, NWO and Shell Travel Services.

Roel Vertegaal *Enschede, Augustus 1998*

Chapter 1

Introduction

EYES THAT FASCINATE

Throughout history, the eye, and its function in providing vision of the world around us, has fascinated researchers. As demonstrated by the works of 13th-century scholar Grosseteste [67], until the late Middle Ages, it was common belief in the western world that the eyes observe things by emitting rays that touch objects. This theory was advanced by Empedocles in the 5th century BC [46], and passed on to Grosseteste by the 2nd-century scholar Galen [56, 57]. It is intriguing that Empedocles' theory prevailed for so long, given that Leucippus and Democritus, contemporaries of Empedocles, had already proposed a particle theory of light, stating that all objects consist of atoms and that objects become visible because their atoms swarm into the eye [93]. It is perhaps *because* the eyes are fascinating, that Democritus' theory was rejected for so long. Greek and Roman myths — like the story of Narcissus [112], who turned into a flower after gazing at his own reflection for too long — show that Empedocles' theory, and its persistence, could well have been based on wide-spread belief in the *oculus fascinus*: the fascinating or evil eye. According to Gifford [61], the fear that the eye is capable of projecting the malignity of its owner through rays, inflicting injury wherever gaze happens to fall, is one of the more ancient and persistent of superstitions*. Sumerian clay tablets, excavated in modern Iraq and dating back to the 3rd Millennium BC, already tell the story of Ereshkigal, goddess of the underworld, who had the power to kill Inanna, goddess of love, with a deadly eye [47]. It is odd, with the power of gaze known to mankind for at least 5000 years, that it took until the 1960s before experimental psychologists such as Kendon [85] and Argyle [6] started to investigate seriously the functions of gazing at others in human social interaction. Although, since then, much has been discovered about the role of gaze directional information in two-person conversations, still very little is known about its functions and effects in group communication. This, and the application of such knowledge in the design of telecommunication systems, are the subjects of this thesis.

---

* Just ask some friends whether they believe they can sense it when someone looks at them from behind.

RESEARCH QUESTIONS AND GOALS

The general questions underlying this thesis were: (1) how important is the conveyance of gaze directional information in mediated group communication, (2) what are the functions of gaze directional information in group communication, and specifically, to what extent does gaze directional information indicate who is talking or listening to whom, (3) what is the isolated effect of gaze directional information on the group communication process, relative to that of other nonverbal visual information provided by the human upper torso, and (4) how can answers to the above questions be applied in the design of mediated systems for group communication (and collaboration). The rationale behind these questions will be discussed in detail in the next chapter.

With regard to addressing the above research questions, the two main objectives were: (a) to contribute to scientific knowledge pertaining the functions and effects of gaze directional information in group communication (b) to improve the design of mediated systems for group communication (and collaboration) based on that knowledge. These two goals are at extreme ends of a continuum, from acquiring fundamental psychological knowledge on human social interaction to developing new technologies and applications. It is in the combination of these two extremes, that we believed the applied science of Cognitive Ergonomics, within which realm this study was carried out, would be best served.

THESIS OVERVIEW AND SUMMARY

This section provides a brief overview of the structure of this thesis, including a summary of main conclusions per chapter. Each chapter may be read as a standalone paper:

Chapter 2 discusses the problem of providing essential characteristics of human communicative behaviour in telecommunication systems. We conclude that in synchronous group (or *multiparty*) communication using mediated systems, there might be problems with, amongst others, the regulation of turntaking and the referencing of other individuals. This might be caused by an absence of certain attention-related information in the audio or video signals that mediate the communication. As a result of this absence, it may be difficult to establish who is talking or listening to whom in a nonverbal fashion. We identified gaze direction — a representation of the visual attention of others — as a candidate for providing such information.

Chapter 3 is an empirical investigation into the extent to which the gaze direction of others — their focus of visual attention — *might* function as an effective indicator of whom they are talking or listening to — their focus of dialogic attention — in four-person face-to-face conversations. We found that, although this is subject to individual and situational differences, the gaze

direction of others may indeed be considered a good indicator of their dialogic attention towards individuals in multiparty conversations.

Chapter 4 is an empirical investigation into the isolated effect of gaze directional information on multiparty mediated communication, relative to that of other nonverbal visual cues provided by the human upper torso. We found a significant positive effect of the presence of *gaze at the facial region* on the number of speaker turns and of head orientation on the number of deictic references. As such, the conveyance of gaze at the facial region may be considered an important requirement in the design of multiparty mediated communication systems.

Chapter 5 is a practical study into the design of multiparty mediated communication and collaboration systems that convey gaze directional information. It discusses how awareness about others may be constituted in an integral fashion by conveying the locus and span of their (visual) attention. We present the GAZE Groupware System, which uses advanced desk-mounted eyetracking systems, rather than a spatial setup of video camera/monitor units, to convey the visual attention of participants metaphorically in a web-based virtual meeting room. The main benefits of this approach are the integration of information about the attention of others towards persons as well as objects in a shared workspace. The separate gauging of gaze directional information also allows a more flexible and scalable use of network bandwidth.

Chapter 6 provides a summary and integration of the main empirical and practical conclusions presented in this thesis. It also provides a discussion of potential future directions of research.

Two appendices at the end of this book provide detailed information on the definition of terms used in this thesis, and materials used in the empirical study presented in Chapter 4.

Chapter 2

Problems in Multiparty Mediated Communication and Collaboration

INTRODUCTION

With recent advances in network infrastructure and computing power, desktop video conferencing and groupware systems are rapidly evolving into technologically viable solutions for remote communication and collaboration. Video conferencing is no longer limited to expensive circuit-switched ISDN networks and is starting to be used over standard Internet connections in conjunction with groupware software. The central premise for the use of video-mediated communication (VMC) over traditional telephony has been that video images improve the quality of communication between individuals by increasing the available sensory bandwidth. In a face-to-face situation, auditory, visual and haptic expressions (or *cues*) are freely combined to convey messages and regulate interaction. It has been presumed that by adding video to an audio-only communication link, mediated communication would bear a significantly closer resemblance to face-to-face communication. In this chapter, we will first discuss why this need not necessarily be the case. We will show why designing mediated systems is a problem of conveying the least redundant cues first. An example of a cue which seems hardly redundantly coded, yet typically not conveyed by mediated systems, is the gaze direction of participants. We will discuss how a lack of gaze directional information in mediated systems may lead to problems in the support of group (or *multiparty*) communication. According to usability studies, because of this lack, participants may have insufficient knowledge on who is talking or listening to whom (the dialogic attention of other participants). As a further investigation of this problem, we will discuss existing empirical evidence regarding the function and effect of gaze directional information in multiparty communication. We will also discuss existing implementations of mediated systems that preserve gaze directional information. We conclude that little is known about the function and isolated effect of gaze directional cues as a provider of information about dialogic attention in multiparty communication. Based on this conclusion, we present the problem definition for this thesis. Firstly, we need to know more about the extent to which gaze directional information might code who is talking or listening to whom in a multiparty setting. Secondly, in order to assess whether such information is actually used, we need to know more about the isolated effect of providing gaze directional information in multiparty (mediated) communication. Finally, if gaze directional information is a requirement in the design of multiparty mediated systems, we need to know more about how such

information could be gauged, mediated, and represented such that it may provide added value in a mediated communication and collaboration setting. As our review of existing systems will demonstrate, achieving this in an integral, transparent and technically scalable fashion is not a trivial task.

CONVEYING THE RIGHT CUES

Face-to-face communication is an extremely rich process in which people have the ability to convey an enormous amount of information to each other. In mediating the process of human communication, it is not obvious that such information richness is easily replicated by adding video images to standard telephony. Indeed, empirical studies (see Sellen [126]) show the difference between face-to-face communication and video-mediated communication to be significantly greater than the difference between video-mediated communication and audio mediated communication. We may indeed attribute such findings to the large difference in sensory bandwidth between face-to-face and mediated conditions. Sensory bandwidth is characterized by the number of *cues* (actions which convey information from one human to another) conveyed by the different media. Verbal cues are the actual words spoken in a conversation, nonverbal cues include the way in which these words are spoken (paralinguistic speech), facial expressions, gaze, gestures, bodily movement, posture and contact, physical proximity and appearance [6]. Theoretically, the notion that we can simulate face-to-face situations under mediated conditions is a correct one. In practice, however, it seems that the number of cues that need to be conserved in order to accomplish a complete replication is far greater than one would expect. Simply adding video is only a minor step. And in conditions where much of the information is redundantly coded, it might be an insignificant step where it comes to improving regulation of conversations or task performance [126]. The notion that the addition of video images should make mediated communication significantly more like face-to-face communication may have been based on a misinterpretation of Short et al.'s Social Presence Theory [129]. In this theory, communication media are ranked according to the degree in which participants feel co-located. Face-to-face communication would provide the greatest sense of social presence, followed by video, multi-speaker audio and monaural audio. This ranking was based on a factor analysis of subjective ratings of dyadic (two-person) conversations using the various media, and does indeed suggest that the amount of social presence is improved by increasing the number of cues conveyed. So why then does the addition of video images to audio-only communication seem to be an insignificant step towards replicating face-to-face conditions where it comes to regulation of conversations or task performance? We believe this may, to a large extent, be attributed to a typical redundant coding scheme for those visual cues that are conveyed by a single stream of video.

Cue	***Category***	***Perceptual Channel***	***Telephony***	***Traditional VMC***
Completion of grammatical clause	verbal	auditory	yes	yes
Sociocentric expression such as 'you know'	verbal	auditory	yes	yes
Drawl on final syllable	paralinguistic	auditory	yes	yes
Pitch shift at end of phonemic clause	paralinguistic	auditory	yes	yes
Drop in loudness	paralinguistic	auditory	yes	yes
Termination of hand gesture	gestural	visual	no	if in view
Relaxation of body position	postural	visual	no	if in view
Resumption of eyegaze	gaze-related	visual	no	no

Table 2-1. *Taxonomy of cues speakers may use when releasing the floor.*

As Short et al. themselves pointed out, when cues are redundantly coded, we can no longer predict the effects of a communication system upon interaction by listing differences in the number of cues conveyed by different media. For example, a speaker preparing to yield the floor to a listener may use a combination of the following expressions (see Table 2-1): completion of a grammatical clause; a sociocentric expression such as 'you know'; a drawl on the final syllable; a shift in pitch at the end of the phonemic clause; a drop in loudness; termination of a hand gesture; relaxation of body position; and resumption of eyegaze towards the listener [44, 85, 129]. Note that we see a merging of verbal, paralinguistic, gestural, postural, and gaze-related cues, all indicating the same thing. When confronted with a different medium, speakers may easily adapt their behaviour by using different combinations of cues or by simply dropping several cues without failing to yield the floor. Indeed, half of the nonverbal cues in the above example are auditory, and five of the total of eight cues could be conveyed by telephone. This makes it extremely hard to find differences between video-mediated communication and audio mediated communication in terms of performance in a joint task or, for that matter, more objective variables of conversational structure (such as number of interruptions, duration of simultaneous speech or number of utterances). Indeed, empirical studies have so far failed to find clear differences in terms of conversational structure or task performance between video- and audio mediated communication (for an excellent overview, see Sellen [126]). When improving mediated communication, should we therefore aim to model face-to-face conditions even closer? We agree with Dennett [36] that it is not very realistic to think that face-to-face situations can, or indeed should be substituted by modelling the world on a one-to-one basis (a question already raised by

Descartes [37]). Although improving mediated communication by means of increased bandwidth for motion video may be considered an important area of research, we believe we should avoid putting too much research emphasis on this. Instead, we should first focus on providing nonverbal cues which seem less redundantly coded in speech, thereby hoping to provide some essential characteristics of face-to-face communication without intending to substitute it completely.

PROBLEMS WITH MEDIATING MULTIPARTY COMMUNICATION

In multiparty conditions (in which more than two persons communicate), gaze direction may well serve as a good example of such a cue. Multiparty conversational structure is much more complicated than its dyadic equivalent. As soon as a third speaker is introduced, the next turn is no longer guaranteed to be the non-speaker. When the number of participants rises beyond three, it becomes possible to have side conversations between subgroups of people. This can pose problems for the regulation of, for example, turntaking. When we consider the above example of a speaker yielding the floor in a multiparty situation, the question arises to whom he would like to yield the floor. With the notable exception of gaze direction (or rather the general orientation of body, head and eyes) and perhaps pointing gestures, such attention-related information is not coded by the eight cues listed in Table 2-1. It can only be conveyed by telephone by means of explicit verbal references (e.g., calling someone by name) or the internal context of conversation. We believe turntaking problems with current multiparty conferencing systems (regardless of whether they use video or audio) may be attributed to a lack of cues about other participants' attention. Isaacs and Tang [78] performed a usability study of a group of five participants using a typical desktop video conferencing system. They found that during video conferencing, people needed to address each other by using each other's names and started to explicitly control the turntaking process by requesting individuals to take the next turn. In face-to-face interaction, however, they saw many instances when people used their eyegaze to indicate whom they were addressing and to suggest a next speaker. Often, when more than one person started speaking at the same time, the next speaker was determined by the eyegaze of the previous speaker without the need for conventions or explicit verbal intervention. Similarly, O'Connaill et al. [109] found that in video conferencing more formal techniques were used to achieve speaker switching than in face-to-face interaction. They too attribute this to the absence of certain speaker-switching cues. This suggests that multiparty communication using video conferencing is not necessarily easier to manage than using telephony.

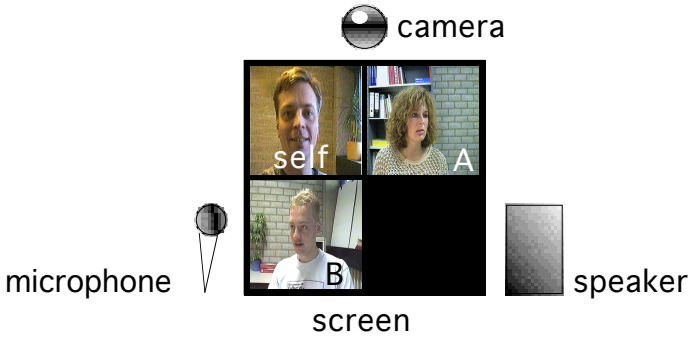


Figure 2-1. *A single-camera video conferencing system.*

Single-camera video systems such as the one shown in Figure 2-1 do not convey deictic visual references to objects (e.g., on the computer screen) or persons (such as the other participants) outside the frame of reference of the camera any more than telephony. To some extent, the participants' presumption that video conferencing is more like face-to-face interaction than telephony may actually lead to inappropriate use of such visual cues. Isaacs and Tang [78] show how, when a participant points to one of the video images on her screen, it is difficult for the others to use spatial position to figure out whom is being addressed. Similarly, subjects may try to establish eye-contact by gazing at the video image of a participant. Although the large angle between the camera and the screen usually prevents looking each other in the eyes (as one would need to look at the camera and the video image simultaneously), even if they were to establish eye-contact, they would establish it with every participant in the group.

Conveying gaze directional cues may be a way of preventing the above usability problems. This might, for example, be done in multiparty mediated systems by conveying the following information [152]:

1) *Relative Position.* Conveying the relative viewpoints of participants based on a common reference point (e.g., around a shared workspace), may provide a common spatial context.
2) *Head Orientation.* Conveying the general orientation of looking might help participants in achieving deixis (e.g., *"What do you think?"*), and might provide support for knowing who is attending to whom.
3) *Gaze.* Conveying the exact position of looking within each other's facial region might also help in achieving deixis, and might provide support for knowing whether others are still attending. *Mutual gaze* constitutes eye-contact.

In the next section, we will review empirical studies into the function of gaze directional cues in (multiparty) human communication.

THE CASE FOR CONVEYING GAZE DIRECTION IN MULTIPARTY COMMUNICATION

According to Argyle and Kendon, in two-party communication, gazing at other persons serves at least five functions: to regulate the flow of conversation; to provide feedback on the reaction of others; to communicate emotions; to communicate the nature of relationships; and to avoid distraction by restricting input of information [7, 8, 85]. Due to technological and methodological complications, most studies into the role of gaze direction in communication were limited to two-person (dyadic) situations. In the early seventies, Argyle [6] estimated that when two people are talking, about 60 percent of conversation involves gaze, and 30 percent involves mutual gaze (or eye-contact). People look nearly twice as much while listening (75%) as while speaking (41%). The amount of gaze is also subject to individual differences such as personality factors and cultural differences. For example, an extravert may gaze more frequently than an introvert. Also, there is more gaze in some kinds of conversations than others. If the topic is difficult, people look less in order to avoid distraction. If there are other things to look at, interactors look at each other less, especially if there are objects present which are relevant to the conversation [10]. In general, however, gaze seems closely linked with speech. According to Kendon [85], person A tends to look away as she begins a long utterance, and starts looking more and more at her interlocutor B as the end of her utterance approaches. This pattern should be explained from two points of view. From the first point of view, in looking away at the beginning, person A may be withdrawing her attention from person B in order to concentrate on what she is going to say. When she approaches the end of her utterance, the subsequent action will depend largely upon how person B is behaving, necessitating person A to seek information about her interlocutor. From the second point of view, these changes in gaze can come to function as signals to person B. In looking away at the beginning, person A signals that she is about to begin an utterance, forestalling any response from person B. Similarly, in looking at person B towards the end of her utterance, she may signal that she is now ceasing to talk yet still has attention for him, effectively offering the floor to person B. According to Argyle [6], however, the main reason why people gaze at the end of their utterances is that they need feedback on the other's response: to see if they are still attending, and to see how the verbal message was received. Argyle argues that although gaze may function as a *minor* stop signal for synchronization purposes, other verbal and paralinguistic cues are more important. However, such other signals may only function effectively in a multiparty setting if information about their destination is included.

So how do the above results hold in a multiparty condition? In one of the few studies on gaze in triads, Exline [50] found less gaze at individuals than in dyadic studies. Averaged across groups, they found 36% of gaze by individuals during listening activity, and 31% during speaking activity. This finding may be attributed to divided attention of individuals during multiparty listening and

speaking activity. It also suggests that differences between gaze while listening and while speaking become smaller with group size. Indeed, in one of the few (unpublished) studies on gaze in groups of more than three, Weisbrod [160] found that attendees of a seven-member seminar gazed over 70% of their speaking time, but only 47% of their listening time. This is a complete reversal of the pattern found in dyadic communication. What is more interesting, is that Kendon [85] attributed this reversal to a need to make clear to whom one is speaking. However, as a means for assessing the extent to which gaze might code such information, one would need to compare percentages of gaze at the person(s) spoken or listened to with percentages of gaze at *others* than the person(s) spoken or listened to. To our knowledge, this has never been the subject of an empirical study.

We conclude that there simply have not been enough studies into the function of gaze directional cues in multiparty communication. In the next section, we investigate to what extent gaze directional cues are actually used in multiparty communication, discussing empirical findings on the effects of their presence in a mediated setting.

The Effect of Gaze Direction on Multiparty Mediated Communication

Very few, if any, studies exist in which the isolated effect of representing gaze direction in multiparty mediated communication has been empirically evaluated. Sellen [127] examined the differences in conversational structure between three multiparty conditions: using face-to-face communication; using a single-camera desktop video conferencing system (similar to the one depicted in Figure 2-1); and using a Hydra system (see Figure 2-2): a setup with multiple cameras, monitors and speakers which preserves relative position (including separation of audio), head orientation and, to a large extent, gaze (see page 15 for details on the Hydra system) [125]. Although Sellen found differences in terms of objective measures (such as amount of simultaneous speech and speaker switching time) between face-to-face and mediated conditions, she did not detect any differences between the two mediated systems. Sellen attributed this, in part, to the small screens of the Hydra system and their separation. As Heath and Luff [73] pointed out, movements in the periphery of vision which appear on a screen lose their power to attract attention*.

Qualitative data and informal discussions with subjects did indicate they preferred the Hydra system over single-camera video conferencing. Reasons given included the fact that they could selectively attend to people, and

* The study presented in Chapter 4 suggests the still-present parallax between camera and screen may have been a factor.

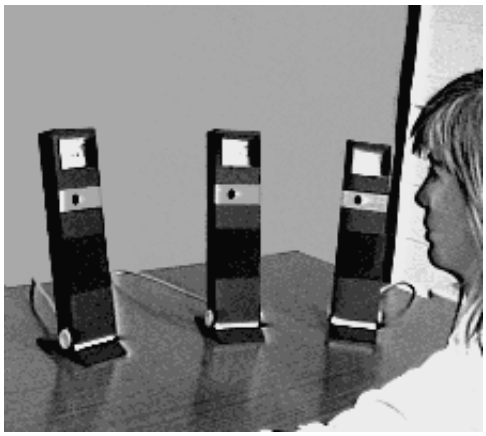


Figure 2-2. *The Hydra System (from Buxton [25]).*

could tell when people were attending to them. They also confirmed that keeping track of the conversation was the most difficult in the single-camera video conferencing condition. However, such conclusions may, in part, also be attributed to the separation of audio sources in the Hydra system. There were many more of these, potentially confounding, differences between conditions in the above study. Although the qualitative results may be considered promising, we therefore cannot regard this study as conclusive with regard to the *isolated* effect of gaze directional cues on multiparty communication.

We conclude that there simply have not been enough studies into the isolated effect of gaze directional cues on mediated multiparty communication. In the next section, we will investigate the role of gaze directional cues in collaboration, discussing empirical evidence for their preservation in mediated collaboration systems.

THE CASE FOR CONVEYING GAZE DIRECTION IN COOPERATIVE WORK

We have so far examined the role of gaze direction in multiparty communication. Although some studies have investigated the role of looking at things during face-to-face collaboration, there are, to our knowledge, few empirical studies examining the effect of conveying gaze direction during computer supported cooperative work. Argyle and Graham [10] found that if a pair of subjects were asked to plan a European holiday and there was a map of Europe in between them, the amount of gaze dropped from 77% to 6.4%. They spent 82% of the time looking at the map. Even when they presented a vary vague, outline map, subjects looked at it for 70% of the time, suggesting that they were keeping in touch by looking at and pointing to the same object, instead of looking at each other. They also found there was little attention for the map if it was irrelevant to the topic of conversation.

Within the realm of computer supported cooperative work, Ishii and Kobayashi [79] demonstrated how the preservation of relative position and the transfer of gaze direction could aid cooperative problem solving through their ClearBoard system. They conducted an experiment in which two participants were asked to solve the "river crossing problem", a puzzle in which two groups of people (typically missionaries and cannibals) should reach the other side of a river with certain restrictions on who can join whom in the boat. According to the authors, the success of this game depends heavily on the point-of-view of the players. Participants could see video images of each other through a shared drawing board on which they could also sketch the problem. Ishii and Kobayashi concluded that it was easy for one participant to say on which side of the river the other participant was gazing and that this information was useful in jointly solving the problem. Colston and Schiano [33] describe how observers rated the difficulty people had in solving problems, based upon their estimates of how long a person looked at a particular problem, and how his or her gaze would linger after being told to move on to the next problem. They found a linear relationship between gaze duration and rated difficulty, with lingering as a significant factor. This suggests that people may use gaze-related cues as a means of obtaining information about the cognitive activities of a collaborator. Velichkovsky [151] investigated the use of eyetracking for representing the point of gaze during computer supported cooperative problem solving. Two people were asked to solve a puzzle represented on their screen as a random combination of pieces which had to be rearranged using the mouse. The two participants shared the same visual environment, but the knowledge about the situation and ability to change it on the way to a solution were distributed between them. One of the partners (the expert) knew the solution in detail but could not rearrange the pieces. The other (the novice) could act and had to achieve the goal of solving the puzzle without having seen more than a glance of the solution. In the first condition, they could only communicate verbally. In the second condition, the gaze position of the expert was added by projection into the working space on the screen of the novice. In the third condition, the

expert used his mouse instead to show the novice the relevant parts of the task configuration. Both ways of conveying the attention of the partners improved performance. The absolute gain in the case of gaze position transfer was about 40%. Approximately the same gain was obtained with mouse pointing. In a second experiment, the direction of gaze position transfer was reversed from the novice to the expert. Here too, a significant gain was found in the efficiency of distributed problem solving. Apparently, experts could see the types of barriers novices confront in their activity and were therefore able to give more appropriate advice. This shows that gaze position transfer may be useful in situations where manual deixis is impossible: the novices could not use their mouse for pointing because they needed it to manipulate puzzle pieces.

We conclude that although the effect of providing a representation of gaze direction in cooperative work may be highly dependent on the task situation, a closer coordination between the communication and cooperation media with respect to conserving such deictic cues can be considered beneficial. In the next section, we will review existing systems in which gaze directional cues are preserved.

MULTIPARTY MEDIATED SYSTEMS THAT PRESERVE GAZE DIRECTIONAL CUES

Over the years, a number of multiparty conferencing systems have been developed that conveyed gaze directional cues by preserving relative position, head orientation and gaze. Negroponte [103, 104] describes a system commissioned by ARPA in the mid-1970s to allow the electronic transmission of the fullest possible sense of human presence for five particular people at five different sites. Each of these five persons had to believe that the other four were physically present. This extraordinary requirement was driven by the government's emergency procedures in the event of a nuclear attack: the highest ranking members of government should not be hiding in the same nuclear bunker. His solution was to replicate each person's head four times, with a life-size translucent mask in the exact shape of that person's face. Each mask was mounted on gimbals with two degrees of freedom, so the 'head' could nod and turn. High-quality video was projected inside of these heads. In this rather humorous setup, each site was composed of one real person and four plastic heads sitting around a table in the same order. Each person's head position and video image would be captured and replicated remotely. According to Negroponte, this resulted in lifelike emulation so vivid that one admiral told him he got nightmares from these 'talking heads'. A technical advantage of this system was that only one camera was needed at each site to capture the video image of the participant's head, resulting into only one stream of video data from each participant (we will further address this issue below). A technical disadvantage was the elaborate setup of the talking heads: the total number of heads required is almost the square of the number of participants ($n^2 - n$; in which n is the number of participants).

Sellen [126] describes the Hydra system, a setup of multiple camera/monitor/speaker units in which relative position (including spatial separation of audio), head orientation and gaze might be preserved during multiparty videoconferencing. Hydra simulates a four-way round-table meeting by placing a box containing a camera, a small monitor and speaker in the place that would otherwise be held by each remote participant (see Figure 2-2 on page 12). Each person is therefore presented with his own view of each remote participant, with the remote participant's voice emanating from his distinct location in space. This way, when person A turns to look at person B, B is able to see A turn to look towards B's camera. According to Sellen, eye-contact (i.e., mutual gaze) should be supported because the angle between the camera and the monitor in each unit is relatively small. The separation of audio in the Hydra system may ease selective listening, allowing participants to attend to different speakers who may be speaking simultaneously. Although Hydra is of course a very elegant alternative to Negroponte's system, it has some disadvantages. One disadvantage is that although participants can see when someone is looking at a (shared) workspace, their estimation of where this person looks within that workspace would probably be worse than possible with, e.g., Negroponte's system. A more technical drawback is that each camera in the setup provides a unique video stream, and that the number of cameras required is almost the square of the number of participants ($n^2 - n$; in which n is the number of participants). For three participants, only six Hydra units are needed, but when this number rises to five, twenty Hydra units are required. In a Multicast network [49], the bandwidth requirements of traditional single-camera video conferencing systems are greatly reduced. With Multicasting, a video stream of an individual user is not sent to individual remote participants by means of multiple connections. Instead, that video stream is 'broadcast' to all other participants simultaneously, requiring only one unit of the total network bandwidth at any time. With the Hydra system, such compression cannot be achieved, causing the amount of network bandwidth used to convey video to rise with almost the square of the number of participants ($n^2 - n$). This may have an effect on usability, as it may lead to problems with proper conveyance of motion video information [154].

Okada et al.'s MAJIC system uses a rather more elaborate setup in an attempt to achieve a seamless integration of life-size images of the other participants with each participant's real work environment [110]. In essence, it is a bigger version of the Hydra system, with a more precise positioning of cameras, *behind* the monitors. In each office, a thin half-transparent curved projection screen is placed behind a computer terminal in front of the user. On this screen, life-size video images of the other participants are projected. Behind each projection screen, video cameras are located at the center of the projected facial region of the other participants, one camera for each participant. This way, head orientational information is conveyed, and users may achieve eye-contact by looking at each other's faces. A corresponding placement of microphones and speakers is used to ease selective listening. We may well consider the MAJIC

system the closest we will get to replicating a face-to-face situation without holographic projection. However, the disadvantages of MAJIC are similar to those of the Hydra system. In addition, due to the large image size, each video stream will require considerably more bandwidth than with the Hydra system, assuming resolution is maintained.

A more recent development has been the embodiment of chat participants in virtual environments [16, 102]. Whereas such systems include efficient ways to pictorially represent users using spatial metaphors, they typically do not comprise a transparent way of capturing gaze directional information. Although we will briefly discuss their use of representations in Chapter 5, we refer to Harrison and Dourish [72] for a more detailed discussion on the issues concerning such Collaborative Virtual Environments.

In summary, it is difficult to design multiparty mediated systems such that gaze directional information is conveyed in communication as well as collaboration, in a manner that is technically scalable.

PROBLEM DEFINITION

We conclude that in synchronous multiparty communication using mediated systems, there might be problems with, amongst others, the regulation of turntaking and the referencing of other individuals. This might be caused by an absence of certain attention-related information in the audio or video signals that mediate the communication. As a result of this absence, it may be difficult to establish who is talking or listening to whom in a nonverbal fashion. We identified gaze direction — as a representation of the visual attention of others — as a possible candidate for providing such information.

However, little is known as to what extent gaze directional cues are actually suitable for coding whom is being addressed or listened to in multiparty conversation. We need to know whether the focus of visual attention of others might predict their focus of articulatory or auditory attention (i.e., their dialogic attention). One approach to studying this is to determine whether observers, during listening or speaking activity, gaze more at the persons they are addressing or listening to, than at others. This is the subject of the empirical study presented in Chapter 3.

If the visual attention of others effectively codes their dialogic attention, one needs to verify whether a representation of such information is actually used in multiparty communication. Very few, if any, empirical studies exist into the isolated effect of representing visual attention — in the form of gaze directional information — on multiparty (mediated) communication. Whether there is an effect on mediated communication, and to what extent such effect might be attributed to (a lack of) knowledge about the dialogic attention of others, is the subject of the empirical study presented in Chapter 4. In order to assess the relative importance of conveying gaze directional information, we will compare any effects with those caused by the other upper-torso nonverbal visual cues for which video-mediated systems already provide support. Results of this study were to be generalized such that they could be used as design recommendations for the preservation of visual attention in multiparty mediated communication systems.

If conveyance of visual attention is a fundamental requirement for multiparty mediated *communication* systems, and given available evidence for conveying such information in mediated *collaboration* systems, could we generalize this into an integrated design model for multiparty mediated systems? How can we gauge, convey, and represent information about the attention of participants for communication and collaboration in an integrated and transparent fashion, such that its function is not only preserved, but possibly augmented? How could this be implemented in a multiparty mediated communication and collaboration system in a manner that allows a scalable and efficient use of network resources? This is the subject of the design study presented in Chapter 5.

Summary of Approach

The emphasis in this thesis is on the functions and effects of knowing the visual attention of others in multiparty mediated communication. The first — empirical — part investigates to what extent visual attention correlates with dialogic attention during multiparty communication, and examines the effect of its representation on variables of the multiparty mediated communication process. The second part of this thesis approaches the coding of attention in multiparty mediated systems as a practical design problem. It discusses a simple design rationale and candidate solutions for conveying the attention of others in mediated communication and collaboration systems in an integrated, transparent, scalable and possibly augmentative fashion.

Chapter 3

Visual Attention as Predictor of Dialogic Attention in Multiparty Communication

ABSTRACT
We investigated to what extent the focus of visual attention of others *might* function as an effective indicator of their focus of dialogic attention. We examined this by measuring the amount of time subjects spent looking at the facial region of conversational partners listened or spoken to during four-way face-to-face discussions. We compared those findings with the amount of time subjects spent looking at others than the individual listened or spoken to. We found that gaze at the facial region may indeed be considered an excellent indicator of dialogic attention towards individuals in multiparty conversations. When someone is listening to an individual, there is an 88% chance that the person gazed at is the person listened to. When someone is addressing a single individual, there is a 77% chance that the person gazed at is the addressed individual. In this more or less dyadic condition, we found about 1.6 times more gaze while listening (62%) than while speaking (40%). When a speaker addresses more than a single individual, it seems likely that gaze may still be considered an effective indicator of his dialogic attention. When addressing a triad, speaker gaze typically seems to be distributed evenly across listeners. However, the total amount of speaker gaze rises significantly to about 59% of time. In such situations, the amount of gaze received by individual listeners (20%) is therefore still significantly more than the amount of gaze they would have received when not addressed (12%). However, our estimates of the effectiveness of gaze as a indicator of dialogic attention may not be considered free of individual differences in, for example, personality. In addition, our

estimates may be generalized only to situations where there is no requirement to look at task objects.

INTRODUCTION

Before discussing our research regarding the effects of a representation of the visual attention of others on multiparty turntaking (treated in the next chapter), we will discuss to what extent such a representation *might* function in multiparty face-to-face communication to indicate the dialogic attention of others, coding whom others are talking or listening to. We examined this by studying the converse situation: the extent to which the visual attention of a person for others relates to his dialogic attention for others, i.e., synchronization between looking at others and listening or speaking to others. We approached this by treating visual attention as a dependent variable of dialogic attention, taking other relevant factors such as the personality of the onlooker into account. We will first discuss the measurement of visual attention as a dependent variable, after which we will discuss each of the independent variables.

Visual Attention as a Dependent Variable in Multiparty Communication

We determined the visual attention of individuals in four-way face-to-face conversations by measuring their point of gaze using a desk-mounted eyetracking system. We will first discuss how point of gaze can be informative with regard to the focus of visual attention of an onlooker. We will then discuss the practicalities of measuring point of gaze in a multiparty communication setting.

The Relation Between Point of Gaze and Focus of Visual Attention

Selective visual attention is an important mechanism in human parsing of visual information. With visual attention, human can allocate their limited resources to the processing of the most relevant visual information in a given situation, without being overloaded by irrelevant aspects of the visual world [105]. Indeed, only a small area of the retina (with a range of approximately 2°) is equipped for acute vision: the *fovea*. The rest of the retina provides parafoveal and peripheral vision, which can be regarded as providing contextual information of low resolution in terms of spatial detail and colour precision, at high speed. So for acute vision of a particular region in the visual array (our projection of the world), this region needs to be foveated first [113]. This process involves the use of small rotating movements of the eyeball to aim a region at the fovea. By detecting potential regions of interest, peripheral vision appears to play an important role in bottom-up guidance of the selection of regions for foveation [63, 140]. This process may be aimed at a need for higher-level identification of simple lower-order feature sets which *pop out* of a visual scene (such as colour, brightness, movement and orientation of visual elements) [142]. In addition, expectations at a semantical level about the relevance or

relations of objects in the visual field are seen as an important factor in the top-down (or cognition-directed) guidance of the foveation process.

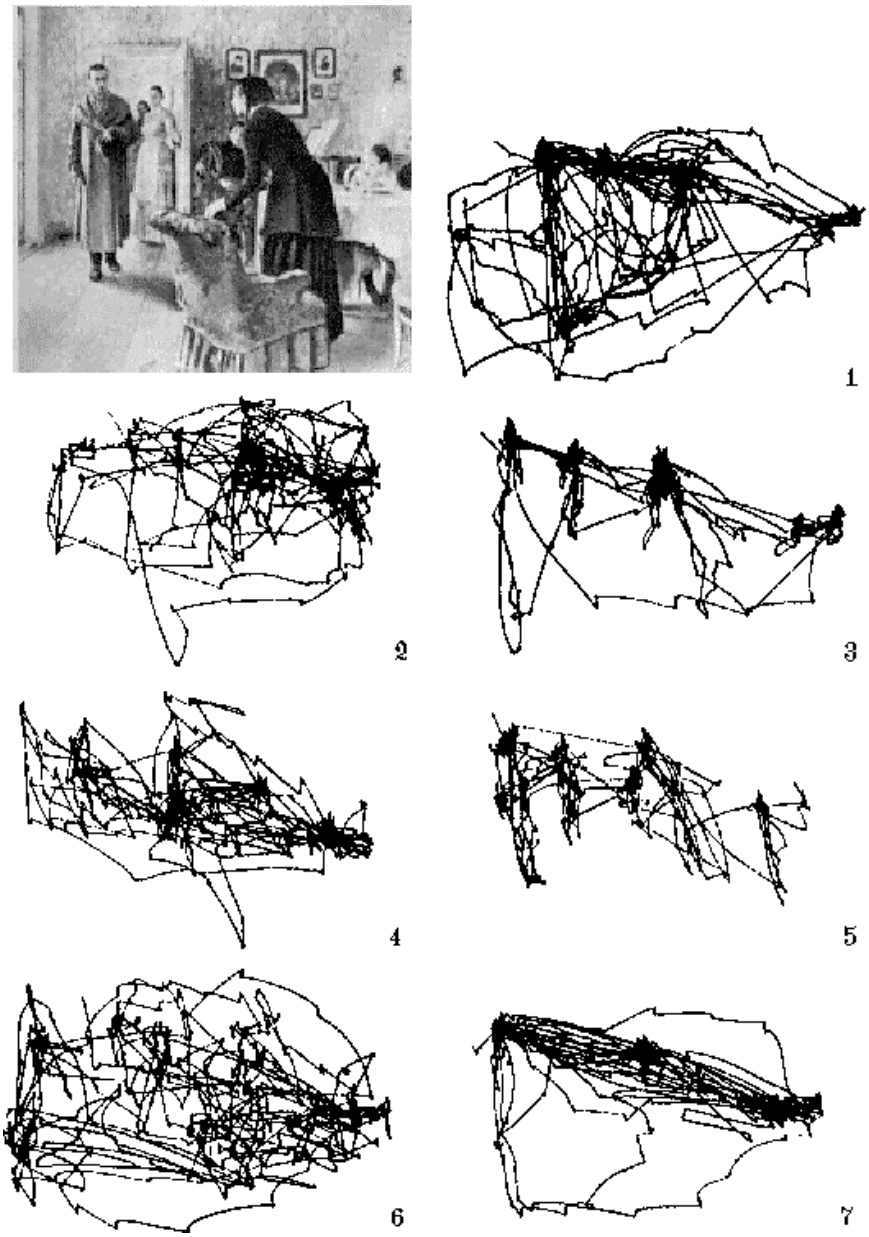


Figure 3-1. *Patterns of foveation at a picture after different instructions: 1. Free examination; 2. Determine material circumstances; 3. Determine age of people 4. Determine activities prior to arrival of visitor; 5. Remember clothing; 6. Remember*

positions of people and objects; 7. How long was the visitor away? (from Yarbus [165]).

Yarbus showed that the semantics of a situation have considerable influence on the pattern of foveation by an observer [165]. The picture shown in Figure 3-1 is scanned quite differently when the observer is asked to estimate the ages of the people in it than when asked to estimate their wealth. Similarly, Biederman et al. [18] found that, using visual search, subjects were able to locate the position of a bicycle in a picture more rapidly if the bicycle could be found in a natural context, i.e., a place where bicycles are usually located.

So is selective visual attention defined by the process of foveation? Although this is still the subject of further investigation, it seems clear that the process of foveation is an important *hardware* filter component of visual attention. However, there is evidence that visual attention also comprises a software filter, or *attentional spotlight*. According to Posner [113], it is indeed possible to direct the focus of attention to other parts of the visual array than the foveated area. However, in general, it seems that the location of the software filter is well correlated with the location of the hardware filter [88]. We may conclude that the location of the hardware filter, or *point of gaze*[*], provides the best available estimate of the locus of visual attention [149, 151]. Since the center of the pupil always corresponds to the center of the foveated area, this location can be gauged by determining the angular position of the pupil(s) relative to the visual scene [149].

The Relation Between Point of Gaze and Interest

Although this is also the subject of further investigation, it seems that both top-down and bottom-up guidance of visual attention are aimed at connecting low-level feature sets with higher-level, semantical information [142, 143]. Although point of gaze by itself cannot reveal what a person is thinking, it does give a good indication of the interest of the observer for objects in the outside world, particularly during spontaneous looking (free examination in Figure 3-1) and task-oriented looking (other assignments in Figure 3-1) [84, 95, 105]. According to Kahneman [84], there is one exception to this rule: foveation during thinking. When an observer is attending to some inner cognitive process, he might not be attending to the foveated information at all. However, according to Argyle et al. [12], in such situations the observer will typically foveate an area containing little information. This can often be identified as looking away, or staring, and therefore need not necessarily confound measurements of attentional focus.

---

[*] The location of the hardware filter is actually a direction. *Point of gaze* refers to the angular position of the filter relative to a visual scene.

Measuring Gaze in Multiparty Face-to-Face Communication

As we have seen in the previous paragraph, the best possible estimate of the locus of visual attention is given by the orientation of the pupil in between eye movements. When the pupil remains relatively still for at least 120 ms we speak of a *fixation* [150]. It is during fixations that visual information in the foveal area is being processed. Relocation of the foveated area is characterized by very rapid ballistic movements of the eyes, or *saccades*. It is very likely that processing of visual information is suspended during saccades [20]. Pupil orientation is therefore most informative with regard to the locus of visual attention during fixations. Since there are other forms of eye movement, during which visual processing is *not* suspended, *fixation points,* as given by the orientation of the pupil (relative to the visual scene) during fixations, are most informative when the visual scene is relatively stable. An operational definition of the amount of visual attention for a particular part of the visual scene is given either by the frequency or total duration of fixation points within that area [86].

We were, however, interested in the visual attention of subjects for their conversational partners. This is known as *gaze*: the act of looking at others [6]. By and large, the facial region — particularly the region of the eyes — seems to be the focal point of visual attention for others during face-to-face communication [8, 85]. The terms *gaze* and *gaze at the facial region* can therefore be seen as virtually synonymous, with *mutual gaze* (i.e., two people gazing at each other) constituting *eye-contact*. Measuring *gaze* thus requires the detection of fixations within the area of the visual field occupied by another person's body, with the center of gravity at the location of that person's eyes. Per observer, this yields a binary variable `gaze` for each of his conversational partners, which is *true* when the observer gazes at that partner, and *false* when not. This variable is a function of time.

Most research on the role of gaze in communication, which took place during the 60s and early 70s, relied on human observers for coding `gaze`. A typical laboratory setup would include a half-silvered mirror placed as a wall between two observers and two subjects. By placing both dyads parallel to the wall, observer 1 would be able to observe the gaze of subject *b* at subject *a*, and observer 2 the gaze of subject *a* at subject *b* (see Figure 3-2). However, the subjects would not be able to see the observers. Each observer would press a button on an interaction recorder to register `gaze` through time. There is an impressive body of work on the reliability of such observational measurements of gaze [50, 86, 92, 121, 138, 156]. In general, the reliability and validity of such measurements is good, but only if the angle and distance between the observer and the observed is small (see the text box on page 115 of Chapter 4 for a discussion). At right angles to the interactors, an observational approach is not acceptable in scientific terms.

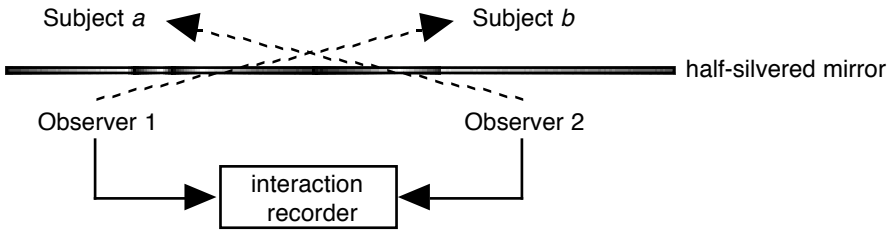


Figure 3-2. *Laboratory arrangements for observing gaze in dyads (from Argyle [6]).*

According to Argyle and Cook [8], ideally, observers should be located at the position of the other interactor, facing the observed person. In some studies, this problem was approached by placing a camera above the head of the other interactor, registering gaze behaviour for later analysis. A clear advantage of this approach is the ease with which one can determine inter-observer reliabilities, since observations can easily be repeated using the same material. However, Stapley [134] demonstrated that if the lens of the camera is not located at the exact position of the interactor's eyes, a systematic error of judgement may occur. It is evident that methodological difficulties in the measurement of `gaze` have been predominant reasons for the limitation of most studies to dyadic communication.

With great foresight, Argyle and Cook [8] suggested a radical new approach, which uses direct measurement of pupil position to register `gaze`. In the mid-70s, several techniques were available to track the position of the eyes with better accuracy than human observers in ideal situations. However, successful application of most of these *eyetracking* techniques was inhibited at the time by the obtrusive nature of the registration equipment. Firstly, the head of the observed subject typically needed to be fixed *completely*. Secondly, rather bulky attachments to the head and sometimes the eyeballs were required for accurate measurement. For recording `gaze` in social interaction, this equipment left much to be desired. Freedom of movement, the ability to speak and a normal appearance of the subject are highly desirable for recording `gaze` in normal human social interaction [8].

Although not all of the above problems have been completely solved with current-day eyetracking systems, in the past 20 years eyetracking has developed into a much more viable measurement technique. Many of the above preconditions are no longer requirements, and although the use of eyetracking still raises methodological issues, many of the problems with observational techniques are effectively circumvented. Current-day eyetrackers are capable of registering fixations with higher spatial resolution than possible with human observers, at a higher temporal resolution than practical with human observers. Stolk [139] gives an excellent overview of available eyetracking techniques. From his study, it becomes apparent that the best eyetracking technique for use

in a face-to-face communication setting is the *pupil center/corneal reflection* method. It has a large measurement range (±40°); it requires no head attachments yet hardly any head stabilization (with freedom of head movement of a few cm); its temporal resolution is satisfactory for studying fixations; its spatial resolution is better than human observers (< 1°); and the impact on subject discomfort, awareness and appearance is minimal. In addition, it typically provides digital output of point of gaze and is relatively cheap. We employed an LC Technologies' Eyegaze system, which uses an infrared camera for registering point of gaze of a single eye. This camera could be placed unobtrusively on a table in front of a subject. We refer to the *Operationalization* section, from page 42 onward, for a detailed description of the used measurement procedure.

The most important methodological problem regarding the use of the Eyegaze system was the limited horizontal head movement of about 5 cm. In order to reduce rotary head movements by a subject during measurement we used a very comfortable headrest, which did not *afford* such behaviour. In addition, we needed to position all his conversational partners within an angle defined by the area of overlapping vision of the left and right eyes: approximately 50 degrees [1]. This could be achieved if only one subject was measured at a time, seated at some distance from his three conversational partners. These conversational partners, in their turn, had to be placed relatively close together. Argyle and Dean [9] found that distance is positively correlated with the amount of `gaze` in dyads. Most studies, however, report a levelling off of this effect at distances larger than 1.8 m [11]. Hall [71] reported this relationship might indeed not be continuous. He suggested that there are four zones of social distance: *intimate* (0- .45 m); *personal* (.45- 1.2 m); *social consultative* (1.2- 3 m); and *public* (over 3 m). We thus placed all four individuals approximately within the *social consultative* range, which matched nicely with the relative unacquaintedness of conversational partners and the distances used in earlier studies [11]. The measurement of only a single person's gaze per session did, however, imply that we could not study mutual gaze.

A second methodological issue was the measurement of body areas of the three conversational partners. If conversational partners were to move a lot, automatic detection of fixations at their body region would be difficult, complicating measurement of `gaze`. We therefore seated the conversational partners around a rectangular table, with their chairs as close to the table as was still comfortable. This setting effectively afforded minimal movement of the upper-torsos of the three conversational partners. We established the average location of the eye region of each partner by asking the subject to fixate on the eyes of each partner, while asking that partner to rotate her head. From this data, we could determine the average center of gravity of a circular region within which fixations would be considered as `gaze`. This process is also discussed in detail in the *Operationalization* section, on page 44.

In the next section, we will consider the relation of `gaze` with each of our independent variables, discussing previous experiments in order to establish our experimental hypotheses.

Independent Variables and Their Effect on Gaze

We studied the effect of a number of independent variables on gaze of individuals at each of their three conversational partners during four-person face-to-face discussion sessions. Firstly, we examined the extent to which the amount of gaze at a conversational partner is affected by having dialogic attention for that partner: do people gaze more at the person they listen or speak to than at others? This would yield a measure of the extent to which gaze of others might be useful as a predictor of whom people are talking or listening to. Secondly, we were interested whether results would hold for both directions of dialogic attention: while listening and while speaking to persons. Thirdly, we studied the influence of the extent of articulatory attention: do people gaze more when speaking to a group than when speaking to an individual? Finally, we gauged the influence of two relevant personality factors on gaze behaviour: extraversion and autonomy.

In order to allow for as natural a communication setting as possible, all independent variables were measured, rather than controlled. Testing of effects took place on subsets of dependent variable data assembled according to categorized measurements of the independent variable. For each independent variable, we will now discuss its relevance, summarize its operationalization and, based on literature, develop experimental hypotheses with regard to its effect on gaze behaviour in multiparty communication.

Gaze and Focus of Dialogic Attention

In order to estimate the chance that the individual a person gazes at might be the individual that person is addressing or listening to, we compared the percentage of time spent gazing at the person on which dialogic attention is focused, with the percentage of time spent gazing at others* than the person on which dialogic attention is focused.

We define *dialogic attention* as having either auditory or articulatory attention for person(s). *Auditory attention* is defined as listening to a person, *articulatory attention* as speaking to one or more person(s). The focus of dialogic attention is thus an identifier of the person(s) spoken or listened to. In dyadic situations, it is more or less evident whom is being addressed or listened to. With the possibility of group as well as individual discussions, conversational structure is far more complicated in a multiparty setting. Consequently, speech activity scores, as used in most dyadic studies (see Argyle and Cook [8] for an

* We did not use the percentage of time spent gazing at other *individuals*, since it is likely that humans are better in estimating gaze at themselves, than in estimating gaze at other individuals (see discussion in Box 4-2 on page 115). Our chance estimate is thus a conservative one.

overview), are insufficient for determining the focus of dialogic attention in true multiparty settings. We approached this problem by asking subjects themselves to indicate whom they were attending to during a discussion session. While watching a video recording of a discussion session in which they participated, they pressed keys on an accord keyboard to score which person(s) they focused their dialogic attention on during speech and listening activity. Details of this scoring procedure, including a discussion of its validity, are given in the *Operationalization* section, on page 46.

Predictions Regarding Gaze and Focus of Dialogic Attention

Since almost all previous studies investigated gaze in dyadic communication only, not much is known about the amount of time spent gazing at others than the individual one speaks or listens to. However, on average in dyadic communication, people may gaze at their conversational partner over 60% of the time when they have dialogic attention for that partner, and are not distracted by visual tasks [11]. Other studies report similar findings, although typically with a considerable variance [85, 106]. In one of the few (unpublished) studies on gaze behaviour in larger groups, Weisbrod [160] found a mean percentage of 58% gaze during dialogic activity in a seven-member group. It therefore seemed probable that the percentage of time left for gazing at others than the person at the focus of dialogic attention should be comparatively low. In order to test this, we defined the following hypothesis with regard to the percentage of time spent gazing at conversational partners other than the individual in the focus of dialogic attention:

H1 *"On average, subjects spend significantly more time gazing at the individual on which their dialogic attention is focused, than at others"*

Next, we will discuss why we evaluated Hypothesis 1 separately for each of the two modes of dialogic attention: auditory (**H1a**) and articulatory attention (**H1b**).

Gaze and Mode of Dialogic Attention: Auditory vs. Articulatory

We defined dialogic attention as being constituted by one of two modes: having auditory attention (listening to a partner) *or* having articulatory attention (speaking to one or more partners). There is evidence that the mode of dialogic attention influences gaze behaviour. According to Argyle and Ingham [11], in dyadic communication within the social consultative range, individuals typically gaze at each other almost twice as much while listening (75% of time) as while talking (41% of time). Nielsen [106] found 62% gaze while listening, and 38% gaze while speaking. Other dyadic studies, such as Kendon [85], report similar differences between modes of dialogic attention. According to Argyle and Cook [8], speakers may avert their gaze from their interlocutors to reduce interference of visual information processing with processing of verbal information. Indeed, Kendon [85] found that gaze is typically averted during moments of hesitant or slow speech. According to him, looking away may aid

the speaker in organizing his utterance as well as in signalling an intent to hold the floor.

In dyadic studies, the modality and occurrence of dialogic attention was typically measured using an observational approach, similar to that used for recording `gaze`. Observers would score utterance boundaries using buttons on an interaction recorder. In these studies, listening activity was simply defined by utterances of the other individual. Similarly, we used utterance data to disambiguate dialogic attention scores. If a subject scored dialogic attention for a person, *and* that person had an utterance, the subject had `auditory attention` for that person. If a subject scored dialogic attention for persons *and* he had an utterance himself, the subject had `articulatory attention` for those persons. In order to reduce the complexity of this disambiguation process, in which all data was to be kept in sync, we used automated analysis of individual speech patterns to obtain utterance data. See the *Operationalization* section on page 45 for details on this procedure.

Predictions Regarding Gaze and Mode of Dialogic Attention

Since the process of listening and speaking to individuals in a multiparty setting might resemble a dyadic situation, we believed it probable that we would find clear differences in gaze behaviour between modalities of dialogic attention. In multiparty communication, gaze might not be as effective a predictor of articulatory attention towards individuals as of auditory attention towards individuals. We therefore tested all hypotheses, including **H1**, separately for these two modalities of dialogic attention. In addition, we formulated the following hypothesis with regard to the difference between the percentage of time spent gazing at individuals while listening and while speaking:

H2 *"On average, subjects spend significantly less time gazing at an individual they speak to, than at an individual they listen to"*

Gaze and Extent of Articulatory Attention

We define the *extent* of dialogic attention as the number of people on which dialogic attention is focused. Assuming one can only listen to a single individual, but speak to many, this variable is meaningful only with regard to articulatory attention. When speaking to a group, one might expect visual attention to be divided equally over group members. One might also expect the average amount of gaze per individual to drop to very low levels, given the already low percentage of gaze by speakers found in dyads. Individuals might have difficulties predicting that they are being addressed when the percentage of gaze at individuals when speaking to a group becomes lower than the percentage of gaze at others when speaking to an individual. Although one might speculate that people could perceive patterns of gaze at individuals by speakers as an indication of dialogic attention to a group with the extent of that pattern, it seems evident that low levels of gaze would raise serious questions with regard to the efficacy of gaze as an indicator of dialogic attention.

We measured the extent of articulatory attention by counting the number of keys pressed simultaneously by subjects during their scoring of the focus of dialogic attention. Thus, we could isolate moments of conversation for any number of conversational partners addressed. Details of the scoring procedure are given in the *Operationalization* section, on page 46.

Hypotheses About Gaze and Extent of Articulatory Attention

One can attempt to predict the percentage of time spent gazing at one individual when talking to a group of *n* individuals (GT_n) from the percentage of time spent gazing at that individual when talking to that individual only (GT_1):

$$Expected\ GT_n = \frac{GT_1}{n} \qquad \textit{(Equation 3-1)}$$

The null hypothesis with regard to speaking to a group of three would thus be that the mean percentage of gaze at individuals should equal one third of the mean percentage of gaze at an individual when speaking to that individual only. There is, however, evidence that this null hypothesis would not hold, and that the observed percentage of gaze at individuals when speaking to a group is in fact higher than the expected percentage of gaze suggested by this formula. In one of the few studies on gaze in triads, Exline [50] found less gaze at individuals than in dyadic studies by Argyle and Ingham [11]. However, the difference between these percentages was smaller than one would expect. In an unpublished study of a seven-person seminar, Weisbrod [160] found that subjects gazed over 70% of their speaking time, but only 47% of their listening time. This is an interesting reversal of the pattern observed in dyadic studies.

Kendon [85] attributed this reversal to a need to make clear to whom one is speaking, indeed, the subject of our study. Since the lower percentage of gaze while speaking was traditionally attributed to gaze avoidant behaviour due to thinking, the high percentage found by Weisbrod might suggest such behaviour is overridden by the need to specify dialogic attention in multiparty behaviour. However, a second, possibly complementary, explanation for such effect might lie in the *Intimacy Equilibrium* hypothesis by Argyle and Dean [9]. One of the most basic effects of gaze is a heightening of arousal of the individual looked at [8]. Depending on context*, that individual may interpret gaze as a communication of interest, liking, loving, dominance or hostility [8]. Similarly, gaze avoidance may be interpreted as indifference, evasiveness, coldness, submissiveness or defensiveness [87]. Argyle and Dean [9] suggested there is an optimal level of intimacy for different communication situations, and that gaze, in regulating the occurrence of mutual gaze, is an important factor in maintaining this equilibrium. Other factors which affect this equilibrium include: physical proximity, intimacy of topic, and amount of smiling. Rather

* Although we believe such link may exist, to our knowledge, a relation between physiological arousal, context and the interpretation of gaze has never been formally verified.

than group size as such, we believed it very likely that *extent* of articulatory attention is amongst these factors. Speakers can typically gaze only at a single individual at a time. Thus, when a group of individuals is addressed, speaker gaze would have to be timeshared between listeners. With large extents, the function of speaker gaze to maintain an appropriate level of intimacy with each listener would thus be impaired, unless the speaker would gaze more than in a dyadic setting. Thus, one would expect speakers to gaze more with larger extents in order to maintain an acceptable level of intimacy with each individual in their audience. A third, possibly complementary, reason for more gaze with larger extents might lie in visual feedback requirements. In order to successfully monitor the nonverbal responses of all individuals, speakers would need to gaze more [8].

So we have three theories (communicating articulatory attention; maintaining equilibrium of intimacy; and increased visual feedback) which *all* predict a positive relationship between gaze and extent of articulatory attention. We could not verify the contribution of each of these, possibly complementary, explanations. Instead, we tested the relationship itself by means of the following hypothesis:

H3 *"On average, the time subjects spend gazing at an individual when addressing a group of three is more than one third of the time spent gazing when addressing a single individual"*

In order to test the effectiveness of gaze as a predictor of articulatory attention when addressing larger groups, we added another hypothesis:

H4 *"On average, subjects spend significantly more time gazing at an individual when addressing a group of three, than at others when addressing a single individual"*

Gaze and Location of Attended Person

According to Argyle and Cook [8], gaze in dyads is related to the spatial relations between the two people. When interaction starts between two people there is an immediate tendency to orient towards each other. Diebold [39] suggested this is due to an orientation reflex. This reflex causes interactors to line up the facial-visual and vocal-auditory channels, to look at the face which is the source of the sound. As discussed, we tried to reduce this behaviour of the tracked subject. His position did not *afford* head reorientation, necessitating use of eye movements only. However, this did result in a situation of which we needed to verify that it did not confound our measurements. Aiello [3] found more gaze when interactors were directly facing each other. We were worried subjects might spend more time looking at partners opposite them, than at partners left and right of their position. We therefore constituted spatial location of the attended person as a control variable. We evaluated this control variable by comparing, for each location in space, the percentage of time spent by subjects gazing at partners seated at that location.

Hypotheses Regarding Gaze and Location of Attended Person

Our hypothesis with regard to the effect of the location of an attended person on the percentage of time spent gazing at the person was simply the null hypothesis:

H5 *"The relative spatial location of a person on which dialogic attention is focused does not significantly influence the time spent looking at that person"*

We evaluated Hypothesis 5 separately for each of the two modes of dialogic attention: auditory (**H5a**) and articulatory attention (**H5b**).

Gaze and Individual Differences: Personality Factors

According to Argyle and Cook [8], gaze, like other aspects of behaviour, is a product of situations, persons, and the interaction between these. Gaze behaviour is extremely complex, and can be accounted for in terms of many situational and personality-oriented factors. Amongst these are distance; availability of other things to look at; topic of conversation; culture; sex; affiliation; age; mental disorders; and personality. Since our hypotheses were evaluated using relative differences in gaze behaviour *within* subjects, our conclusions need not *necessarily* be affected by these variables. However, before generalizing our results, we did wish to gain *some* insight into the effect of individual differences on the absolute levels of gaze found. We chose to evaluate two attributes of personality which are frequently associated with gaze behaviour: extraversion/introversion and dominance/dependence.

We measured these variables using the Five Factor Personality Inventory [74]. We operationalized extraversion/introversion by means of its *extraversion* scale, and dominance/dependence by means of its *autonomy* scale [15]. We evaluated the effect of these variables by correlation between percentage of time spent gazing and personality score. Details of personality measurements are given in the *Operationalization* section, on page 51.

Hypotheses Regarding Gaze and Personality Factors

Extraversion has been the subject of many studies into gaze in dyadic communication. The excellent overview by Argyle and Cook [8] shows the most strongly confirmed result is that extraverts gaze more frequently, especially while talking. Extraverts *may* also gaze for a greater percentage of time [98], although Rutter [119] found no significant effect in this respect. According to Argyle and Cook [8], extraverts may need to gaze more due to a lower level of cortical arousal [54], which needs to be compensated by gaze. Alternatively, extraverts may have a higher need for affiliation, which is expressed by greater use of gaze [50]. We believe these explanations are possibly complementary, as they are of different levels.

Gaze seems to be often interpreted as an act of dominance or power. One would therefore expect dominant individuals to gaze more than dependent individuals. Oddly enough, studies which investigated this found an opposite effect. Exline and Long [52] studied gaze in dyads in which a power differential was artificially created by the task situation. They found that higher power subjects gazed *less* than low power subjects. Since some subjects involved army officers, they also investigated the relative rank of subjects, finding the same effect. Other studies seem to confirm this. Exline and Messick [53] found that overall, dependent subjects gazed more. A possible explanation for this lies in dominant individuals not following the rules of maintaining the Equilibrium of Intimacy [9]. According to this theory, there would be two ways in which dominant individuals can control their listeners using gaze: by denying an appropriate level of intimacy (gazing too little), *or* by maintaining too high a level of intimacy (gazing too much). Either way, they can demonstrate their independence. By typically gazing less, they ensure maximum arousal effect when their gaze *is* required to outstare a conversational partner. Indeed, there is evidence that dominant individuals will gaze more, but only when their conversational partner gazes more than the Intimacy Equilibrium requires [51].

We therefore formulated the following hypotheses with regard to the effect of extraversion and autonomy on the percentage of time spent gazing at individuals:

H6 *"On average, subjects with a higher score on extraversion spend significantly more time gazing at the individual on which their dialogic attention is focused"*

H7 *"On average, subjects with a higher score on autonomy spend significantly less time gazing at the individual on which their dialogic attention is focused"*

We evaluated these hypotheses separately for each of the two modes of dialogic attention: auditory (**H6a** and **H7a**) and articulatory attention (**H6b** and **H7b**).

METHODS

Our experiment involved groups of four participants discussing current-affairs topics of their choice in a face-to-face meeting. This section discusses the methods used to conduct the experiment, describing experiment design, subjects, experimental task, instructions and session procedure.

Experiment Design

Each subject participated in four discussion sessions: one in which their point of gaze was measured and three in which it was not. Since we expected all subjects to be familiar with the task of discussion, we did not expect any learning or order effects. Even so, we randomized the order in which subjects were assigned to their seats, with one constraint: each person had to be seated behind the eyetracker *once*. In order to allow for as natural a communication setting as possible, all independent variables were measured, rather than controlled. To evaluate our hypotheses, we used a within-subjects procedure, comparing subsets of dependent variable data assembled according to categorized measurements of the independent variable.

Experimental Subjects

Our experimental subjects were paid volunteers, mostly university students from a variety of technical and social disciplines. Prior to the experiment, we tested all subjects on eyesight, personality, and their ability to operate the eyetracking system. We also asked their opinion on a range of discussion topics. In order to reduce variance between discussion groups, we allocated each subject to a group in a way that matched groups on the following matching variables:

- *Extraversion*. Extraverts not only seem to gaze more, but also to speak more than introverts [119]. Since gaze behaviour of individuals may be related to gaze behaviour of their conversational partners, we ensured the mean extraversion per group was as close to normal as possible. By distributing extraverts evenly over groups, we would also increase the chance of having animated discussions in all groups. We used subject scores on the Five-Factor Personality Inventory [74] to reduce between-group variance on this variable.
- *Autonomy*. Since we expected autonomous individuals to behave differently with regard to gaze than dependent individuals, we ensured the mean autonomy per group was as close to normal as possible. By distributing autonomous individuals evenly over groups, we would also increase the chance of having someone take the initiative in each group, while it would decrease the chance of having too much debate between two highly autonomous individuals. We used subject scores on the Five-Factor Personality Inventory [74] to reduce between-group variance on this variable.

- *Sex*. Females seem to gaze more than males, and there may be an interaction between the amount of gaze and group composition with regard to sex [50]. Given the limited availability of females, we attempted to assign one female to each group. 5 of the 7 groups used for further analysis consisted of 1 female and three males, two consisted of males only.
- *Age*. We also minimized age differences between groups.

The subset used for further analysis consisted of 24 subjects (5 female, 19 male, mean age 24) from 7 discussion groups, out of a total of 48 subjects and 12 discussion groups. Of the remainder, early sessions were used to fine-tune apparatus and methodology, and some were skipped due to individual problems with calibration of the eyetracker or training procedures. For every experiment, a fifth subject was invited. This person knew he would only be required if one of the other subjects in his group was missing, but was otherwise treated as a normal subject. This safety measure was needed four times (over 12 groups). To avoid distorted measurements by unnatural behaviour of subjects, they were not informed of the exact purpose of the experiment. Instead, they were told we were interested in the relation between pupil dilation and speech behaviour. This deceit was chosen as subjects would be unable to control pupil dilation [75, 135]. All subjects were informed of the true purpose of the experiment after treatment.

Task

Each group of four subjects participated in four 8-minute discussion sessions, resulting in a total conversation time of 32 minutes per group. Prior to the experiment, we asked the opinion of subjects on a range of controversial current-affairs issues, including the competence of our future king, the effect of computer use on child development, etc. Topics did not include intimate or undesirable issues, as this might have influenced looking behaviour [9]. Per topic, each subject indicated his opinion on a 5-point scale ranging from *strongly agree* to *strongly disagree*. We selected the most controversial topics (as illustrated by a bipolar frequency distribution of answers) for use as discussion material. This way, we increased the chance of having animated discussions. Although groups used the same pool of topics, each group was allowed to skip to a topic of preference at any moment during a session.

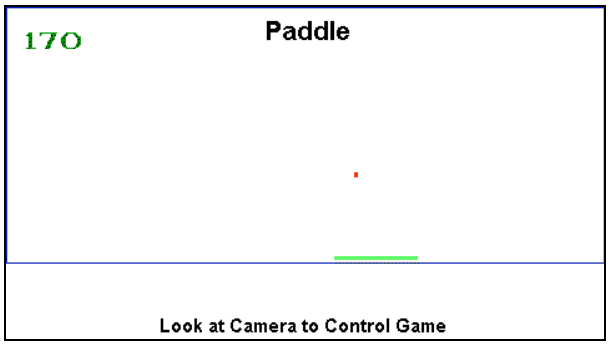


Figure 3-3. *Paddleball. When tracking the ball with the eyes, the paddle would automatically follow its position, making this a game you cannot loose.*

Instructions and Session Procedure

Before participating in the experiment, each group was given 15 minutes to get acquainted in an informal setting. After having spent 10 minutes with the others, the first subject was asked to join the experimenter for eyetracker training. The person seated behind the eyetracker will from now on be called *subject*. Others will be called *conversational partners*.

Prior to Each Session

The subject was seated behind the eyetracker and asked to the use the headrest in such a manner that she would not need to rotate her head. The headrest was adjusted to support the head and neck. The eyetracker was then calibrated in order to correct for individual differences in eyeball geometry. This procedure was repeated until the calibration error levelled (see page 43 for details).

In order to train the subject on the range within which she could move her head, she played a game of *Paddleball* projected on a video screen behind the discussion table (see Figure 3-3). The purpose of this game is to prevent a ball from reaching the bottom of the screen. This is done by bouncing the ball using a *paddle*, moved horizontally along the bottom of the screen. Since the horizontal component of the subject's point of gaze was used to move this paddle, the subject could prevent the ball from reaching the bottom of the screen simply by tracking the ball with her eyes. However, when she moved her eye out of eyetracker range, she would loose control of the paddle until her eye was moved back again. In order to give feedback on this process, subjects could see the eyetracker camera image of their eye during this game. With each successful bounce the subject scored points. The game ended when failing to bounce the ball. We repeated the game until the subject's score levelled (with at least 5 consecutive bounces), or after more than 30 consecutive bounces.

Subsequently, the conversational partners were admitted to the room and seated according to plan. All participants were given a microphone headset, which was adjusted for use (see the *Operationalization* section on page 45 for details). They were asked not to tinker with the microphone during the session.

Next, the center of gravity of the facial region of each of the three conversational partners was determined. We asked the subject to track the eyes of each conversational partner, while registering point of gaze (see the *Operationalization* section on page 44 for details). In order to avoid biasing the subject, we also asked her to look at the hands of each conversational partner.

Group members were given three discussion topics, and were asked to involve everyone as much as possible. They were also told they were free to move, as long as they remained seated.

After Each Session

After 8 minutes of conversation, the experimenters interrupted the discussion. The three conversational partners took a break, while the subject was directed to the control room. There, she was asked to score whom she had spoken or listened to during the session.

To practice for this task, the subject watched a 5-minute pre-recorded video of an enacted session in which the focus and occurrence of dialogic attention was specified according to a script. Care was taken to include all possible permutations of dialogic attention in this video, including side conversations. The video was shot from the point of view of one of the actors, as in Figure 3-4. The subject was instructed to score whom this actor might have been speaking or listening to by pressing identifier keys on an accord keyboard whenever there was speech activity (see the *Operationalization* section on page 46 for details). While scoring, subject key presses were automatically compared with the score specified by script. Performance was measured by calculating, for each 15-second interval, the percentage of time in which the subject score was equal to the pre-specified score. When the percentage of overlap was better than 60% for at least 3 consecutive samples (45 s), training was completed. All subjects were trained for a minimum of 2 minutes, which was sufficient for most. We kept the exercise as brief as possible in order to allow the subject to retain her memory of the prior session and to avoid fatigue.

After training, the subject was shown a video recording of the prior discussion, registered from approximately her point of view (see Figure 3-4). The subject was instructed to score whom *she* had spoken or listened to during that session. She was asked *not* to score looking behaviour, and to use her memory to score what she did *then*, not what she saw *now*. After scoring, the video was played a second time for correction purposes. For details of the scoring procedure, see page 46. Finally, to correct the timing

Figure 3-4. *Conversational partners as seen from a camera located above the subject's head.*

of scores, we measured the response time of the subject. We did this using a simple stimulus-response test (see the *Operationalization* section on page 49 for details).

MATERIALS

This section discusses the apparatus used to conduct the experiment, describing the layout of the discussion room, video and audio registration equipment, and equipment used for the registration and synchronization of data.

Experimental Setup

The experimental setup in the discussion room consisted of two tables, around which the subject and the three conversational partners were seated (see Figure 3-5). The subject was seated behind the table on the left in Figure 3-5. The eyetracker camera was placed in front of the subject, mounted on a 30 cm tripod so that it was not in the line of sight between subject and conversational partners. The conversational partners were seated around the right table. The distance between the subject and the two closest conversational partners was about 2.2 m, and the distance between conversational partners was about 1 m. The subject was seated with his head against an adjustable headrest, typically at 119 cm, with individual adjustments made of up to 3 cm. Care was taken that there were no objects to look at on any of the tables throughout the experiments.

The video projector used during calibration and subject training was placed between the two tables, tilted 5° and projecting onto a screen located behind the discussion table. The top of the projector was only 13 cm above the tables, so that it was not in the line of sight between subject and conversational partners.

The dashed lines in Figure 3-5 indicate the approximate range used for point-of-gaze measurements. The left side of the horizontal range was determined by the left side of eyetracker measurement range, the right side of the horizontal range was approximately the right side of the area of overlapping vision of subjects. Note that since we measured the right eye, the subject was positioned slightly off-axis. This was to ensure the conversational partner on the left was within eyetracking range. Looking outside the measurement range could result in the eyetracker not being able to report point of gaze.

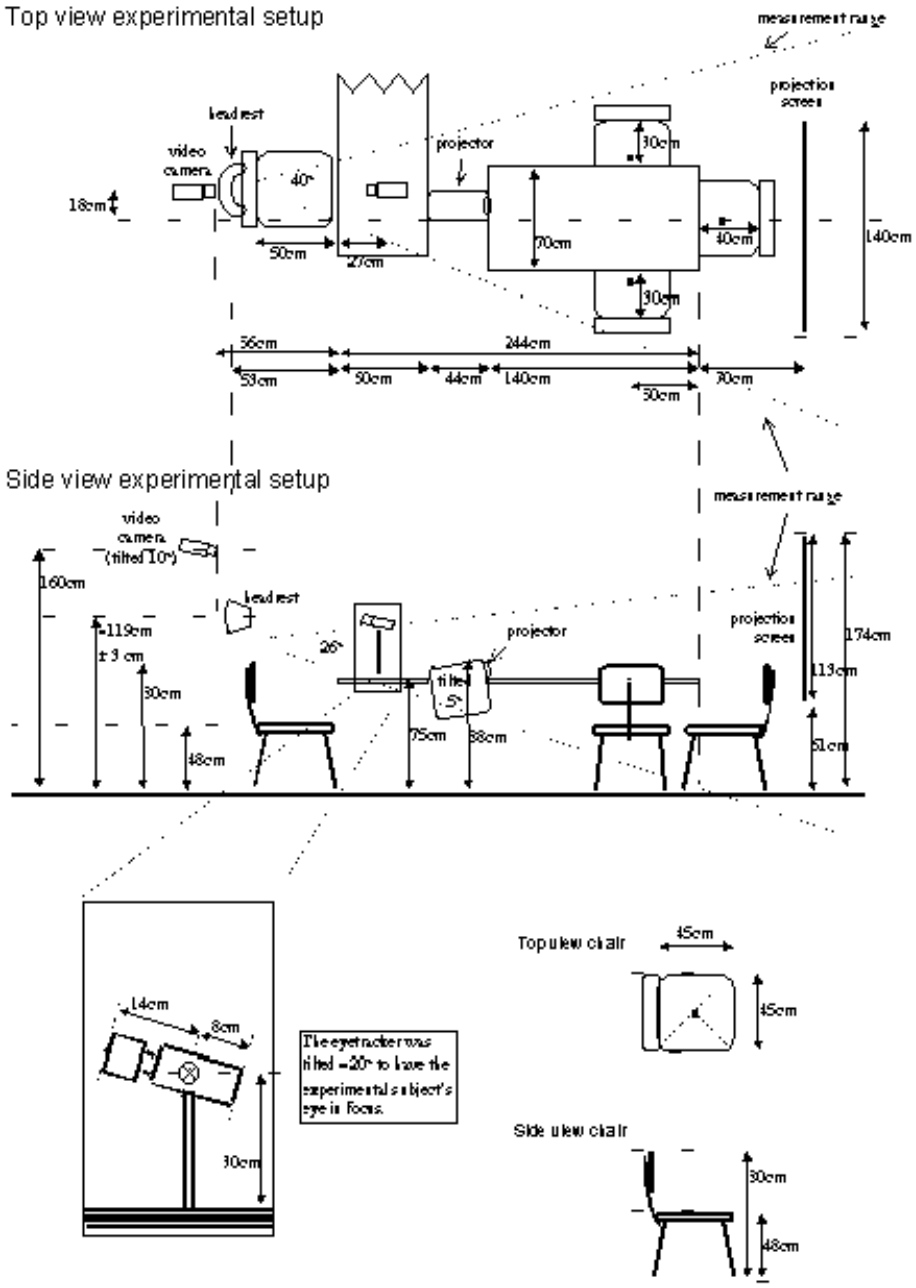


Figure 3-5*. Overview of the experimental setup.*

Video & Audio Registration Equipment

A camera with an internal microphone was placed just above and behind the subject's head (see Figure 3-5). The audio and video signals of this camera were used to register sessions for subsequent scoring of dialogic attention. The camera angle and position were so that this video registration approximated the subject's point and angle of view as closely as possible. The video and mono audio signals of the camera were fed to the control room, where they were recorded on a Hi8 recorder, with a SMPTE time code signal added on the right audio channel for data synchronization purposes.

The signal of the infrared eyetracking camera was also fed to the control room. This allowed the experimenter to monitor the correctness of the position and focus of the subject's eye, and to interrupt the session if necessary. This image was also registered using a VHS recorder, with a SMPTE time code signal added on the right audio channel for data synchronization purposes. This registration was used in the analysis phase to identify the error component in eyetracker measurement data (see the *Analysis* section on page 52 for details).

The lag of the video registration and recording equipment was negligibly small (< 1 unit of measurement).

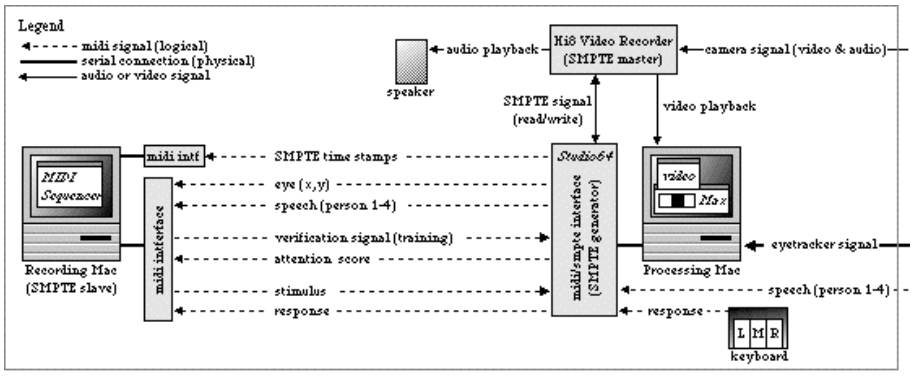


Figure 3-6. *Overview of the recording equipment with data streams.*

Data Recording Equipment

An overview of the processing and recording equipment and the rather complex streams of data between them is given in Figure 3-6. All incoming data was first processed on a PowerMac 6100 AV (located on the right in Figure 3-6). This machine ran *Max*, a real-time MIDI processing environment [114], and during subject scoring of dialogic attention it also ran video display software. After processing, all data was sent as MIDI [35] to a Macintosh IIci computer (located on the left in Figure 3-6). There, it was recorded with 40 ms (1 video frame) accuracy using Performer MIDI sequencing software [96]. This software added a SMPTE time stamp [35] to each data event for synchronization and timing purposes. Time stamps were generated by an Opcode Studio64 MIDI/SMPTE interface [111], which also generated time code signals for registration on the Hi8 video recording of the session. During subject scoring of dialogic attention, this Hi8 video recorder acted as a SMPTE master, allowing subject scores to be registered with the absolute time at the moment of session video registration. This effectively ensured synchronization of *all* measurement data.

OPERATIONALIZATION

In the next section, we will discuss what was measured to constitute the dependent and independent variables, and how these measurements took place. First, we will discuss the measurement of gaze, after which we will discuss measurement of dialogic attention and personality factors.

Operationalization of Gaze Measurements

In order to constitute a binary variable `gaze` for each conversational partner, which was *true* whenever the subject gazed at that person's facial region, and *false* when not, we needed to measure the following variables:

- *Point of Gaze*. We measured this by registering the position and duration of subject fixations throughout the discussion session using an eyetracker.
- *Center of Gravity of the Facial Regions*. We measured this by asking the subject to track the eyes of each partner, while registering the position of subject fixations with the eyetracker. Retroactively, we fitted a circle around the mean eye position of each partner. Looking within this circle would yield `gaze` for that partner.

We will first discuss point of gaze registration, then registration of center of gravity of the facial regions.

Measuring Point of Gaze

To measure subject fixations, we used the LC Technologies' Eyegaze System [91], a desk-mounted imaging eyetracker. The Eyegaze system consists of a 486 computer processing the images of a high-resolution infrared video camera. This camera unit was mounted on a tripod on the table in front of the subject (see Figure 3-5 on page 39), at approximately 58 cm from the right eye, at which it was typically aimed (see Figure 3-7).

The system implements the *Pupil-Center/Corneal-Reflection* method in the following way. On top of the lens of the infrared camera, an infrared light source is mounted which projects invisible light into the eye. This infrared light is reflected by the retina, causing a bright pupil effect (the large circle in Figure 3-7) on the camera image. The light is also reflected by the cornea of the eye, causing a small glint to appear on the camera image (the small dot in Figure 3-7). Because the cornea is approximately spherical, when the eye moves, the corneal reflection remains roughly at the same position. However, the bright pupil moves with the eye. By processing the image on the computer unit, the vector between the center of the pupil and the corneal reflections can be determined. In order to correctly translate this vector into real-world coordinates, the system needs to be calibrated. That procedure will be treated separately in the ensuing section.

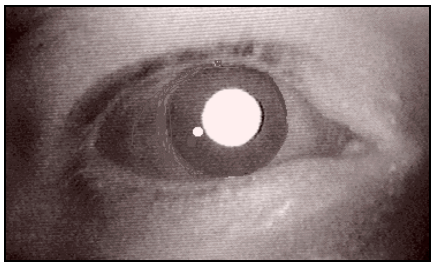


Figure 3-7. *The eyetracker infrared camera image.*

Every 120 ms, the system reported one of three kinds of events:

- *Fixation Coordinate*. If the eye had remained within a radius of 20 camera pixels for at least 120 ms (3 consecutive camera frames), it would report the average location over 3 frames as an *(x, y)* fixation coordinate.
- *Saccade*. If the eye had moved outside of the radius, the eyetracker would report a saccade event until another fixation was detected.
- *Eye Not Found*. If the system could not detect the eye (i.e., one of the reflections was missing), it would report this until another fixation or saccade was detected. See the *Analysis* section on page 52 for a discussion of the interpretation of these events.

These messages were sent via a serial link to the control room, where they were recorded with a time stamp using MIDI sequencing software [35].

Calibration and Error of Point of Gaze Measurement

In order to relate the relative eye vector coordinate to coordinates in the experimentation room, the eyetracker was calibrated with each new subject. This calibration also effectively removed measurement errors due to individual differences in eyeball geometry. Before the session, subjects were asked to fixate on nine pre-determined positions, successively projected as dots on a video screen behind the discussion table (see Figure 3-5 on page 39). After calibration, the system calculated the match between the grid of fixation points and the grid of pre-determined positions as a weighted error. The calibration procedure was repeated at least three times, until this error levelled under a value of .45°.

Spatial Accuracy

The resolution of the image projected on the screen during calibration was 640 x 480. However, at about 40° horizontally, and 26° vertically, the actual range used for measurement was much larger than this projected image. In the virtual plane at the location of the projection screen this range was 2.6 m (1258 pixels) horizontal and 1.68 m (830 pixels) vertical, relating to a spatial resolution of measurement of 4.9 pixels/cm in this plane (see Figure 3-5 on page 39). Since our subjects scored an average calibration error of .38° (SD .7°), the actual mean bias error of fixation points was 2.4 cm in this plane.

Therefore, it was theoretically possible to determine whether the subject was looking at the left or the right eye of the conversational partner located furthest away.

Temporal Accuracy

The Eyegaze system processed images at 25 full video frames per second, yielding a maximum temporal resolution of 80 ms. At 120 ms (or 3 frames), our actual temporal resolution was well within that range. This resolution corresponds to the minimum human fixation time [150]. We determined the lag of the eye registration subsystem by comparing time stamps of the infrared camera image, as recorded on video tape, with time stamps of recorded data events. These time stamps were synchronized by SMPTE time code. Thus, of a corresponding fixation, we could determine the time between the end of the previous saccade on video and the registration of the data event reporting the new fixation. We did this for a set of 22 controlled fixations by the experimenter. Frame-by-frame video analysis of these fixations yielded a mean latency of .46 s (SD .06 s), corresponding to 4 units of measurement. This lag was corrected for during analysis.

Measuring Center of Gravity of the Facial Regions

In order to establish the average location of the eye region of each conversational partner during the discussion, we registered subject fixations at these eye regions. Retroactively, we fitted the largest possible non-overlapping circles around the mean positions of these eye regions, yielding, per conversational partner, a circle within which his facial region was very likely to be located throughout the session. During analysis, fixations within such a circle would yield `gaze` for the corresponding partner.

After calibration and training, we asked the subject to track the eyes of a conversational partner, while that partner looked successively for 3 s at:

- the person at his right-hand side;
- at the person in front of him;
- at the person at his left-hand side;
- the table and ceiling.

This procedure was repeated for all three conversational partners. After each measurement, the experimenter, on his display, saw a graphic representation of the location and size of the circle which contained 95% of the measurements for that partner. If this circle had a radius of about 140 pixels for the partners on the left and right, and about 105 pixels for the partners opposite the subject, the measurement procedure was complete. If radii differed more than 40% from these values, the procedure was repeated. For each orientation of a partner's head, we measured approximately 25 fixation position samples, yielding a minimum of 100 samples per center of gravity. For a discussion of the circle

fitting procedure, and the accuracy of determining `gaze` using this method, see the *Analysis* section on page 53.

Operationalization of Dialogic Attention Measurements

In order to constitute, for each conversational partner, a binary variable `articulatory attention`, which was *true* only when the subject spoke to that person, and a binary variable `auditory attention`, which was *true* only when the subject listened to that person, we needed to measure the following variables:

- *Speech Activity*. We measured this by registering the isolated speech energy of each person in the group throughout the discussion session. Retroactively, we analyzed this data for the occurrence of *utterances*: moments of at least 1.5 second duration where one person would speak while others were silent. For each person, this analysis yielded a binary variable `utterance`, which was *true* when that person had an utterance, and *false* when not (see the *Analysis* section on page 55 for details).
- *Dialogic Attention Score*. We measured this by asking the subject to score whom he was talking or listening to during playback of a video recording of the discussion session. For each person, this yielded a binary variable `dialogic attention`, which was *true* when the subject scored he was dialogically attending to that person, and *false* when not.

For each conversational partner, the variable `auditory attention` would be constituted by the *logical and* of the variables `utterance` and `dialogic attention` for that partner. For each conversational partner, the variable `articulatory attention` would be constituted by the *logical and* of the variable `utterance` of the subject and the variable `dialogic attention` for that partner.

With auditory and articulatory attention of the subject known for each of his conversational partners, we could constitute all independent variables regarding dialogic attention retroactively (see *Calculating Results* on page 59 for details). We will now discuss how we measured speech activity and dialogic attention scores.

Measuring Speech Activity

To register individual speech activity, each person in the group wore a head-mounted microphone (see Figure 3-4 on page 37). Each microphone was connected to a sound level meter (a filtered full wave rectifier). Using a capacitor with a discharge time of 120 ms, it converted the alternating microphone signal into a direct current, outputting +5 V when speech energy was above an adjustable threshold, and 0 V when not. Prior to each session, this threshold was calibrated individually to ensure maximum signal level without crosstalk of other participants' voices. This was done by asking each person to count to five with a loud and soft voice. When crosstalk occurred, this procedure was repeated.

The state of all sound level meters was sampled every 30 ms by an I-Cube digitizer [77]. This yielded four bits, with each bit *true* if there was speech activity by that person during that period, and *false* if not. Using MIDI [35], these bits were sent to a Macintosh computer in the control room (see Figure 3-6 on page 41). At arrival of these bits, *Max* software [114] on this machine determined, for each individual, the *speech activity state* over the previous 120 ms interval. A state was set to *true* if one of the four samples of the corresponding individual had been *true* during this interval, and set to *false* if not. If the *speech activity state* of an individual changed, a MIDI message (identifying individual and state) was sent to a second Macintosh (see Figure 3-6 on page 41), where it was recorded with a time stamp using MIDI sequencing software [96].

Accuracy

We determined the accuracy of the speech sampling equipment by sending a set of 66 1-second bursts of white noise through a microphone at 2-second intervals and threshold level. Since the time at which these bursts were sent was known, we could determine the latency of the speech recording subsystem. Mean latency was .12 s (SD < .04 s), corresponding to 1 unit of measurement. This lag was corrected for during utterance analysis. Signal duration was accurate to within 1 unit of measurement. Maximum signal drop was .24 s, which occurred infrequently. This was typically corrected for during utterance analysis.

Measuring Dialog Attention Score

In order to gauge the dialogic attention of each subject during the last five minutes of his session, we replayed a video registration just after the session, asking the subject to score whom he had listened or spoken to at the time. This video was registered from the approximate point and angle of view of the subject, using a camera placed just above his head (see Figure 3-5 on page 39). The subject was seated behind a computer screen in the control room, which displayed the video image of the three conversational partners (see Figure 3-4 on page 37). Below this image, there were three indicator lights, with each light placed such that it corresponded to the location of one of the conversational partners. This setup is shown in Figure 3-8. Between the subject and the computer screen, a MIDI accord keyboard was placed with all keys covered but three. Each of these three keys corresponded spatially to one of the indicator lights below the video image. Thus, when the subject would press the left key, the indicator light below the left partner would burn, etc. Pressing multiple keys would activate all corresponding indicator lights. During scoring, each key press and release was recorded using MIDI sequencing software [96], which added the time stamp of the video signal for synchronization purposes. The subject was to press keys only during listening or speaking activity of himself. He was to press those keys that corresponded to the conversational partners he was listening or speaking to. Keys were to be pressed at the onset of his

listening or speaking activity, and released only when this activity stopped, or when his focus of dialogic attention switched to different person(s).

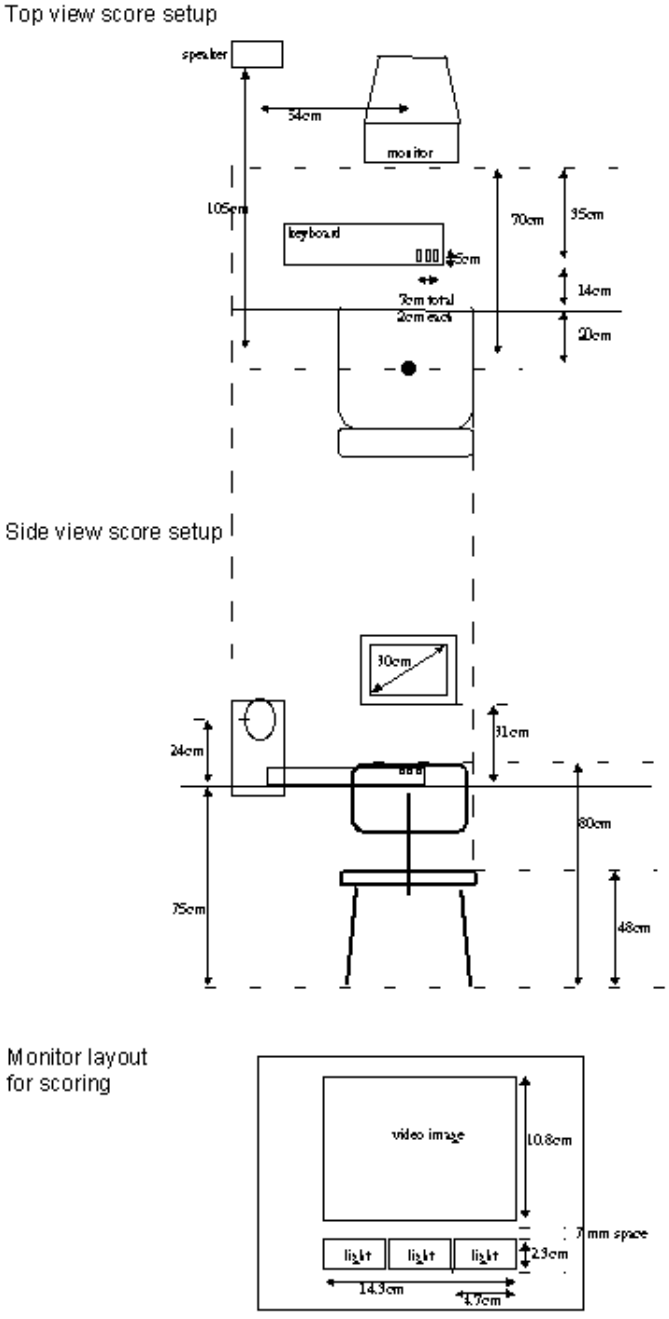


Figure 3-8. *Overview of the setup used for scoring dialogic attention.*

After scoring, the subject was asked to review the video. This time, the indicator lights below the image burned automatically, showing his previous score. The subject was asked to indicate the occurrence of scoring errors by pressing a key on the keyboard for the duration of that error. These key presses were registered along with the time stamp of the video signal, allowing us to skip these scoring errors during analysis.

Validity of Scoring

Before scoring, subjects were trained and carefully instructed on the task (see *Session Procedure* page 36 for details). During this training, subjects scored a video of a session in which four actors played the role of the subject and conversational partners, according to a script which specified the dialogic attention of each actor. Subject key presses during training were compared with key presses pre-specified by the experimenter, according to the actual enactment of this script. For each 15-second interval of training, the agreement of key presses was calculated as a percentage of time in which key presses overlapped exactly. Before being admitted to the scoring procedure, each subject had to reach a 60% agreement with the pre-specified score for at least 45 consecutive seconds.

During the subsequent scoring procedure, subjects could use three sources of information: their memory, and the internal and external contexts of conversation as provided by the video. Internal context is information provided by earlier utterances, while external context is information provided by the environment [40]. Part of the external context was the head orientation of the conversational partners. We decided not to hide this information in order to aid memory as much as possible. However, this may have resulted in a situation in which the subjects used such information to specify focus of dialogic attention, particularly during speech activity of the subject. One might argue this was problematic, as the head orientation of the conversational partners may have correlated with the visual attention of the subject. However, this argument is irrelevant, since all subjects reached a concurrent validity of at least 60% accordance with *pre-specified* focus of dialogic attention during training. Given that subjects could rely on their memory during scoring, we expect that the actual concurrent validity of scoring was better than 60% agreement.

Correcting Scores for Response Time

In order to achieve exact synchronization between dialogic attention scores and other data, we corrected each score for the response time of individual subjects. To measure this response time, we used a stimulus-response (S-R) test which was an abstraction of the original scoring task. After scoring, the video display on the computer screen (as shown in Figure 3-8), was replaced by a second set of three indicator lights, positioned right above the existing three indicators. During the S-R test, one of the upper three indicator lights would start burning at random intervals, accompanied by a clearly identifiable auditory stimulus. These auditory stimuli consisted of a 130 Hz

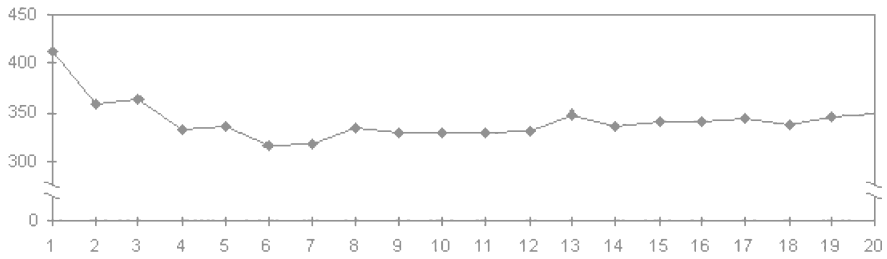


Figure 3-9. *Mean response time (mean of on- and off-times) in the stimulus-response test, per stimulus, across subjects.*

sine tone (left light); a 262 Hz sine tone (middle light); and a 523 Hz sine tone (right light). The subject was asked to place his hand on the keyboard and press the key corresponding to the audio-visual stimulus for the length of that stimulus. While doing so, he would receive visual feedback of his key presses by means of the lower set of indicator lights, as in the scoring task. Subjects were asked to respond as quickly as possible, both when pressing and releasing a key. We used a set of 25 stimuli, with only one of the three upper indicator lights, chosen randomly, burning at a time. The duration of each stimulus and the interval between stimuli were randomly chosen between 1 and 2 seconds. However, since randomization was done prior to the experiment, each subject used the same set of stimuli. Both stimuli and responses were recorded with time stamps for later analysis. We calculated two response times for each subject: the mean *on-time* needed to select and press a key; and the mean *off-time* needed to release a key. A graph of mean response time (mean of *on-* and *off-times*) across subjects is shown in Figure 3-9. To remove learning effects, we excluded the first 5 stimuli of each subject from analysis. Errors were also excluded: the average number of correct responses was 19.2 (SD 2.7), the minimum 14.

The mean *on-time* across subjects was .38 s (SD .06 s); the mean *off-time* was .25 s (SD .03 s). All *on-times* and *off-times* of subject scores were corrected for these latencies on an individual basis. Typically, *on-times* were corrected by 3 units of measurements, while *off-times* were corrected by 2 units of measurement. The latency added by the recording equipment was negligibly small (< .04 s).

Operationalization of Personality Measurements

Subjects' scores on personality factors were gauged prior to the actual experiments by means of a Dutch version of the Five Factor Personality Inventory, developed by Hendriks, Hofstee, De Raad & Angleitner in 1995 [74] and largely based on the Big-Five model of personality [34, 144]. By answering a questionnaire of 100 items, subjects were asked to evaluate their own personality. The answers were analyzed using an automated procedure available with the test. We thus obtained data on the extraversion, autonomy, mildness, orderliness and emotional stability of subjects. Of these, we only used the first two measures to constitute independent variables: extraversion and autonomy. The mean score across subjects was .06 (SD .90) for extraversion, and .72 (SD .75) for autonomy, with 0 being the national average on all parameters of this test [74]. Unlike extraversion, dominance, as discussed in our literature study on page 32, is not a parameter in the FFPI. However, recent studies show that the correlation between the autonomy parameter in the FFPI and dominance parameters as measured by more traditional tests (in this case, the Dutch Personality Inventory [94]), is sufficiently high to allow their substitution [15]. Even so, we will treat any conclusions as appertaining to autonomy, rather than dominance.

ANALYSIS

During this phase, all data was compiled in order to constitute the dependent and independent variables. In order to keep all data synchronized, we used automated analysis procedures only, checked by human observers. We will first discuss the analysis of gaze measurements, after which we will explain the analysis of dialogic attention measurements. Finally, we will discuss, for each independent variable, how we calculated the results.

Analyzing Gaze Measurements

We will first discuss how we interpreted ambiguous eyetracker measurements, after which we will explain how we determined `gaze` at the facial region.

Interpreting Measurements Where the Eyetracker Lost Track

Whenever the eyetracker lost track of the eye it suspended measurement of gaze position, reporting this by means of an *eye not found* message. Measurement of gaze position would be resumed the moment the eyetracker located the pupil and corneal reflections again. These temporary interruptions of measurements yielded ambiguity in the measurement data, as the eyetracker software could not determine the cause of interruption. We identified three likely causes for temporary interruptions of measurement:

1) *Extreme Angles of Looking.* When the angle of eye rotation becomes too large, the corneal or pupil reflection may disappear.
2) *Blinks.* When the eyelid obstructs the eyetracker image of the pupil for more than 120 ms during blinking, the eyetracker can no longer report point of gaze.
3) *Error of Measurement.* The eyetracker image of the pupil can be obstructed by reflections on glasses, incorrect positioning of the eye (outside camera aperture), rapid adjustment of head position (within camera aperture), squeezing of eyelids and other interference such as eye rubbing.

We verified the interpretation and relative proportion of *eye not found* messages by comparing video recordings of subjects' eye movements with actual measurement data. This was possible as all video and data recordings contained SMPTE time descriptors, allowing them to be synchronized after the fact. We superimposed a graphical representation of *eye not found* messages and utterance analysis results onto a video recording of the infrared eyetracking camera image. The result was recorded for subsequent frame-by frame video analysis. This way, we could determine the occurrence and length of eye not found messages, the actual movements of the eye, and whether the subject was classified as speaking or listening. We analyzed 8 minutes of conversation from 4 randomly selected sessions. 50% of the analysis data involved speech activity of subjects, and 50% listening activity of subjects. Pauses were skipped from analysis. We categorized each *eye not found* message, attributing them to one of the above three causes. We also measured the length of each message using

video frame counts. This yielded estimates of the relative contribution of each type of *eye not found* message to gaze position measurements for two cases: while listening and while speaking.

The average duration of *eye not found* messages was brief at .68 s. We estimate the total contribution of *eye not found* messages at approximately 6% of total measurement time while listening, and approximately 17% of total measurement time while speaking. In both cases, the error component was small, accounting for less than 1 percent of total measurement time, and less than 6% of the total duration of *eye not found* messages. Errors were typically due to rapid adjustment of head position within the camera aperture or squeezing of eyelids. We attribute the large increase while speaking to an increase of extreme looking behaviour with a factor 3.5, and an increase in long blinks — with an average length of .4 s — with a factor 2. Although one might argue about the meaning of long blinks when listening, the large increase when speaking in both blinking and extreme looking behaviour might suggest purposeful obstruction of eyesight, or a signalling function that may be attributed to this, to be important contributing factors. This would imply that *eye not found* measurements should not be excluded as errors of measurement, but instead be interpreted as *not* looking at a person. We made one exception to this rule: *eye not found* messages of a single unit length (120 ms) were typically caused by short blinks. When this occurred while gazing at a facial region it was treated as gaze.

The overall decision to interpret *eye not found* messages as *not* looking did not significantly affect results of hypothesis testing, since data sets excluding *eye not found* messages produced the same test results as data sets including *eye not found* messages. It also did not significantly affect our estimates for the efficacy of gaze as a predictor of dialogic attention.

Determining Gaze at the Facial Region

Per subject, we constituted a binary variable `gaze` for each of his conversational partners, which was *true* whenever the subject gazed at that partner's facial region, and *false* when not. We did this by fitting the largest possible non-overlapping circles around the mean centers of gravity of the partners' facial regions. Per unit of time, the variable `gaze` of a partner yielded *true* when a subject fixation fell within the corresponding circle.

For each subject, we first determined the mean of all center of gravity measurements for each partner, as recorded prior to the session. This yielded three *(x,y)* points in space, around which we fitted the three largest possible non-overlapping circles. The radii of these circles were calibrated for the distance at which the corresponding partner was seated, yielding

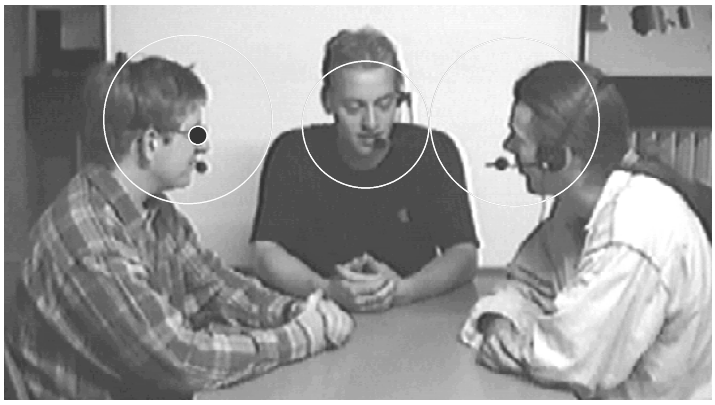


Figure 3-10. *A graphical representation of the algorithm used for determining gaze, superimposed on the actual video image. The black dot in the left circle shows the point of gaze of the subject.*

proportionate radii of 4:3:4 respectively for the left, middle, and right conversational partner. This process is shown in Figure 3-10. Per unit of measurement (120 ms) of the last five session minutes, we subsequently determined whether there was a subject fixation, and if so, whether its position was within one of these circles. If so, `gaze` for the corresponding partner would be set to *true*. If not, it was set to *false*. The example in Figure 3-10 yields *true* for the left conversational partner, since the black dot (indicating the location of a subject fixation) is within the left circle. In this process, two corrections were made:

1) When a single *eye not found* message was found in between two fixations within the same circle, the corresponding unit of time was evaluated as a fixation within that circle. This was to compensate for short blinks that triggered such messages.
2) When a *saccade* message was found in between two fixations within the same circle, the corresponding units of time were evaluated as fixations within that circle. This way, we compensated for iterative looks at the left eye, right eye and mouth of the same partner.

Validity of Gaze Analysis

The results of the analysis process were evaluated by the experimenter by superimposing the fixation position of the subject, the circles fitted around each partner, and the value of the `gaze` variable for each partner onto the real-time video image of each session, as in Figure 3-10. The experimenter could indicate an error by pressing a key for the duration of that error. These key presses were registered in sync with data, so that errors could subsequently be discarded from analysis. An error could be due to one of two reasons: the subject fixated within the circle but outside the actual facial region of a partner, or the partner had moved his facial region outside the circle. No errors were found.

Analyzing Dialogic Attention Measurements

We will first discuss how we analyzed speech activity registrations of each individual for the occurrence of utterances. We will then explain how the resulting `utterance` variable was used to determine `auditory` and `articulatory attention` of subjects.

Analyzing Speech Activity for Utterances

Per group member, we constituted a binary variable `utterance`, which was *true* whenever the person had an utterance, and *false* when not. We did this by analyzing speech activity throughout the last five session minutes by means of the automated utterance analysis algorithm outlined below. We selected the last part of the session since subjects would be more acquainted with one another, and were more likely to focus on the discussion rather than the situation.

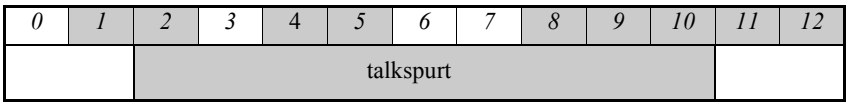
It is not evident that a registration of the energy produced during speech activity is a good indicator of what we consider to be an *utterance*. This is because throughout the articulation of speech, the speaker may introduce various moments of silence:

1) *Pauses within words.* Words are constituted by a string of basic vocalization units, called *phonemes* [141]. In between phonemes, pauses may occur. One cause for this is *stop consonants*. For example, the word *"stop"* itself consists of four phonemes: */s/ /t/ /o/* and */p/*. While the first three are typically pronounced consecutively, a small pause may occur between */o/* and */p/*. The */p/* in this example is a stop consonant. It is evident we should consider such pauses within words as belonging to an utterance. The length of such pauses is typically less than 200 ms [19].
2) *Pauses between words.* Propositions are constituted by a string of words. In between words, pauses can occur. Whether or not we should consider these as part of the utterance depends on whether the words belong together. In order to identify this, one can ascertain whether they belong to a *phonemic clause* [127]. The phonemic clause is regarded as the basic syntactic unit of speech [81]. According to Jaffe [81], it is a string of 2-10 words (typically 5, with a typical duration of about 1.5 s), in which there is one primary stress and which is terminated by a juncture, a slight slowing of speech, with slight intonation changes at the very end. This definition corresponds to Givón's definition of the oral expression of a *mental proposition*, the basic unit of mental information storage [62]. Although this is still the subject of further investigation, we will consider the phonemic clause the smallest string of words necessary for expressing a proposition. We therefore interpreted pauses within a phonemic clause as belonging to an utterance.

The Talkspurt Filter

The procedure counts the number of samples with a value true (indicated as gray boxes) in a 13-sample window (1.56 seconds) of the speech activity signal after word analysis. If this total is less than 7, the procedure does nothing and shifts one position ahead in time. If it is greater, the mean position of samples with a value true within the window is calculated. In the below example, this amounts to: (1+2+4+5+8+9+10+11+12)/9 = 6.89.

If the mean value is between 5 and 7 inclusive, it decides these samples are evenly spread over the window. It now sets all samples in the talkspurt signal between 2 and 10 to true (as indicated by the large gray box). Then, the windows shifts one position ahead in time.



Box 3-1*. A graphical illustration of the talkspurt analysis algorithm.*

3) *Pauses between phonemic clauses.* Talkspurts are constituted by a string of phonemic clauses. In between phonemic clauses, pauses can occur. Whether or not we should consider these as part of the utterance depends on whether the phonemic clauses are produced by the same speaker [123, 44]. This is based on the idea that typically, only one person is considered to be a speaker at any moment in time (the floor-holder). The process that constitutes this, *turntaking*, is effectively designed to ease the focusing of auditory attention in conversations [123, 124, 44]. However, when there are more than three participants, *side conversations* may occur, allowing subgroups to have simultaneous conversations [124, 127]. By our definition, side conversations are seen as simultaneous talkspurts, rather than utterances. This methodological problem is treated in the *Validity* section below.

We designed a fuzzy algorithm which analyzed the speech activity of all group members simultaneously, designating moments of silence and moments of utterance activity by a floor-holder using the above definitions. First, *word* analysis filled in all pauses smaller than 2 units of measurement (240 ms) to account for stop consonants [19]. Then, *talkspurt* analysis removed pauses between words, but only if those words could be considered part of a phonemic clause. To do this, the talkspurt filter moved a window with the length of the mean duration of a phonemic clause (1.56 s or 13 samples) over the speech data of each individual, shifting it one unit at a time (see Box 3-1). Samples within a 70% confidence interval around the center of the window were filled with speech energy if more than half of the samples in the window indicated speech activity, and if this speech activity was well-balanced over the window (i.e., if the mean position of samples indicating speech activity was between 5 and 7 inclusive). The 70%

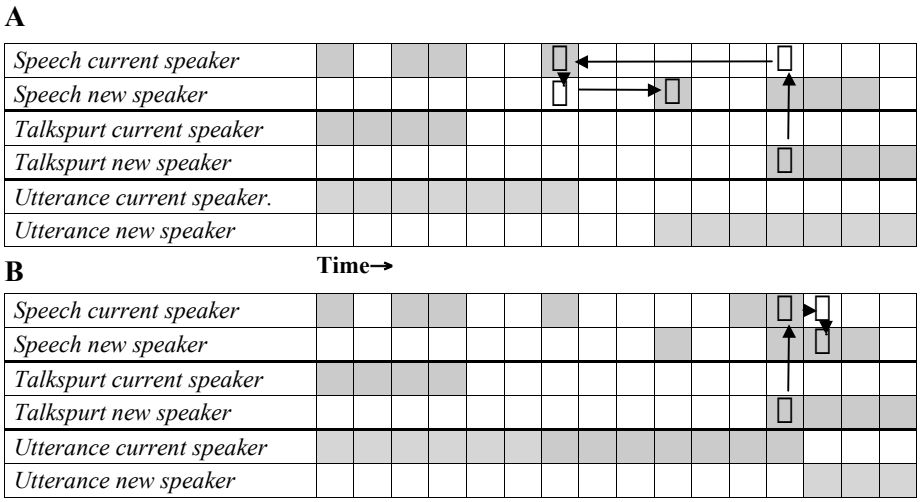


Figure 3-11. *A graphical representation of the utterance algorithm for two cases: non-overlapping speech (A) and overlapping speech (B).*

confidence interval corresponded to the start and end of the two most unbalanced talkspurts of 7 samples that could pass the balancing criterion, padding them with no more than two samples, while preserving correctly either their start or end position. Finally, if one of the speakers had a talkspurt, longer than a phonemic clause (i.e., 1.56 s or 13 samples), with everybody else not having any talkspurts during that time, an *utterance* was assigned to her. This utterance would end with a speaker switch*: i.e., when a new speaker had a talkspurt, longer than a phonemic clause, while everybody else (including the previous speaker) was not having a talkspurt during that time.

Figure 3-11 illustrates how the exact location of utterance boundaries was determined, showing speech activity assignments for the previous and new speakers after *word* analysis, *talkspurt* analysis and *utterance* analysis. In order to predict these boundaries more exactly, we used the actual speech activity registrations (after *word* analysis) for this. Figure 3-11A shows how boundaries were located when there was a silence before the speaker switch. When a new utterance was found, the algorithm first located the start of the *talkspurt* in which this new utterance was located (1). Then, it would look back in time (2) to find speech activity of the previous speaker in the *word* analysis signal (3). This was designated the end of the previous utterance. It would then look forward in time again (4) to locate the first speech activity after this position by the new speaker in the *word* analysis signal (5). This was designated the start of the new utterance. Figure 3-11B shows how boundaries were located when there was overlapping speech during the speaker switch. Again, the algorithm

* Note that in this definition an utterance differs from a turn in that it does not include the pause that may occur before a speaker switch.

first located the start of the *talkspurt* in which the new utterance was located (1). Then, it would look forward in time (6) to find a pause of the previous speaker in the *word* analysis signal (7). The sample before this pause was designated the end of the previous utterance. From this position, it would then look forward in time again to locate the first speech activity by the new speaker in the *word* analysis signal (8). This was designated the start of the new utterance.

Note that the utterance algorithm is *conservative*. A speaker will retain the utterance, and thus the floor, during any overlapping speech activity, unless he falls silent. The algorithm thus effectively filters out any unsuccessful interruptions and back channel responses [40].

Validity of Utterance Analysis

The results of the analysis process were evaluated by the experimenter by superimposing the value of the `utterance` variable for each partner onto the real-time video image of each session. The experimenter could indicate an error by pressing a key for the duration of that error. These key presses were registered in sync with data, so that errors could subsequently be discarded from analysis. Errors were typically due to speakers retaining the floor during joint laughter or side conversations. Of each session, an average of two 5-second periods were skipped from analysis as a result of this review process (3.3% of session time).

We also checked the concurrent validity of the utterance analysis algorithm (albeit within a triadic mediated communication setting). We did this by calculating the correlation between a turn classification (with a turn defined as an utterance including the pause before a speaker switch) produced by the algorithm, and that produced by a trained linguist. This yielded a significant concurrent validity of .64 ($p<.001$, 2-tailed). For details of this procedure see Chapter 4, page 102.

Determining Auditory and Articulatory Attention

For each conversational partner, the variable `auditory attention` was constituted by calculating, for each moment in time, the *logical and* of the variables `utterance` and `dialogic attention` score for that partner. For each conversational partner, the variable `articulatory attention` was constituted by calculating, for each moment in time, the *logical and* of the variable `utterance` of the subject and `dialogic attention` score for that partner. These variables were calculated only for the last 5 minutes of each session, and only for the subset of data where subjects agreed with their dialogic attention score and experimenters agreed with results from the utterance analysis algorithm.

Calculating Results

Results were calculated over moments of time in which the following statements were true:

- Data was from the last 5 minutes of the session;
- The subject's eye was not moving;
- The experimenters agreed with results for the `gaze` variable;
- The subject had scored `dialogic attention` for someone;
- The subject agreed with this `dialogic attention` score;
- A group member had an `utterance`;
- Experimenters agreed with this `utterance`.

Results based on less than 50 samples were omitted. 50 samples corresponds to 2% of the total number of samples in the 5-minute session. This 2% threshold was non-critical for results: a slightly higher or lower percentage did not influence results significantly.

We will now summarize how results were calculated with regard to each independent variable.

Gaze and Focus of Dialogic Attention

H1. In order to validate Hypothesis 1, stating that on average, subjects spend significantly more time gazing at the individual on which their dialogic attention is focused, than at others, we compared the mean percentage of gaze at the individual on which dialogic attention was focused with the mean percentage of gaze at others. We did this separately for auditory and articulatory attention by determining the following statistics:

1) *Gaze at individual listened to.* We calculated this mean percentage by dividing, per conversational partner, the amount of time in which `gaze` was *true* for this partner while `auditory attention` was *true* for this partner by the amount of time that `auditory attention` was *true* for this partner. This yielded percentages of gaze at the partner listened to *left, opposite* and *right* of the subject. Since these percentages did not differ significantly between conversational partners ($F(2, 66)=.257$, $p>.05$), we averaged them into a single mean per subject. We then took the mean across subjects.
2) *Gaze at others than individual listened to*. We calculated this mean percentage by dividing, per conversational partner, the amount of time in which `gaze` was *true* for another partner while `auditory attention` was *true* for this partner by the amount of time that `auditory attention` was *true* for this partner. We averaged percentages over partners into a single mean per subject. We then took the mean across subjects.
3) *Gaze at addressed individual.* We calculated this mean percentage by dividing, per conversational partner, the amount of time in which `gaze` was *true* for this partner while `articulatory attention` was *true* for this partner by the amount of time that `articulatory attention` was *true*

for this partner. This yielded percentages of gaze at the addressed partner *left, opposite* and *right* of the subject. Since these percentages did not differ significantly between conversational partners ($F(2, 53)=.075$, $p>.05$), we averaged them into a single mean per subject. We then took the mean across subjects.

4) *Gaze at others than addressed individual.* We calculated this mean percentage by dividing, per conversational partner, the amount of time in which `gaze` was *true* for another partner while `articulatory attention` was *true* for this partner by the amount of time that `articulatory attention` was *true* for this partner. We averaged percentages over partners into a single mean per subject. We then took the mean across subjects.

Gaze and Mode of Dialogic Attention

H2. In order to validate Hypothesis 2, stating that on average, subjects spend significantly less time gazing at an individual they speak to, than at an individual they listen to, we compared the mean percentage of *gaze at individual listened to* with the mean percentage of *gaze at addressed individual*.

Gaze and Extent of Articulatory Attention

H3. In order to validate Hypothesis 3, stating that on average, the time subjects spend gazing at an individual when addressing a group of three is more than one third of the time spent gazing when addressing a single individual, we compared the mean percentage of gaze at an individual in an addressed triad with the mean percentage of *gaze at addressed individual*, divided by three. We did this by determining the following statistics:

5) *Gaze at individual in addressed triad*. We calculated this mean percentage by dividing, per conversational partner, the amount of time in which `gaze` was *true* for this partner while `articulatory attention` was *true* for all partners by the amount of time that `articulatory attention` was *true* for all partners. We averaged percentages over partners into a single mean per subject. We then took the mean across subjects.

6) We divided, for each subject, the mean percentage of *gaze at addressed individual* by 3. We then took the mean across subjects.

H4. In order to validate Hypothesis 4, stating that on average, subjects spend significantly more time gazing at an individual when addressing a group of three, than at others when addressing a single individual, we compared the mean percentage of *gaze at individual in addressed triad* with the mean percentage of *gaze at others than addressed individual*.

Gaze and Location of Attended Person

H5. In order to validate Hypothesis 5, stating that the relative spatial location of a person on which dialogic attention is focused does not significantly influence the time spent looking at that person, we compared mean percentages, across subjects, of *gaze at individual listened to left, opposite* and *right* of subjects. We also compared mean percentages, across subjects, of *gaze at addressed individual left, opposite* and *right* of subjects.

Gaze and Personality Factors

H6. In order to validate Hypothesis 6, stating that on average, subjects with a higher score on extraversion spend significantly more time gazing at the individual on which their dialogic attention is focused, we performed a one-tailed evaluation of the correlation across subjects between extraversion score and the following percentages of subject gaze:

- *gaze at individual listened to;*
- *gaze at addressed individual.*

In order to gain insight into the nature of these relationships, we also performed a two-tailed evaluation of the correlation across subjects between extraversion score and:

- *gaze at others than individual listened to;*
- *gaze at others than addressed individual.*

H7. In order to validate Hypothesis 7, stating that on average, subjects with a higher score on autonomy spend significantly less time gazing at the individual on which their dialogic attention is focused, we evaluated the same correlations as above, but with autonomy scores.

Variable	Amount of Gaze and Dialogic Attention *mean (s.e.)*		
	Listening to individual x	*Addressing individual* x	*Addressing triad*
Amount of gaze at individual x *(% time)*	62.4 (3.8)	39.7 (4.7)	19.7 (1.8)
Amount of gaze at others (% time)	8.5 (1.2)	11.9 (2.4)	
Amount of gaze at triad (% time)			59.0 (5.4)

Table 3-1. *Means and standard errors for the percentage of time spent gazing at partners during dialogic attention of subjects in the last 5 session minutes.*

RESULTS

The results were calculated over 24 sessions, distributed over 7 experiments. Only the last 5 minutes of each session were analyzed. All data was normally distributed (Kolmogorov-Smirnov test, $p>.05$). Planned comparisons were carried out using 1-tailed paired t-tests, evaluated at $\alpha=.05$ level. Pearson's correlation coefficients were evaluated 1-tailed at $\alpha=.05$ level, except where indicated. Hypothesis 5 was tested using a one-way analysis of variance evaluated two-tailed at the $\alpha=.05$ level, with equal variances between conditions (Levene test for Homogeneity of Variance, $p>.05$).

Gaze and Dialogic Attention

Table 3-1 shows the data summary for the mean percentage of time spent gazing at partners during different forms of dialogic attention of subjects in the last five session minutes (see also Figure 3-12 on page 64). We will now discuss the results of planned comparisons per independent variable of dialogic attention.

Gaze and Focus of Dialogic Attention

Planned comparison showed that subjects gazed approximately 7.3 times more at the individual listened to (62.4%), than at others (8.5%) ($t(23)=12.92$, $p<.001$, 1-tailed), thus confirming Hypothesis 1a. They gazed approximately 3.3 times more at the addressed individual (39.7%), than at others (11.9%) ($t(23)=5.2$, $p<.001$, 1-tailed), thus confirming Hypothesis 1b.

	Amount of Gaze and Location *mean (s.e.)*	
Variable	*Listening to individual* x	*Addressing individual* x
Amount of gaze when x = left *partner (% time)*	61.6 (4.6)	36.8 (8.5)
Amount of gaze when x = opposite *partner (% time)*	60.1 (4.6)	37.9 (8.5)
Amount of gaze when x = right *partner (% time)*	64.6 (3.7)	40.1 (5.7)
Results	not sign.	not sign.

Table 3-2. *Means and standard errors for the percentage of time spent gazing at the attended individual during the last five session minutes, for different locations of that individual.*

Gaze and Mode of Dialogic Attention

Planned comparison showed that subjects gazed approximately 1.6 times more at an individual listened to (62.4%), than at an addressed individual (39.7%) ($t(23)=5.49$, $p<.001$, 1-tailed), thus confirming Hypothesis 2.

Gaze and Extent of Articulatory Attention

Planned comparison showed that time spent gazing at an individual when addressing a triad (19.7%) was approximately 1.5 more than one third of time spent gazing at a single addressed individual (13.2%) ($t(22)=-4.47$, $p<.001$, 1-tailed), thus confirming Hypothesis 3.

Planned comparison showed that subjects gazed approximately 1.7 times more at an individual when addressing a triad (19.7%), than at others when addressing a single individual (11.9%) ($t(22)=2.71$, $p<.01$, 1-tailed), thus confirming Hypothesis 4.

Gaze and Location of Attended Person

Table 3-2 shows the data summary for the mean percentage of time spent gazing at the attended individual during the last five session minutes, for different locations of that individual.

Analysis of variance showed no significant differences in gaze at the individual listened to between locations of that individual ($F(2, 66)=.257$, $p>.05$), thus confirming Hypothesis 5a. There were also no significant differences in gaze at the addressed individual between locations of that individual ($F(2, 53)=.075$, $p>.05$), thus confirming Hypothesis 5b.

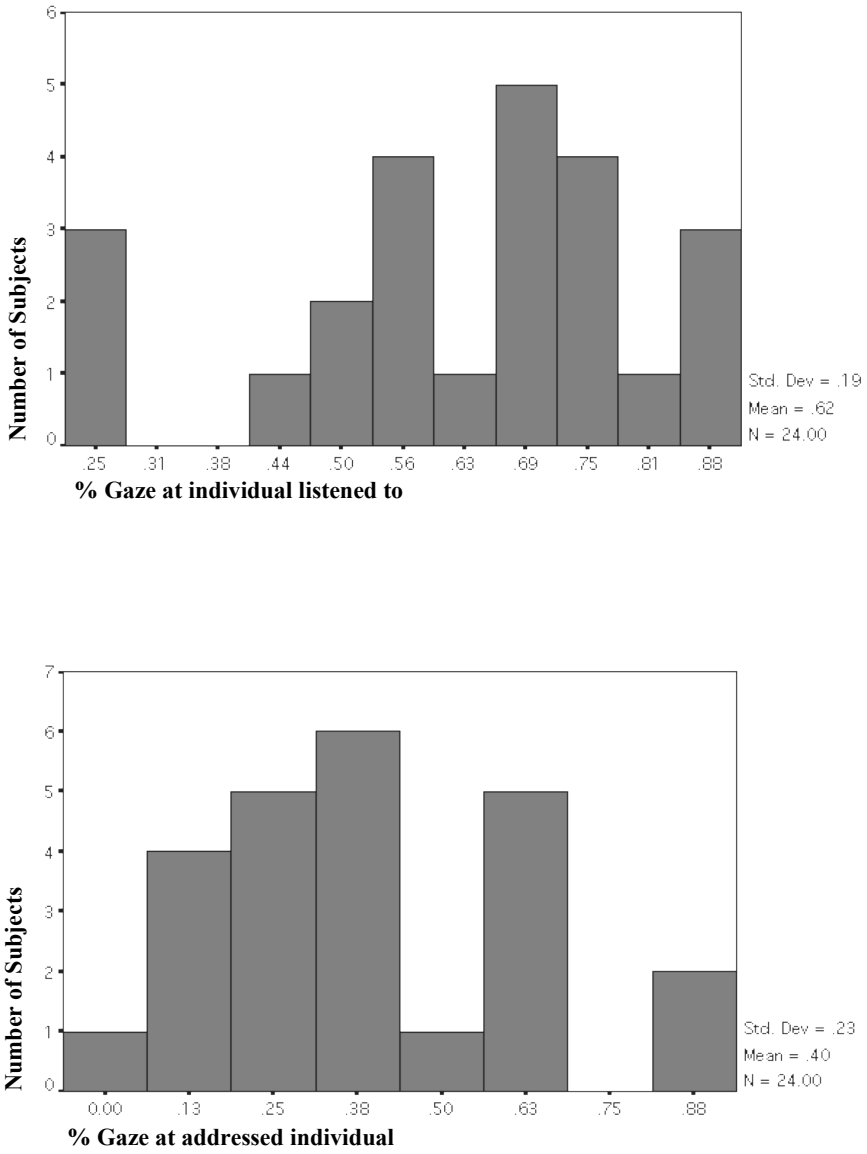


Figure 3-12. *Differences between subjects in the percentage of time spent gazing at individuals when listening to or addressing that individual were considerable.*

	Correlation between Gaze and Personality ρ	
Variable	*Extraversion Score*	*Autonomy Score*
Gaze at individual listened to	.26 not sign.	**-.48** p<.01
Gaze at addressed individual	.15 not sign.	**-.35** p<.05
Gaze at others than individual listened to	**-.45** p<.05 *	-.19 not sign.*
Gaze at others than addressed individual	.06 not sign.*	-.19 not sign.*

Table 3-3. *Correlation between personality of subjects and the mean percentage of time spent gazing at attended or unattended individual(s) during the last five session minutes. Significant correlations are printed in boldface.*

Gaze and Individual Differences: Personality Factors

Figure 3-12 shows there were considerable differences between subjects in the time spent gazing at attended individuals, with standard deviations of 19% while listening and 23% while speaking. Multiple regression showed that about 34% (multiple r^2) of variance in *gaze at individual listened to* can be attributed to the combined effect of extraversion and autonomy of subjects. Multiple regression of personality variables was not significant for *gaze at addressed individual*.

Table 3-3 shows the data summary for the correlation between personality of subjects and the mean percentage of time spent gazing at attended or unattended individual(s) during the last five session minutes. We will now discuss the results of correlations per independent variable of personality: extraversion and autonomy.

Gaze and Extraversion

Results show no significant positive relation between extraversion of subjects and the percentage of time spent gazing at the individual listened to, thus rejecting Hypothesis 6a. There was also no significant positive relation between extraversion of subjects and the percentage of time spent gazing at the addressed individual, thus rejecting Hypothesis 6b. However, there was a moderate, but significant, *negative* relationship between extraversion of subjects and the percentage of time spent gazing at others than the individual listened to.

* 2-tailed significance.

Gaze and Autonomy

Results show a moderate, but significant, *negative* relation between autonomy of subjects and the percentage of time spent gazing at the individual listened to, thus confirming Hypothesis 7a. There was also a moderate, but significant, *negative* relationship between autonomy of subjects and the percentage of time spent gazing at the addressed individual, thus confirming Hypothesis 7b.

DISCUSSION

In this section, possible explanations for our findings will be discussed for each of the independent variables. First, we will discuss results regarding the relation between gaze and dialogic attention of subjects. Then, we will discuss situational effects on gaze behaviour, focusing on effects of the personality variables extraversion and autonomy.

Gaze and Dialogic Attention

All findings with regard to the relation between gaze and dialogic attention were according to expectations. For each of the dialogic attention variables, we will now discuss our results, providing potential explanations based on literature.

Gaze and Focus of Dialogic Attention

H1. Results regarding gaze and the focus of dialogic attention were according to expectations. Hypothesis 1 was confirmed stating that on average, subjects spend significantly more time gazing at the individual on which their dialogic attention is focused, than at others. This was confirmed for both auditory (1a) and articulatory (1b) modes of dialogic attention. Given a four-person face-to-face communication setting similar to ours, gaze seems to be an excellent predictor of dialogic attention towards individuals. When someone is listening to an individual, there is an 88% chance (≈7:1 ratio) that the person gazed at is the person listened to. When someone is addressing a single individual, there is a 77% chance (≈3:1 ratio) that the person gazed at is the addressed individual. Although this experiment was not designed to ascertain whether a representation of visual attention in the form of gaze is *actually used* as a cue for predicting the focus of dialogic attention of others, results show it would function well as such.

However, such predictive power should be regarded as applying to pure communication situations only. If there are things that need to be looked at other than faces, people will do so [10]. Even so, faces can be considered a powerful attractor of visual attention. Qualitative observations of eye fixations superimposed onto video recordings of sessions showed that even during periods of heavy gesticulation by an attended conversational partner, subjects would typically produce only a few fixations at the hands of that partner, focusing on the facial region instead.

Gaze and Mode of Dialogic Attention

H2. Results regarding gaze and the mode of dialogic attention were according to expectations. Hypothesis 2 was confirmed stating that on average, subjects spend significantly less time gazing at an individual they speak to (40%), than at an individual they listen to (62%). Indeed, it seems that gaze behaviour during discussions *between individuals* in a multiparty group essentially reflects dyadic behaviour. Our results match those found by Nielsen [106] and are in

line with percentages found by Argyle and Ingham [11] (41% gaze while speaking; 75% gaze while listening), both dyadic studies.

Gaze seems to be a slightly less effective predictor of articulatory attention than of auditory attention. Our investigation of eyetracker *eye not found* messages suggests that a considerable part of the difference in gaze behaviour between listening and speaking to individuals is due to extreme looking and increased blinking activity. Although the latter may simply be explained by a higher arousal state of speakers, there does seem to be a pattern of purposeful obstruction of gaze at the listener during dyadic speech activity. As discussed in the introduction, we have a number of, possibly complementary, explanations for this:

1) *Visual Feedback.* Video observations revealed a tendency to gaze at the lips when listening to a partner with a soft voice seated opposite the subject (furthest away). Visual input in the form of lip movements and facial expressions might be less of a requirement during preparation and articulation of utterances, as such activity might be regarded as more output- than input-oriented. Consequently, there might be a lesser need for speakers to gaze during their utterances.
2) *Orientation of Thought*. Kendon [85] and Argyle & Cook [8] suggested gaze aversion while speaking is due to a need to focus attention on organization of utterances. Speaker would look away in order to reduce interference of visual input with the preparation of utterances. The tendency to blink more while speaking may also be partly attributed to this.
3) *Boundary Floor Control.* According to Kendon [85], gaze aversion during speech activity is synchronized with utterance boundaries. He showed that a speaker tends to look away when beginning an utterance, and starts producing sustained gazes at her interlocutor at the end of her utterance, possibly inviting him to take the floor. However, although gaze might be used to communicate dialogic attention, suggesting *whom* should take the floor [78], the question is to what extent the mere indication of utterance boundaries as such is not simply a derivative of function 2 (orientation of thought). Gaze avoidance during utterance preparation could indeed be regarded by listeners as a signal that the speaker has not finished her utterance. However, as we have seen in the previous chapter, utterance boundary information seems redundantly coded by many cues. According to Argyle [6], gaze functions only a minor full-stop signal for turn synchronization, with verbal and paralinguistic cues being more important.
4) *Regulating Equilibrium of Intimacy*. Speakers might be in a position where it is necessary for them to lower the level of intimacy by avoiding gaze. As this function might interact with extent of articulatory attention, we will address it in the next section.

Gaze and Extent of Articulatory Attention

H3. Results regarding gaze and the extent of articulatory attention were according to expectations. Hypothesis 3 was confirmed stating that on average, the time subjects spend gazing at an individual when addressing a group of three (20%) is more than one third of the time spent gazing when addressing a single individual (40% / 3 ≈ 13%). In fact, the total amount of gaze rises dramatically when addressing a triad, to about the same level (59%) as gaze while listening (62%). As discussed in the introduction, we have a number of, possibly complementary, explanations for this effect.

1) *Visual Feedback.* Argyle et al. [12] provided evidence that visual feedback is a predominant reason, if not the most important reason, to gaze at an interlocutor. Since, when addressing triads, there is less time to collect information on the nonverbal responses of each addressed individual, speakers might gaze more in order to satisfy their visual feedback requirements. This, however, does not preclude the existence of other functions. Indeed, Argyle et al. [12] showed that interactors who cannot see each other still attempt to send nonverbal visual information by accompanying glances at the presumed location of their interlocutor with nodding, raising eyebrows, etc.
2) *Regulation of Intimacy.* By avoiding or seeking gaze, speakers may well attempt to keep the arousal state of themselves and the addressed individuals at a mutually satisfactory level [9]. Since, when addressing triads, there is less time for speakers to gaze at each addressed individual, speakers would need to gaze more in order to maintain as satisfactory a level of intimacy with their audience as possible.
3) *Communicating Dialogic Attention.* Since in dyadic situations, it is generally clear whom the speaker is addressing, the communication of articulatory attention signals is typically not cited as a reason for gazing at interlocutors. In a multiparty situation, however, a speaker could be addressing any subset of individuals. As we have seen, gaze would seem quite effective a signal for indicating the focus of articulatory attention towards individuals. Since, when addressing triads, there is less time for speakers to signal their articulatory attention to each individual, they need to gaze more in order to ensure all individuals are aware they are being addressed (Conversational Awareness, see next chapters). This reason was cited by Kendon [85] as an explanation for Weisbrod's findings of increased speaker gaze in larger groups [160].

If we compare the above set of explanations with those in the previous section, we see that *orientation of thought* and *boundary floor control* functions, which should cause gaze avoidant behaviour, do not seem to be operative when the extent articulatory attention is large. If gaze avoidance would be *required* whenever focusing attention on the organization of utterances, or in order to indicate utterance boundaries, one would not expect to find very similar percentages of overall gaze between listening (62%) and speech activity (59%

with triadic extent). Our findings suggest that if gaze avoidance by speakers in dyadic settings is triggered by, e.g., orientation of thought, such triggering is either overridden by the above functions, or subject to an interaction effect with the above functions. We believe that a likely candidate for a potential interaction effect is *regulation of intimacy*. If Argyle and Cook [8] are correct, than gaze avoidance due to orientation of thought is caused by a need to reduce interference of visual input with the preparation of utterances. Although there may be many sources of visual input which could cause such interference, prolonged speaker-oriented gaze by a listener, particularly prolonged mutual gaze, seem excellent candidates. Both sources of visual information may cause a direct heightening of speaker arousal [8]. It seems evident that overarousal may prevent a successful allocation of attentive resources to the preparation of utterances [164]. By looking away, speakers can easily avoid prolonged mutual gaze with a listener. However, when the extent of articulatory attention is large, such behaviour would not be necessary. Instead, speakers can produce alternating fixations on different listeners, thus increasing the already low level of intimacy with each listener without becoming distracted by prolonged mutual gaze. Thus, when the extent of articulatory attention is large, we should see an alternating pattern of speaker gaze distributed both in time and space across listeners. Qualitative observations of eye fixations and articulatory attention scores superimposed onto video recordings of sessions indeed suggest the existence of such a pattern. Speakers typically seem to fixate alternatively on each addressee, using an iterative pattern of relatively short fixations. One might believe the *visual feedback* function of gaze as such would not *require* such alternating patterns of fixations. For visual feedback, it would seem far more efficient to fixate only on those addressees providing meaningful visual responses. However, we cannot rule out that alternating fixations are used as a detailed check of visual responses of individuals to the utterances of the speaker.

The above discussion does not explain why gaze avoidance also occurs in *multiparty* conversations, when the extent of articulatory attention is small (i.e., aimed at a single listener). This is where function 3, the communication of articulatory attention, might play a role. By averting their gaze when speaking to an individual, rather than projecting it iteratively at individuals outside their extent of articulatory attention, speakers may avoid giving those individuals the impression that they are being addressed. The above discussion also implies that the *boundary floor control* function of gaze avoidance is merely an output derivative of orientation of thought. As such, the occurrence of gaze avoidance during hesitant speech has come to signal an intent to hold the floor [85]. When the extent of articulatory attention is large, the need to avoid gaze for orientation of thought evaporates, and with it, its indication of utterance

boundaries*. Indeed, Argyle and Cook [8] already suggested other forms of coding of utterance boundaries are more important. We realize the above discussion is of course rather tentative. We therefore recommend further investigation of this matter.

H4. Results regarding the efficacy of gaze as a predictor of articulatory attention towards individuals in a triadic extent were according to expectations. Hypothesis 4 was confirmed stating that on average, subjects spend significantly more time gazing at an individual when addressing a group of three (20%), than at others when addressing a single individual (12%). Since speakers indeed seem to gaze more with larger extents of articulatory attention, the amount of gaze received by each individual might still be sufficient to allow addressees to discriminate it as a signal of articulatory attention. We note that this is no evidence that gaze is actually used as such.

Gaze and Location of Attended Person

H5. Results regarding gaze and the location of attended person were according to expectations. Hypothesis 5 was confirmed, stating that the relative spatial location of a person on which dialogic attention is focused does not significantly influence the time spent looking at that person. This was confirmed for both the auditory (5a) and articulatory (5b) modes of dialogic attention. This effectively allowed us to average gaze measurements across locations of conversational partners. It also suggests that subjects' use of eye movements, rather than a combination of eye *and* head movements, did not confound the experiment (at least in terms of subject observation).

Gaze and Individual Differences: Personality Factors

We will now discuss results regarding gaze and individual differences, with a special focus on personality effects. We will first discuss the relation between gaze and the extraversion and autonomy scores of subjects, after which we will discuss the impact of situatedness on gaze as a predictor of dialogic attention.

Gaze and Extraversion

H6. Results regarding gaze and the extraversion of the onlooker were contrary to expectations. Hypothesis 6 was rejected, stating that on average, subjects with a higher score on extraversion spend significantly more time gazing at the individual on which their dialogic attention is focused. It was rejected for both auditory (6a) and articulatory (6b) modes of dialogic attention. Our findings might be explained by a potential negative relation between frequency and duration of the fixations constituting gaze, discussed by Argyle and Cook [8]. However, introverts did seem to gaze more at *others* than the individual listened

* We should note that we separate the potential function of gaze to indicate utterance boundaries from its potential function to indicate dialogic attention. As such, we see the ability to indicate *whom* is addressed or expected to speak as a communication of the *mode* and *focus of dialogic attention*, rather than as a communication of *boundary floor control*.

to. This finding was confirmed by qualitative observations of fixation data superimposed onto video recordings of sessions. Introverts demonstrated an interesting tendency to observe others than the attended speaker. The theory of Argyle and Cook [8] that extraversion is related to the amount of cortical arousal and affiliative need might provide an explanation. By gazing at listeners, rather than speakers, introverts would aim to reduce the chance of mutual gaze occurring, while not giving the impression that they are not attending. They would thus lower their state of arousal. We recommend further investigation of this matter.

Gaze and Autonomy

H7. Results regarding gaze and the autonomy of the onlooker were according to expectations. Hypothesis 7 was confirmed, stating that on average, subjects with a higher score on autonomy spend significantly less time gazing at the individual on which their dialogic attention is focused. This was confirmed for both auditory (7a) and articulatory (7b) modes of dialogic attention. These results are in line with earlier findings, and consistent with the theory that autonomous individuals communicate their independence by not following the rules of maintaining equilibrium of intimacy, typically by denying others their gaze. However, our results provide no conclusive evidence for this theory. Further investigation is therefore recommended.

Situatedness of Gaze as a Predictor of Dialogic Attention

With regard to individual differences in general, we found considerable variance in the percentage of gaze during dialogic activity, with standard deviations of 19% when listening and 23% when speaking to individuals. However, distributions of gaze across subjects were very close to normality, suggesting that mean percentage of gaze *can* be considered a meaningful indicator. Personality factors seemed to have a larger impact on gaze behaviour during listening, than during speaking activity. The combined effect of extraversion and autonomy of subjects accounted for about 34% of variance in gaze at the individual listened to, which may be regarded as considerable. Autonomy alone accounted for about 12% of the variance in gaze at the addressed individual. As to the effect of individual differences on gaze as a predictor of dialogic attention, extraversion alone accounted for approximately 20% of the variance in gaze at others than the individual listened to. All in all, it seems our mean estimates of the efficacy of gaze as a predictor of dialogic attention cannot be considered free of personality and other individual effects. In addition, we realize that our findings are not free of environmental influences either. Depending on the task situation, gaze may loose much of its power as a predictor of dialogic attention. Argyle and Graham [10] found that if a pair of subjects was asked to plan a European holiday and there was a map of Europe in between them, the amount of gaze dropped from 77 percent to 6.4 percent. 82 percent of the time was spent looking at the map, but only if that map was relevant for the task situation. Argyle and Graham suggested that subjects were

keeping in touch by looking at and pointing to the same object, instead of looking at each other. As such, there is evidence that rather than evaporating, a predictive function of gaze might be transformed into a more generic indicator of joint interest when the task situation requires this [116, 151].

CONCLUSIONS

In this chapter, we investigated how well the gaze of others — their visual attention — might predict whom they are speaking or listening to — their dialogic attention — in multiparty conversations. We examined this by measuring the amount of time subjects spent looking at the facial region of conversational partners spoken or listened to during four-way face-to-face discussions. We compared those findings with the amount of time subjects spent looking at others than the individual listened or spoken to. On average, when listening, subjects gazed at the facial region of the speaker about 62% of time, and at the facial region of others only 9% of time. When speaking to an individual, subjects gazed at the facial region of that individual about 40% of time, and at the facial region of others only 12% of time. Thus, gaze indeed seems an excellent predictor of dialogic attention towards individuals in multiparty conversations. When someone is listening to an individual, there is an 88% chance that the person gazed at is the person listened to. When someone is addressing a single individual, there is a 77% chance that the person gazed at is the addressed individual. Gaze seems a somewhat more effective predictor of auditory attention (indicating whom one listens to) than of articulatory attention (indicating whom one speaks to).

When a speaker addresses more than a single individual, it seems likely that the potential predictive function of gaze is preserved. When addressing a triad, speaker gaze typically seems to be distributed evenly over all listeners. However, the total amount of speaker gaze rises significantly to about 59% of time. In such situations, the amount of gaze received by individual listeners (20%) is therefore still significantly more than the amount of gaze they would have received when not addressed (12%).

The study presented in this chapter does not provide evidence that gaze *is* in fact used as a predictor of dialogic attention. In addition, we should note that our estimates of its predictor function can only be generalized to pure communication situations. If a task situation requires visual attention for things other than faces, the amount of gaze may drop significantly. However, there is evidence that in such situations, the potential predictor function of a representation of visual attention might be transformed into a more generic indicator of joint interest. Our estimates are also subject to considerable individual differences. Our investigation of these differences focused on the effect of personality factors extraversion and autonomy on gaze behaviour of subjects. We found that introverts have a tendency to gaze relatively more than extraverts at other individuals than the speaker listened to. We also found that autonomous individuals tend to look relatively less at the individual in the focus of their dialogic attention. The combined effect of these personality variables on gaze behaviour seems greater while listening than while speaking to individuals. When listening, they may account for as much as 34% of the variance in gaze at the speaker. When speaking to an individual, autonomy alone accounted for about 12% of the variance in gaze at that individual. This

might indicate that in practice, gaze is as effective a predictor of articulatory attention as of auditory attention. However, extraversion alone also accounted for about 20% of the variance in gaze at others than the speaker listened to. This means that, for certain individuals, gaze may be a considerably less effective predictor of their dialogic attention than our mean efficacy estimates suggest.

Chapter 4

Effects of Representing Visual Attention in Multiparty Mediated Communication

ABSTRACT
We investigated whether the presence of a representation of visual attention — in the form of gaze directional cues — would have an isolated effect on multiparty mediated communication, relative to the availability of other nonverbal upper-torso visual cues. We studied this by gauging parameters of the communication process during interaction of groups of three partici-pants (two actors and one subject) solving language puzzles under three mediated conditions. Towards the subjects, each condition simulated the use of a video-mediated system, preserving effectively one of the following sets of visual cues: (1) nonverbal upper-torso visual cues other than gaze directional, (2) nonverbal upper-torso visual cues with head orientation, (3) head orientation, gaze and appearance only, using still images. The presence of head orientational cues caused the number of deictic verbal references to persons (deictic use of second-person pronouns) to increase significantly by a factor two. We believe this was due to differences between conditions in the subjects' estimate of the effectiveness of head pointing in disambiguating verbal deixis. We found no effects of the presence of nonverbal upper-torso visual cues other than gaze direction on any of our dependent variables. We did find a significant positive linear relationship between the amount of actor gaze at the facial region of subjects and the number of speaker switches (r=.37) and subject turns (r=.34). As such, the presence of a representation of the visual attention of others increased turn frequency, but only if it could be recognized by subjects as being aimed at themselves. As demonstrated by subject behaviour in the still image condition, the potential increase in turn frequency may be in the order of 25% when gaze at the facial region is conveyed in a manner that preserves its

temporal and spatial characteristics as observed in multiparty face-to-face communication.

	motion video	**still images**
gaze direction	nonverbal visual cues including gaze direction	only gaze directional cues and physical appearance
no gaze direction	nonverbal visual cues other than gaze direction	*physical appearance only (not used)*

Figure 4-1. *Our incomplete 2x2 factorial design.*

INTRODUCTION

In this chapter, we present an empirical study into the isolated effect of a representation of the visual attention of others on multiparty mediated communication in a triadic collaborative setting. As we have seen in Chapter 2, turntaking problems with traditional multiparty mediated systems might be due to a lack of information about whom others are talking or listening to (their dialogic attention). As we have seen in the previous chapter, gaze directional cues seem quite capable of coding such dialogic attention information. However, the isolated effect of the presence of such cues on the multiparty communication process was, to our knowledge, never demonstrated. We therefore conducted an experiment in which the availability of a representation of the visual attention of others in the form of gaze direction was a controlled factor. Our second factor was the availability of other nonverbal visual cues, such as lip movements, as conveyed by motion video. We gauged the isolated effect of these factors on a variety of dependent variables in a triadic mediated collaborative setting. We will first discuss our factors, and how they were used to constitute experimental conditions. For each dependent variable, we will then discuss why it was measured, how this was operationalized, and our predictions towards treatment effects.

Independent Variables

We studied the effects of two factors, or independent variables, on multiparty mediated communication and collaborative performance. Firstly, we were interested in the isolated effect of a single nonverbal visual cue: the presence of a representation of the visual attention of others in the form of gaze direction. This constituted the first factor. As the presence of other nonverbal visual cues is traditionally stressed in design recommendations for video-mediated systems, a second factor was included: the presence of nonverbal visual cues other than gaze direction, such as facial expressions and lip movements (as conveyed by motion video). This would yield the 2x2 factorial design shown in Figure 4-1. However, we did not use a fully factorial design, as this would have led to a condition in which no visual cues (other than physical appearance) would be represented, no longer constituting a video-mediated system by our definition. We therefore defined the following three conditions:

1) A condition in which all nonverbal upper-torso visual cues other than gaze direction were present (hereafter referred to as *motion video-only*).

Although we realized one cannot separate nonverbal eye signals from their gaze directional function, we believed it possible to render this function ineffective by not conveying any head and body orientations other than frontal.

2) A condition in which all upper-torso nonverbal visual cues including gaze direction were present (hereafter referred to as *motion video with gaze direction*).
3) A condition in which only gaze directional cues and upper-torso physical appearance were present (hereafter referred to as *still images with gaze direction*).

With regard to nonverbal visual cues, we decided to restrict ourselves to the area of the body normally conveyed by video-mediated systems: the upper torso. Although we do not wish to trivialize the function of hand and other gestures in communication, this restriction would ease control of treatment variables. Also, as we have seen in the previous chapter, the facial region seems to be the most foveated area of the body during face-to-face communication.

As Sellen's study [126] illustrates, the use of different mediated systems to constitute the above conditions is impossible without introducing other, potentially confounding, differences between conditions. The only way in which we could control the independent variables towards subjects using a similar system in all conditions was to use actors as their conversational partners. These actors would then alter their behaviour towards subjects in accordance with the experimental conditions. Using triads of one replaceable subject and two reusable stooges, we constituted the simplest form of multiparty communication, thereby keeping both the number of subjects as well as the number of stooges required to an absolute minimum.

Gaze at the Facial Region: a Confounding Variable?

As Sellen's study also indicates, video-mediated systems generally do not support gaze at the facial region of a conversational partner. This is because you cannot place a camera behind the video monitor on which the conversational partner is shown. Thus, when a person looks at a conversational partner on his screen, the conversational partner will not perceive this as gaze at his facial region. In the previous chapter, it is suggested that gaze at the facial region is an inherent element of gaze direction as a cue in multiparty conversation. The technology required to allow this cue to be conveyed in all conditions was not at our disposal (for a discussion of that technology, see the *Implications for Design* section, page 119). This was only a problem in the motion video conditions. We decided to simply instruct the stooges to spend as much time as possible looking into the camera lens when looking at their video monitor in those conditions. After the fact, we could then control for differences between conditions in the conveyance of gaze at the facial region using, for example, the amount of gaze at the facial region as a covariate. The confounding nature of

this variable did, however, make experimental predictions with regard to most dependent variables difficult, requiring the use of post-hoc testing in most cases.

Dependent Variables and Predictions

As Monk et al. [99] and the experiments discussed in Chapter 2 demonstrate, results obtained in comparing different mediated settings may depend very much on the experimental task used. Tasks that are highly personal and/or involve conflict are much more sensitive to differences in mediation than, e.g., problem-solving tasks. Thus, they are more likely to affect dependent variables other than task performance itself.

Most measures of task performance itself are typically sensitive only to gross manipulations of the facilities available for communication. Chapanis [27], for example, found differences in task performance between face-to-face and written communication, but not between audio-visual and audio-only mediated communication. One explanation for this is that people tend to perform the primary task at the expense of any secondary task or (mental) effort. A second explanation is that people are extremely flexible in finding ways to obtain or convey the required information. If mediated information is heavily redundantly coded, hardly any effort is needed to supplement the loss of information inflicted by experimental treatments. This makes it difficult to gauge the effect of a particular treatment on mediated communication and collaboration. Most significant effects reported in studies of mediated communication have tended to be measures of the communication process, rather than its outcome. We therefore decided to measure a variety of process variables from the semantical/syntactic levels of communication (number of deictic verbal references) down to conversational structure (number of turns, speaker switches and amount of simultaneous speech). This way, we would obtain as objective and broad a picture of the effect of experimental treatments on the process of mediated communication as possible. This information was supplemented with measures of task performance and subjective experience. For each dependent variable, we will now discuss why it was measured, how this was operationalized and our hypotheses towards treatment effects.

Task Performance

In designing an experimental task, we needed to take into account two conflicting requirements:

1) The task should be sensitive enough to experimental treatment to qualify as a measure of its effect.
2) In order to be able to generalize findings, any detected effect should not be directly induced by the nature of the task.

Realizing the importance of the second criterion, we decided *not* to use a task which was highly personal, involving conflict or debate. This was less

conservative than it seems, as we assumed gaze directional cues to be far less redundantly coded than the other nonverbal visual cues typically assessed in experimentation. Also, we believed the effect of this information would be directly on speech communication as such, rather than on nonverbal communication. By following this strategy, we would have a much stronger case for this argument if any effects were found. We could not use a highly visual task either, for three reasons:

1) Visual attention for the task would affect visual attention for the representation of the participants, possibly reducing the impact of treatment variables [10].
2) We wished to reserve the video-mediated system for the communication of nonverbal visual cues as expressed by the human body, making it easier to control treatment variables.
3) If the task itself required the video-mediated system and full-motion video to be completed, at least one of our independent variables would be confounded.

With regard to the first criterion, we needed a task which required communication between all participants in order to be completed. Effects of treatment variables on task performance would thus be constituted via their effect on the communication process itself, giving the best possible trade-off between our two main criteria. The task should provide some sort of objective score, be novel to the subjects, and not be too difficult to understand or complete. Considering these requirements, we devised a collaborative problem-solving task based on *language puzzles*. For each problem, each session participant would obtain one of three pieces of information required to solve that problem. Participants would need to put these pieces in the correct order to score a point. By communicating pieces and permutations of pieces, participants would collaborate to perform the task. Our performance measure would be based on the number of correct permutations of puzzle pieces given during the length of the experimental session. For a more detailed explanation of this process, see the description of the experimental task on page 91.

Predictions Regarding Task Performance

As our task was not based on the exchange of nonverbal information, we did not expect the presence of nonverbal visual cues other than gaze direction to significantly affect task performance. We did, however, expect a significant effect of the presence of gaze-directional cues on problem solving. Such cues would make it more apparent who was speaking to whom during the exchange and manipulation of puzzle pieces, thus simplifying this process. Due to our confounding variable, however, we did not plan any comparisons.

Deictic Verbal References

In their usability studies on video-mediated vs. face-to-face communication, Isaac and Tang qualitatively observed that there were many instances in face-to-

face interaction when people used their eyegaze to indicate whom they were addressing [78]. However, when using a video-mediated system, which did not convey gaze direction, they observed participants using each other's names to indicate whom they were addressing, instead of their eyegaze. In general, the use deictic references to persons or objects (such as "*You* can try", or "What is *that?*") may be problematic when visuo-spatial cues are not conveyed. In order to disambiguate the meaning of deictic references, either the internal or external context of conversation is required [40]. Internal context is typically provided by a common knowledge about the situation, often based on earlier utterances [30]. For example, if "*You* can try" is a direct response to something the addressed person just said, the meaning of the word *you* is easily disambiguated by knowledge about the identity of the previous speaker. If, however, such knowledge is unavailable or not applicable, external context (i.e., information from the outside world) is needed. In deictic references, such external context can be provided by visuo-spatial cues such as hand pointing or gaze direction [40, 79]. For example, if "*You* can try" is used imperatively, extra information is needed to ascertain whom is being addressed. It is therefore likely that the availability of visuo-spatial cues affects the usefulness of deictic referencing.

We decided not to measure name frequency because we did not intend to use participants that knew each other prior to the experiment. Instead, we measured the ability to use deixis towards persons by counting singular deictic use of second-person pronouns (i.e., the *you* in "Do *you* think?"). A detailed description of the measurement procedure is given in the *Analysis* section on page 98, with contextual examples of deictic references in Appendix B.

Predictions Regarding Deictic Verbal References

We expected the use of deictic verbal references to be affected by the availability of gaze directional cues of the head. Compared to other conversational behaviour, such as interruption, we expected the nature of deixis to be rather low-frequency. Therefore, the use of eye signalling, which is much faster than head signalling, would not be a *precondition* for achieving deixis. We therefore did not expect our confounding variable to be of *significant* influence on deixis. As only a single independent planned comparison was allowed, we formulated the following hypothesis:

"The presence of gaze directional cues in the form of head orientation causes the number of personal deictic verbal references used to rise significantly."

Speaker Switching, Turn Frequency and Turn Duration

Isaacs and Tang [78] observed that during videoconferencing, people tend to control the turntaking process explicitly by requesting individuals to take the next turn. In face-to-face interaction, however, they saw many instances where people used their eyegaze to indicate whom they were addressing and to suggest a next speaker. These qualitative observations are in line with our findings in the previous chapter. With Kendon [85], Argyle and Cook [8] suggested that gaze directional cues play an important role in keeping the floor, taking and avoiding the floor, and suggesting who should speak next. As such, Short et al. [129] attributed problems in turntaking behaviour with mediated systems to a lack of gaze directional cues. We therefore decided to measure the isolated effect of gaze-directional and other nonverbal visual cues on the turntaking process by comparing the number of turns between conditions.

Like Sellen, we measured the frequency and duration of turns using automated analysis of individual speech patterns. The used procedure is explained in detail in the *Analysis* section on page 100. The total number of turns minus one yielded the number of speaker switches, an indication of overall turntaking efficiency.

Predictions Regarding Turntaking Variables

There is little comparable evidence on which to base predictions for speaker switching, turn frequency and turn duration with regard to the presence of gaze directional cues in multiparty communication. Firstly, to our knowledge, there has only been one study (by Sellen [126]) in which the availability of gaze directional cues was, to some extent, part of the treatment. Secondly, most studies, particularly the early ones, were based on dyadic (two-person) communication. Finally, most studies, including the one by Sellen, compared communication settings that differed on too many variables at once.

The most confirmed result from early studies on dyadic communication is a significant increase in the number of turns or a significant decrease in the length of turns in face-to-face conditions, as compared with audio-only conditions [13, 120]. These results may well be explained by a lack of gaze directional cues in audio-only conditions yielding a worse synchronization of turntaking [85]. However, Short et al. [129] believed that if this was true, the effect should be opposite, as he assumed more turns to be an indication of *breakdown* of turn synchronization. With Rutter [118], we believe this assumption is incorrect. When turntaking is better synchronized, it simply becomes *easier* for each participant to get a turn at any given moment in time. There have been reports of significant increases of turn frequency in face-to-face conditions with regard to video-mediated (multiparty) conditions [32, 109], but those findings are confounded by the presence of lag in the mediated system. Sellen [126] failed to find significant differences in the number of turns between several multiparty conversational contexts, all without lag: face-to-face, video-mediated with gaze direction, video-mediated without gaze direction, and audio-only

communication. Even so, if gaze directional cues code who is talking or listening to whom with high entropy, and if this information *is* used in multiparty communication, one might expect its presence to ease turn synchronization and speaker switching, increasing turn frequency, while decreasing turn duration.

Most studies suggest that in comparison with face-to-face communication, the effect of motion video (without gaze direction) on turntaking is similar to the effect of audio-only conditions (see Sellen [126] for an excellent overview). One would therefore expect the effect of nonverbal cues other than gaze direction on turntaking variables to be insignificant. Although they may contain information used in the turntaking process, it is likely such information is either redundantly coded, or easily made available by speech [129].

In the previous chapter, we found that gaze at the facial region is indeed well-synchronized with the articulatory and auditory attention of participants. It seemed likely that differences between conditions in the ability to convey gaze at the facial region might confound results on turntaking variables. We therefore decided not to use planned comparisons for any turntaking variable.

Speech and Simultaneous Speech

Like turn frequency, the amount of simultaneous speech is generally seen as an indication of the degree of interactivity of a conversation [109, 126]. It gives an indication of the ability of conversational partners to interrupt each other, attempting to take the floor (get a turn) when there is no indication that the current speaker is about to relinquish control. According to O'Connaill et al. [109], simultaneous speech may also occur when people try to anticipate or help the current speaker finishing his turn (*projections/completions*), when people simultaneously try to take the floor (*simultaneous starts*), or when people try to hold the floor by producing utterances that do not contain information, e.g., by repeating words (*floorholding*). For reasons of efficiency, we did not use O'Connaill's classification technique. Instead, we simply measured the overall percentage of overlapping speech (i.e., speech by more than one person at a time). The used procedure is explained in detail in the *Analysis* section on page 100.

The overall percentage of speech per participant was measured mostly as a control variable. This way, we would be able to correct the number of deictic verbal references used by participants for differences in the total amount of speech uttered.

Predictions Regarding Speech and Simultaneous Speech

The picture with regard to the amount of simultaneous speech is similar to that of turn frequency. In general, studies have found more simultaneous speech in face-to-face conditions than in video-mediated conditions (without gaze direction) [122, 162]. Sellen [126] found more simultaneous speech in face-to-face conditions than in any of her mediated conditions. Although these mediated conditions did include a setting in which gaze direction was conveyed, Sellen believed its effect might have been reduced by the small size and large separation of the video monitors used. Also, there *was* a small parallax between the camera and the video monitor in her system. This combination may well have rendered gaze at the facial region ineffective.

Expectations with regard to treatment effects on the amount of simultaneous speech were therefore the same as for turn frequency. One might expect the presence of gaze directional cues to ease interruption, thereby increasing the amount of simultaneous speech, with no *significant* effect of the presence of other nonverbal visual cues. As we expected the total amount of speech to be too gross a measure, we did not anticipate any significant treatment effects regarding this variable. Due to our confounding variable, we did not plan any comparisons.

METHODS

The experiment involved groups of three participants (one subject and two stooges) solving language puzzles using a video-mediated system conveying one of three different subsets of nonverbal visual cues. This section discusses the methods used to conduct the experiment, describing experiment design, subjects and stooges, conditions, experimental task, session procedure and the questionnaires used.

Experiment Design

We used an independent samples design comparing performance between three groups of subjects, each group treated with one of the following three conditions:

1) *Motion video-only.* Simulated a single-camera full-motion videoconferencing system, essentially conveying all upper-torso nonverbal visual cues except gaze direction.
2) *Motion video with gaze direction.* Simulated a multiple camera full-motion videoconferencing system conveying all upper-torso nonverbal cues and gaze directional information.
3) *Still images with gaze direction.* Simulated a videoconferencing system which conveyed the gaze direction of the stooges, but no other moving upper-torso nonverbal visual cues. Each stooge could manually select one of three images for display: looking at the subject, looking at the other stooge, and looking at a computer terminal.

As we did not use a fully factorial design, and were unable to make predictions for most of the dependent variables, we treated the design as being single-factor, using post-hoc tests for most of the dependent variables (with the exception of *number of deictic verbal references*).

Experimental Subjects and Stooges

Our experimental subjects were paid volunteers, mostly university students from a variety of technical and social disciplines. Prior to the experiment, we tested all subjects on eyesight and a number of relevant matching variables. In order to minimize the variance between conditions, we allocated each subject to a treatment group in a way that matched the groups on these variables:

- *Dutch language competence*. As our experimental task involved solving language puzzles, the subjects' command of the Dutch language was likely to influence their performance. We used a pen-and-paper language aptitude test (in Dutch [55], based on the Differential Aptitude Test [17]) to gauge subjects' ability in order to reduce between-group variance on this variable.
- *Spatial ability*. Spatial ability is a component of intelligence which is related to the ability to create a visuo-spatial mental image of a task situation. According to Rothkopf [117] and Van der Veer [146, 147] this variable might predict how well a subject was able to imagine the task space,

deducting the potential use of the video-mediated systems from their appearance, as well as deducting the correct task procedure. We used a pen-and-paper aptitude test (in Dutch [55], based on the Differential Aptitude Test [17]) to gauge subjects' ability in order to reduce between-group variance on this variable.

- *Age, sex and field of study*. We did not match groups on personality factors. However, we attempted to minimize potential differences in cognitive style between groups by matching them on these variables.

The subset used for further analysis consisted of 56 subjects assigned to three treatment groups:

- *Motion video-only* group. 20 subjects (13 male, 7 female, mean age 21.4);
- *Motion video with gaze direction* group. 19 subjects (13 male, 6 female, mean age 21.7);
- S*till images with gaze direction* group. 17 subjects (11 male, 6 female, mean age 22.2).

Subjects believed the stooges were subjects also. None of the subjects in this subset knew or had any suspicion regarding the stooges. None had any previous experience with video-mediated communication. Subjects believed we were interested in how people cooperate via the Internet, and were only informed of the true purpose of the experiment after treatment.

We used two stooges, one female and one male. Stooges were seated in a separate room from the experimental subject. The setting was such that the subject believed that both others (the stooges) were each in a different room. Both stooges were about the same age as the subjects. The difference in sex between the stooges may have aided identification of voices in the still images with gaze direction condition, and may have reduced variance within groups that might have occurred due to sex-preferences.

Conditions

In all conditions, stooges used exactly the same video-mediated system to communicate with the subject. Differences on treatment variables were presented only to the subject. As stooges were seated in the same room, they did not use a video-mediated system to communicate with each other. As will be explained, care was taken this would not confound the experiment. The subject assumed the stooges were in two separate rooms, and that everyone was using the same type of video mediation to communicate. For each condition, we will now describe how differences in the behaviour of stooges and system constituted the experimental treatment:

Figure 4-2. *Three different directions of stooge gaze as experienced by the subjects: a) facing the subject; b) looking at their computer screen; and c) looking at the other stooge.*

1) *Motion video-only.* In this condition, the subjects saw a full-motion video image of the stooges, with the stooges always facing the subject (as shown in Figure 4-2a with eyegaze slightly lowered most of the time). Thus, they simulated a single-camera full-motion videoconferencing system, in which all nonverbal visual cues are conveyed except for gaze directional cues of head and body orientation. As will be discussed below, this also rendered the gaze directional function of the eyes ambiguous.
2) *Motion video with gaze direction.* In this condition, the subjects saw a full-motion video image of the stooges. By turning their heads in different directions, stooges indicated whom or what they were looking at: the subject (Figure 4-2a, with eyegaze slightly lowered most of the time), their computer screen (Figure 4-2b), or the other stooge (Figure 4-2c). Thus, they simulated a multiple camera full-motion videoconferencing system conveying all nonverbal cues with gaze directional information. As stooges were in the same room, it would have been possible to achieve eye-contact between them in this condition. To avoid this potentially confounding

effect, when looking at each other, they looked at a common reference point instead (a puppet placed at the position marked with a + in Figure 4-3 on page 94).

3) *Still images with gaze direction.* At any moment in time, stooges would manually select one of three still images for display to the subject: stooge looking at subject (Figure 4-2a), stooge looking at computer screen (Figure 4-2b), or stooge looking at other stooge (Figure 4-2c). Stooges were instructed to base their selection on whom or what they would actually be looking at. This looking behaviour essentially replicated that of condition 2. Thus, they simulated a videoconferencing system conveying gaze direction, but no other moving nonverbal visual cues.

As will be explained further in the *Materials* section, the video-mediated system each stooge used to communicate with the subject had a considerable parallax between the camera lens (used to capture images of the stooges in both full-motion conditions), and the center of the video monitor (which displayed a full-motion image of the subject in all conditions).

This parallax made it difficult for the stooges to convey gaze at the facial region of the subject in both motion video conditions (1 and 2). In the still image condition, parallax was not an issue, as the image displayed when looking at the subject always featured the stooge looking into the camera lens (Figure 4-2a). This introduced a potentially confounding difference between conditions. Stooges were instructed to spend as much time as possible looking into the camera lens when looking at their video monitor in conditions 1 and 2. In condition 2, this signal would be correctly perceived by the subject as gaze at his facial region. In condition 1, the meaning of this signal was ambiguous, as subjects would not be able to discern whether this was meant as gaze at their facial region, or gaze at the facial region of the other stooge. As will be explained in the *Analysis* and *Results* sections, differences between conditions in the amount of gaze by the stooges at the facial region of the subject were controlled for after the fact.

Suppose the subject gets the following fragment on her screen: 'it was cold'. The subject asks what the stooges have on their screen. Stooge 1 says: 'the good weather' and Stooge 2 reads: 'despite'. First, the subject asks to enter the sentence 'it was cold despite the good weather'. Stooge 1 enters the sentence. Stooge 2 confirms it is correct, but as nothing else happens, they decide to try another permutation. The subject attempts 'was it cold despite the good weather?' but it appears to be wrong. 'was it cold' is a manipulation of 'it was cold', which is not allowed. After discussing this, the subject asks again what each stooge had. Stooge 2 replies 'despite', Stooge 1 says 'the good weather'. Entering 'despite the good weather, it was cold' completes the problem, with a new fragment appearing on the subject's screen.

Box 4-1. *An example of a typical problem solving process during the experimental task.*

Task

Based on our requirements, we constructed a group problem solving task in which each subject was asked to join the stooges — perceived as being subjects also — in solving as many language puzzles as possible within a time span of 15 minutes. For each language puzzle, each participant (the subject and each stooge) was presented a different fragment of a sentence (yielding a total of 3 fragments per puzzle). To solve each puzzle, they had to construct as many meaningful and syntactically correct permutations of the sentence fragments as possible (yielding a theoretical 6 possible solutions per puzzle). After having given all correct answers to a particular language puzzle, another set of fragments would be presented. For the creation of each permutation, participants had to use the following rules:

1) Each permutation had to be grammatically correct.
2) Each permutation had to be meaningful.
3) To make a permutation meaningful, they were allowed to add punctuation marks, as long as the permutation remained one sentence.
4) The order of the words inside each fragment should not be altered.

Each language puzzle was presented in Dutch, as this was the participants' native language. For the subject, each sentence fragment appeared on a computer screen. The stooges pretended this was the case for them also, having their fragments listed successively on a piece of paper instead. In order to prevent a practice effect, this piece of paper also listed all correct answers to each puzzle. The paper also prescribed which correct solutions they were allowed to give away, and when to give incorrect solutions. This was done to minimize the influence of stooges on task performance while keeping their act credible towards the subject. We took the following precautions to ensure an exchange of information between the subject and each stooge was necessary to complete each puzzle:

1) None of the participants could see the sentence fragment of the other participants.
2) Each fragment remained on the subject's screen for only 10 seconds.
3) Each participant had a specific role. The subject's role was to officially submit each solution they collectively agreed on to be correct. Stooge 1 would then pretend to enter this solution for verification by a computer system, while Stooge 2 would report its correctness, pretending this was indicated on her computer screen.

When all correct permutations were given, a computer system would provide a new sentence fragment on the subject's computer screen, generating an audio signal to inform the stooges of this. The number of correct permutations generated per 15 minute session was used as a measure of task performance. Correct permutations that were given more than once counted only once, and uncompleted language puzzles were disregarded. In Box 4-1, a typical example of the experimental task is given. The subset of used language puzzles is presented in Appendix B.

Instructions and Session Procedure

Prior to the experiment, stooges were instructed with regard to their behaviour in the different conditions, which they practiced in several training sessions. Stooges memorized all answers to all problems solved in the experimental task prior to the experiment to avoid practice effects. They were not informed until after the experiment of the purpose of the experiment or reasons behind the experimental treatments. Stooges were instructed to behave as if they were subjects, with a similar system setup. However, stooges were told to allow the actual subject to take the initiative as much as possible. This resulted in a situation in which much of the interaction was between the subject and one of the stooges, rather than between stooges only.

For each subject, the experimental session was structured in the following way. Before the session, the subject was taken to the experimentation room by a host, and seated in front of the video-mediated setup, the components of which were briefly described to each subject. After introducing the subject to the system, the host would ask the experimenter if all subjects had arrived, after which the video links were turned on. The session started with the experimenter addressing the subject through the speaker system, in a way which suggested he addressed all participants. From this moment onwards, participants could see and hear each other. After introducing themselves, the experimenter explained the role of each participant using a simple practice game, in which as many names of mammals, starting with a 'D', were to be listed as possible. After exactly one minute, the experimenter interrupted the game and started to explain the rules of the actual task. This included an example puzzle, in which all participants were asked to read their sentence fragment, after which the experimenter explained the correct answers. The session would then proceed

with the first puzzle, with the experimenter ending the session 15 minutes later. After each session, the subject filled in a questionnaire and was debriefed by the host. First, the subject was asked if he had noticed anything out of the ordinary, after which the host explained the role of the stooges and the purpose of the experiment. The subject was also asked to keep this information confidential for the length of the experiment.

Questionnaire

After completing their experimental session, subjects were given the questionnaire listed in Appendix B. It contained four open questions and seven multiple choice questions. Each multiple choice question was answered by selecting one of five options, ranked from strongly positive to strongly negative statements (e.g., 'strongly agree', 'agree', 'undecided', 'disagree' and 'strongly disagree'). To reduce response set, we arranged the order of the options from strongly negative to strongly positive for about half the questions, and from strongly positive to strongly negative for the other half. One question was used to verify that the subject did not know any of the stooges prior to the experiment.

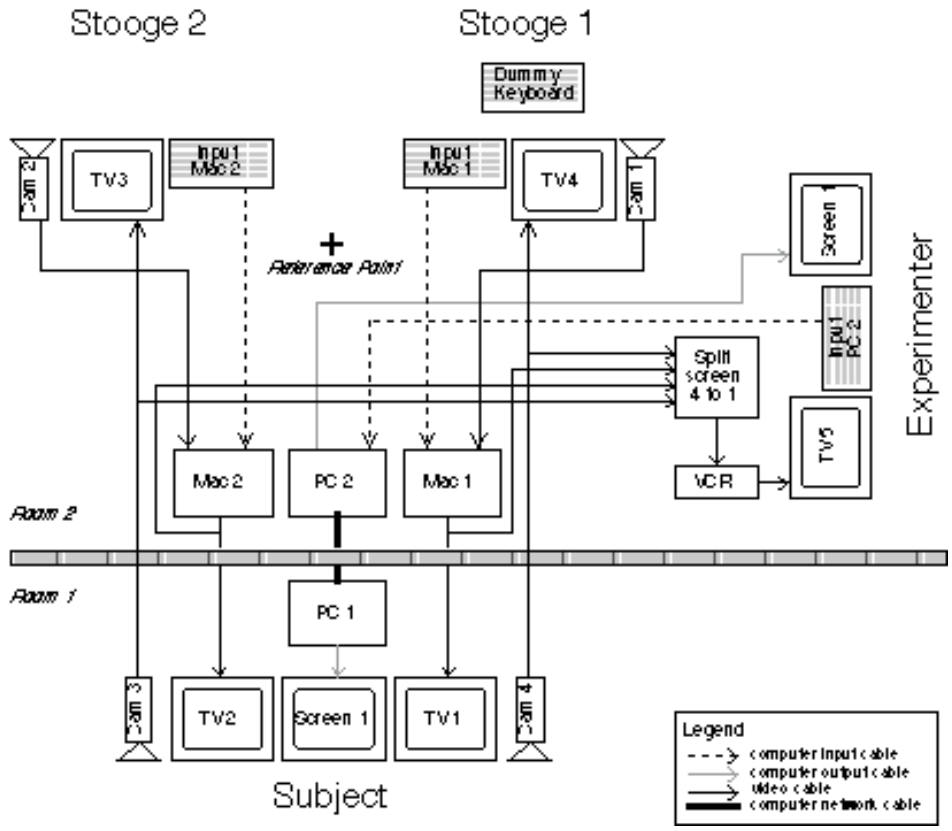


Figure 4-3. *Overview of experimental setup showing video and computer network configuration.*

MATERIALS

All equipment for this experiment was set up in a way that minimized differences between conditions to the treatment variables only. All video and audio equipment was analog to minimize lag. To make the experiment as cost-effective and reliable as possible, we attempted to construct the absolute minimal setup required to simulate our conditions. We will first discuss the experimental setup for subjects and stooges, after which the registration equipment is described as operated by the experimenter.

Experimental Setup

An overview of the equipment used in the experiment is shown in Figure 4-3. Two rooms were used in the experiment: one in which the subject was seated, and an adjacent room in which the stooges and experimenter were seated. We will first describe the subject's configuration, after which the setup for the stooges is discussed.

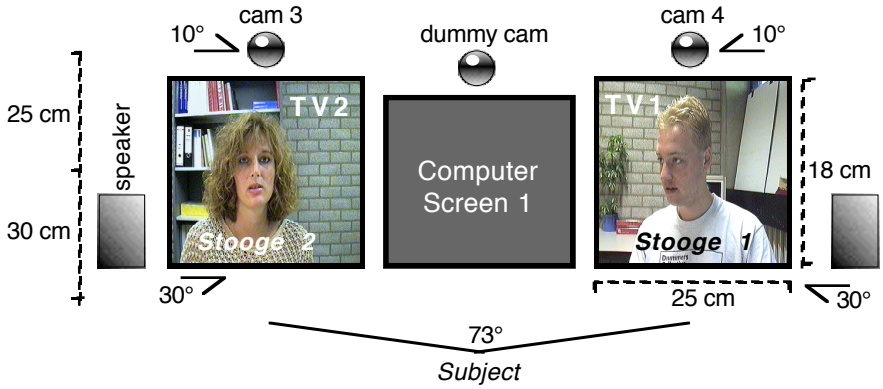


Figure 4-4. *The video-mediated system as experienced by the subjects.*

Subject Configuration

Figure 4-4 shows the experimental setup as experienced by the subjects. The subject was seated in front of two video monitors, with the right monitor (*TV1*) displaying the image of Stooge 1 and the left monitor (*TV2*) the image of Stooge 2. Between these video monitors, a computer monitor (*Screen 1)* was placed, used to display the subject's sentence fragment.

For the two motion video conditions, the source of the stooge image on each video monitor (14" Sony Trinitron) was an *Apple Videoconferencing Camera* placed on top of the video monitor of the respective stooge, with TV1 displaying the image of camera 1 and TV2 the image of camera 2. Before display on the subject's video monitors, camera signals were fed through Apple PowerMac 6100 and 5300 A/V computers (*Mac 1* and *Mac 2*) running *Apple Video Monitor* software. This was to ensure compatibility of the images in both motion video conditions with the images in the still image condition, and introduced no discernable delay as *DMA* was used (with peripheral video input written directly into display memory). Images were converted back to PAL video signals by means of the internal video circuitry of the PowerMacs, again without any discernable delay. In the still image condition, stooges used *Microsoft Powerpoint*, running on Mac 1 and Mac 2 respectively, to select one of three images for display on the subject's video monitor: stooge looking at subject; stooge looking at other stooge; and stooge looking at computer screen (see Figure 4-2 on page 89). These still images were captured before the experiment using the same *Apple Videoconferencing Cameras* used in the other conditions, with no discernable differences in image quality, resolution, colour, size or physical appearance of stooges between conditions (stooges used the same clothes and haircut for all sessions). The video monitors were placed in an angular setting of 30˚ respective to the edge of the table, tilted 5˚ backwards. The size of the images on the video monitors was approximately 25 x 18 cm in

all conditions, with 16-bit colour and 768 x 576 resolution in 25 frames per second, 50 Hz interlaced PAL TV format. The center of these images was located approximately 30 cm above the table surface. Subjects could hear the stooges' microphone signals by means of two amplified HiFi speakers, each placed next to the respective video monitor for that stooge, preserving the spatial separation of their voices equally in all conditions. A third amplified speaker, located on top of the computer monitor, conveyed the voice of the experimenter during instruction. As all signals were essentially analog, audio and video were always synchronized, and without any discernable lag.

The 15" monochrome computer monitor (*Screen 1*) was used to display a new sentence fragment for the subject, generated by the experimenter's DOS computer (*PC2*) after completion of the previous problem. Each sentence fragment remained visible for approximately 10 seconds, after which this monitor turned black. The characters of the sentence fragments were white on black and approximately .5 cm tall (height of the letter 'x'). On top of each monitor, a camera was placed, with the lens at approximately 25 cm above the center of the monitor image, pointing downwards at the subject with an angle of about 10°. The video and internal microphone signals of each of the two outermost cameras (Sony *Handycams*) was fed to the stooges, with camera 3 fed to Stooge 2 and camera 4 to stooge 1. This setup allowed the stooges to see and hear the subject, and get an idea of his gaze direction. It also allowed subjects to believe they had exactly the same setup as the stooges — which they believed to be subjects also — in each condition. The camera on top of the computer monitor was a dummy, placed there to make this full symmetry more credible in the motion video-only condition, where subjects always saw a frontal shot of the stooges.

Subjects were seated on a chair with a seat 47 cm above the floor. The back of the chair was approximately 37 cm from the edge of the table. The height of the table surface was 75 cm above the floor. The average distance from the head of the subject to each monitor was approximately 60 cm. From the subject's point of view, the angle between the center of the left monitor image and the center of the right monitor image was approximately 73°.

Stooge Configuration

The two stooges were seated together with the experimenter in a room adjacent to the subject room. The equipment for each stooge was about half the subject configuration. Stooges each had only one video monitor (*TV3* and *TV4*, on which they *always* saw live video images of the subject) with a videoconferencing camera on top, a numeric keyboard for selecting images in the still-image condition, and a sheet holder with a sheet of paper listing their sentence fragment for each successive problem sentence, all solutions, and instructions when to give away an answer. In addition, Stooge 1 had a disconnected computer keyboard with which he pretended to feed answers into a computer for verification. Stooge 2 used her sheet to pretend verifying the

correctness of those answers on a computer screen. The stooges' video monitors were identical to the ones used by the subject. In all conditions, Stooge 1 got the live image of camera 4, and Stooge 2 got the live image of camera 3. The center of these images was located approximately 22 cm above the stooge's table surface. This setup preserved the subject's gaze direction for the stooges, albeit without subject gaze at their facial region. An Apple videoconferencing camera was placed on top of each video monitor, with its lens about 17 cm. above the center of the monitor image, pointing almost horizontally at the eyes of the stooges. Thus, we tried to minimize the camera lens/monitor center parallax for the images viewed by the subject. Each stooge could hear the subject's voice by means of a small in-ear headphone, invisible to the subject. The signal for Stooge 1 came from the internal microphone of camera 4, for Stooge 2 from camera 3. The stooges each had a unidirectional microphone placed in front of them on the table. The signal from each microphone was amplified and fed to the respective speaker in the subject room.

Video and Audio Registration Equipment

The experimenter, seated in the same room as the stooges, operated the video tape recorder (*VCR*) on his table. All 4 video signals, two of the experimental subject and one of each stooge, were fed into a video splitter. This video splitter generated a colour image split into four quadrants, each quadrant showing one input. This image was recorded on a PAL VHS video tape recorder (*VCR)* for later analysis, and was monitored by the experimenter using *TV5*. All microphone signals were mixed and fed into the video recorder for registration along with the video signals. The voice of Stooge 1 was recorded on the left channel, the voice of Stooge 2 on the right channel and the voice of the subject on both audio channels of the video tape recorder. The experimenter also wore a microphone, connected to the amplified speaker on the subject's computer screen. This audio link was used for instructions, and to end each session. The experimenter used a button to turn all systems on at the start of each session, turning them off again at the end. The experimenter used a stopwatch to time each session.

Task Performance Registration Equipment

Like the stooges, the experimenter had a piece of paper on which all correct answers to each problem were listed. When all correct answers to a problem were given, the experimenter would register this by pressing a key on his computer keyboard (*Input PC2*), which was connected to *PC2*. Using these keystrokes, this DOS computer would count the subject's score, saving them to disk after each session. If all correct solutions to a problem were given, *PC2* would generate a clearly audible beep, displaying the next problem sentence fragment on the subject's computer monitor (*Screen 1*) via an ethernet link with *PC1*. Using his computer screen, which displayed a copy of *Screen 1*, the experimenter could monitor this process.

ANALYSIS

During this phase, video tapes were analyzed to obtain measurements of the following dependent variables: the number of deictic verbal references made by subjects and stooges; the percentage of speech by subjects and stooges; the number of turns by subjects and stooges; and the percentage of gaze at the subject's facial region by the stooges. We used a mix of human and automated classification methods, all of which are discussed below. The reliability of each method was checked by correlation of results between different observers.

Analysis of Deictic Verbal References

The number of deictic verbal references used by subjects and stooges was scored by the experimenter and an independent observer for each session. Of each session, the full 15-minute video recording was analyzed. During analysis, both experimenter and the independent observer were blind to the experimental conditions (i.e., they could not see the stooges' video signal). The independent observer was also blind to any experimental predictions or other details. Before scoring, rules were agreed between observers on what constituted a correct reference. Only deictic second-person pronouns were scored (i.e., the words *you* and *your* in "Are *you* sure that was *your* sentence?"), according to the following criteria:

- the reference was to one person only. In Dutch, this is easy to discriminate as there are different words for plural (*jullie*) and singular (*jij*) use.
- the reference was not directly preceded or followed by a name (as in "*Tom*, *you* said").
- the reference was not used in a generic way (i.e., depending on context, it should not be possible to substitute the word *one* for the word *you:* "*You* would think it's possible — *One* would think it's possible.").
- repetitions were scored only once (e.g., "*You*, *you* said").
- references in task sentences (the actual problems solved) were not scored.

Occasionally, internal context of conversation was used to determine whether the use of a pronoun was deictic. Examples of this are given in Appendix B.

Before scoring, both observers practiced the use of the above criteria on a subset of experimental sessions not used for further analysis. After these training sessions, the inter-observer reliability was determined. Both observers independently scored 45 minutes of conversation taken from 3 sessions which were not used for further analysis. Each minute per participant constituted one data cell containing *true* if one or more deictic verbal references were scored for that person during that minute, and *false* if not. This resulted in 135 (3 participants * 45 minutes) independent dichotomous observations per observer. Averaged between observers, 26 observations (19.3 %) was scored *true*. After calculating the tetrachoric correlation between these two sets of data we obtained a significant reliability of φ=.86 between observers ($p<.001$, 2-tailed).

This indicates the two observers agreed well on the rules of scoring, and that scoring can be replicated reliably by future experimenters.

Subsequent analysis of 56 experimental sessions, averaged between observers, yielded a mean number of deictic verbal references for the subject, Stooge 1 and Stooge 2 in each session.

Speech Data Preparation

We analyzed only the first five minutes of each session for turntaking behaviour and speech activity of subjects and stooges. The first five minutes were selected because observation of video recordings indicated subjects were unlikely to have developed a routine for problem solving during this time, yielding a relatively rich interaction process between subject and stooges.

In order to obtain as objective and reliable statistics as possible, we used an automated procedure to analyze the speech patterns of individual speakers. As we did not, at the time of this experiment, have the equipment to digitize voices of individual participants in real-time as outlined in Chapter 3, we recorded the subject and stooge microphone signals by mixing them onto two analog tracks along with the video signal. Stooge 1 was recorded on the left track, Stooge 2 on the right track, with the subject's voice recorded on both tracks. As automated analysis could only be carried out on the isolated speech data of individual speakers, the two-track recordings were separated by hand into three separate digital audio tracks (22 KHz, 8-bit) for each session. Only a minimum amount of interpretation was required in this process. The following rules were used:

- if only one person was speaking, her voice data was copied to her individual track, padded with silence before and after the utterance.
- if more than one person was speaking at a time, the best-quality audio signal for each person was copied to each individual track. Although crosstalk was inevitable at the occurrence of simultaneous speech, we were only interested in whether a person was speaking or not (i.e., the presence of speech energy). We therefore only copied speech data to an individual track if speech energy of that individual was present in the signal (with an accuracy of approximately 120 ms). Simultaneous speech accounted for less than 9 percent of total speech activity.
- laughter by an individual was filtered out if it lasted longer than approximately 1.5 seconds. This happened infrequently.
- joint laughter and unintentional coughs were filtered out. All other voice activity, including backchannels, was preserved. Typing and environmental noises were filtered out.
- if a speaker could not be identified, speech data was skipped. This happened infrequently.

To check the reliability of this separation process, the experimenter and an independent observer, after agreeing on the above rules, classified the voice data of an unused session of about five minutes into three individual tracks. After classification, the subject's voice track was down-sampled to 2477 1-bit samples of 120 ms length, each sample constituting one data cell containing *true* if speech energy was present, and *false* if not. Averaged between observers, 297 data cells (12 %) were scored *true*. After calculating the tetrachoric correlation between these two sets of data we obtained a significant reliability of φ=.97 between observers ($p<.001$, 2-tailed). This means observers could almost exactly replicate each other's results, indeed suggesting only a minimum amount of interpretation was required in this process.

Analysis of Speech and Turntaking Behaviour

Speech and turntaking behaviour was analyzed automatically using the algorithm outlined below. The automated analysis process consisted of three steps:

1) Downsampling of the signal to 1-bit samples of 120 ms length containing *true* if speech energy was present, and *false* if not (described below).
2) Utterance analysis as described on page 55 of Chapter 3.
3) Calculation of counts and percentages (described below).

Step 1. Downsampling

After separation of voices into individual audio tracks (22KHz, 8-bit), each track was automatically processed to downsample the signal to 1-bit samples of 120 ms length containing *true* if speech energy was present for at least half the 120 ms interval, and *false* if not.

First, the signal/noise ratio of the audio signal was improved by removing background noise and by amplifying faint speech. The input signal was noise-gated with a threshold of 3% of maximum signal strength (max. gain), effectively removing all noise recorded as silence. Input signals between 3% and 25% of max. gain were amplified on average by a factor 2. Louder signals received almost no amplification. The transfer function for this kind of operation is shown in Equation 4-1 for positive values of input signal x only (ranging from 0 to 127), yielding an output signal y (for negative values of x, use negative factors):

$$y(x) = \begin{cases} 0, & 0 \geq |x| \leq 2 \\ 4(|x| - 2), & 3 \geq |x| \leq 8 \\ 0.86(|x| - 8) + 24, & 9 \geq |x| \leq 127 \end{cases} \qquad \textit{(Equation 4-1)}$$

Then, the envelope of the signal was calculated for each 10 ms interval. This effectively removed periodicity from the signal — conserving speech energy only — by finding the highest absolute sample value in the interval and filling

the interval with that value. The resulting envelope signal, consisting of positive values only, was then downsampled to a 1-bit signal with a sample rate of 8.3 Hz, yielding 2500 samples of 120 ms length per 5 minutes, with each sample *true* if average speech energy was above zero, and *false* if not.

Step 2. Utterance Analysis

This was the most complicated part of analysis. We used the fuzzy algorithm described in detail in Chapter 3 on page 55. First, this algorithm filled in 240 ms pauses to account for stop consonants, effectively removing pauses within words [19]. Then, talkspurt analysis removed pauses between words, but only if those words were spoken consecutively. This way, talkspurts with a length of at least one phonemic clause could be identified (the phonemic clause is regarded as a basic syntactic unit of speech, an uninterrupted vocalization of 2-10 words with a typical duration of approximately 1.5 s, see Chapter 3, page 55 for a discussion). To do this, a 13-sample (1.56 s) window moved over the speech data, filling samples within a 70% confidence interval around its mean position with speech energy if more than half of the samples in the window indicated speech activity, and if this speech activity was well-balanced over the window. Finally, if one of the speakers had a talkspurt of longer than a phonemic clause (i.e., about 1.5 seconds or 13 samples) with everybody else being silent for the same length of time, an utterance was assigned to him. We realize this algorithm would not be able to handle side conversations. However, as we only had three participants per session, side conversations did not occur.

Step 3. Calculation of Counts and Percentages

The percentage of speech by a person equalled the percentage of samples with a value *true* in her speech signal *after* talkspurt analysis. The percentage of simultaneous speech equalled the percentage of time in which more than one person shared a value *true* in their speech signals *after* talkspurt analysis.

The number of turns for a person equalled the number of utterance starts by that person, except for the last speaker in the session, who's last turn was not counted as its length could not be verified. The mean turn duration was determined by summating the time between utterance starts, and dividing it by the number of turns. Thus, turn duration included the pauses between utterances. The number of speaker switches equalled the total number of turns per session minus one.

Meaningfulness of Automated Turntaking Analysis

We checked the concurrent validity of the above turntaking analysis algorithm by calculating the correlation between a turn classification produced by the algorithm and that produced by a trained linguist, yielding a measure for the meaningfulness of automated turntaking analysis. We asked the human expert to classify subject turns during 5 minutes of an experimental session not used for further analysis, based on the following definition of a turn:

"When a person utters at least one phonemic clause, while others are silent for at least that one phonemic clause, that person gets a turn, starting with the first speech of the first phonemic clause, and ending with the start of the next turn (by a different speaker)."

The linguist used his own definition of a phonemic clause based on intonation and semantics, not length. It consisted of the following rules, the first of which was adapted from Jaffe [81]:

1) *"A string of words in which there is one primary stress and which is terminated by a juncture, a slight slowing of speech, with slight intonation changes at the very end."*
2) *"This string of words expressed a proposition."*

Although the original definition by Jaffe concerned the English language, the linguist insisted such definition would be valid for the Dutch language also. The human expert and the algorithm both indicated for each 120 ms of audio whether the subject had a turn or not. Each 120 ms sample thus constituted one data cell containing *true* if the subject had a turn during that 120 ms interval, and false if not. This resulted two sets of 2304 independent dichotomous observations, one from the automated classification method and one from the human expert classification method. Averaged between methods, 483 observations (21 %) was scored true. After calculating the tetrachoric correlation between these two sets of data we obtained a significant concurrent validity of φ=.64 between classification methods ($p<.001$, 2-tailed). This score is quite acceptable. The algorithm, which identified phonemic clauses simply by checking the duration of consecutive speech based on the *average* duration of a phonemic clause, did well against the much more flexible human expert, who identified phonemic clauses using intonation and semantics of speech, regardless of its length. We therefore consider the algorithmic classification of turns to be quite acceptable as a substitute for human classification.

Analysis of Gaze at the Facial Region

The study outlined in the previous chapter shows that people tend to spend a considerable amount of time looking at the facial regions of their partners during conversations, both while listening as well as while speaking.

We expected the three systems emulated in the experiment to differ considerably in their support for gaze by the stooges at the facial region of the subject. In both motion video conditions, stooges were instructed to spend as much time as possible looking into the camera when looking at the subject, thus appearing to gaze at the subject's facial region. However, in the still image with gaze direction condition, stooges would *always* appear to look the subject in the eyes when facing her. As we considered this difference a potential confounding variable in the experiment, we needed to measure these differences in order to control for them after the fact. For each stooge, we therefore measured the percentage of time spent gazing at the camera lens during the first 5 minutes of each session. An independent observer, blind to any experimental predictions or other details, counted the number of video frames per stooge in which the eyes of that stooge appeared to look straight at her. Blinks while looking into the camera lens were included in the frame count, unless the stooge would look elsewhere after the blink.

Before scoring, both the observer and the experimenter practiced observation on a subset of experimental sessions not used for further analysis. After these training sessions, but before scoring, the inter-observer reliability was determined. Both the observer and the experimenter independently scored about 15 minutes of video taken from 3 unused sessions. This material contained approximately 5 minutes of each experimental condition.

The observer and the experimenter independently indicated for each 120 ms of video whether the stooges looked into the camera lens or not. Each 120 ms sample thus constituted one data cell containing *true* if a stooge looked into the camera lens during that 120 ms interval, and *false* if not. For each stooge, this yielded 3 tracks of approximately 2500 data cells, one track per condition. Averaged between observers and stooges, approximately 332 cells (14%) were scored *true* in the motion video-only condition, 75 cells (3%) in the motion video with gaze direction condition, and 899 cells (36%) in the still image with gaze direction condition.

An overview of the resulting correlations between the observer data sets is given in Table 4-1 for each stooge and each condition. Inter-observer reliabilities averaged across conditions (using the Fisher *Z* transformation) were φ=.94 for Stooge 1 gaze and φ=.87 for Stooge 2 gaze. All correlations were significant ($p<.001$). Overall results indicate the scoring method was reliable, allowing replication by future experimenters.

Variable	Inter-observer Reliability of Gaze Score φ		
	Motion	Motion+Gaze	Still+Gaze
Stooge 1 Gaze	.88	.83	.99
Stooge 2 Gaze	.70	.77	.97

Table 4-1. *Inter-observer reliabilities for scoring the percentage of stooge gaze at the subject's facial region per stooge per condition.*

Subsequent scoring by the independent observer of the first 5 minutes of each of the 56 used sessions yielded a mean percentage of gaze at the subject's facial region per stooge per session. This percentage was calculated by dividing the number of frames by the total number of frames analyzed. As differences between stooges were insignificant and small, percentages were averaged between stooges, yielding a single measure of gaze at the subject's facial region per session.

Variable	Task Performance *mean (s.e.)*			Results
	Motion	Motion+Gaze	Still+Gaze	
Number of correct answers	32.3 (1.4)	30.9 (1.7)	28.8 (1.3)	not sign.

Table 4-2. *Means and standard errors for the number of problems solved per session.*

RESULTS

Results for each variable were calculated over the same 56 sessions; 20 sessions in the motion video-only condition, 19 sessions in the motion with gaze direction condition and 17 sessions in the still image with gaze direction condition. Task performance and deictic verbal references were measured over the full 15 minutes of each session, turntaking and speech variables were measured only for the first 5 session minutes.

Where appropriate, analyses of variance (one-way ANOVAs) were carried out, evaluated throughout this chapter at α=.05 level. Post-hoc comparisons were carried out using Student-Newman-Keuls (SNK) evaluated at α=.05 level. Where samples could be considered independent, we used analysis of variance to test differences in percentage of absolute time. We did not consider this problematic as all compared percentages had an absolute equivalent in seconds, relating to the same absolute time frame for each condition.

One planned comparison was carried out using a one-tailed t-test evaluated at α=.05 level. All data was normally distributed (Kolmogorov-Smirnov test, $p>.05$) with equal variances between conditions (Levene test for Homogeneity of Variance, $p>.05$) unless indicated.

Task Performance

Table 4-2 presents the data summary for the number of correct answers given per session (i.e., the total number of correct permutations of sentence fragments for all problems completed during the session).

Analysis of variance showed no significant differences between conditions with regard to the number of problems solved ($F(2, 53)=1.39$, $p=.26$). Neither of the stooges demonstrated a significant practice effect over sessions regarding this variable.

Variable	Deictic 2nd-pers. Pronouns *mean (s.e.)*			Results
	Motion	Motion+Gaze	Still+Gaze	
Subject number of verbal references	1.3 (.3)	2.6 (.7)	3.6 (.9)	MG> M
Mean stooge number of verbal references	1.8 (.3)	2.4 (.3)	2.2 (.3)	not sign.

Table 4-3. *Means and standard errors for the number of deictic verbal references using second-person pronouns per first 5 session minutes.*

Deictic Verbal References

Table 4-3 presents the data summary for the number of deictic verbal references using second-person pronouns (i.e., the *you* in "What do *you* think") during the first five session minutes. Numbers were obtained by averaging the double-blind scores of a research assistant with the single-blind scores of the experimenter (inter-observer reliability on a separate pilot sample φ=.86, p<.001).

A planned comparison showed that subjects used twice as many deictic verbal references in the condition conveying motion video with gaze direction than in the condition conveying motion video only (t(26.93)=1.82, p<.04, uneq. var., 1-tailed), thus confirming our hypothesis.

Analysis of variance showed no significant differences between conditions with regard to the average number of deictic verbal references made by the stooges (F(2, 53)=.88, p=.42). Neither of the stooges demonstrated a significant practice effect over sessions regarding this variable.

Variable	**Amount of Speech** *mean (s.e.)*			**Results**
	Motion	Motion+Gaze	Still+Gaze	
Subject amount of speech (% time)	13.6 (1.1)	16.5 (1.7)	15.8 (1.4)	not tested*
Amount of simultaneous speech (% time)	3.4 (.4)	4.2 (.5)	4.0 (.4)	not sign.

Table 4-4. *Means and standard errors for the amount of subject and simultaneous speech per first 5 session minutes. Amounts are given as percentages of that absolute time frame.*

Variable	**Speaker Distribution** *mean (s.e.)*			**Results**
	Subject	Stooge 1	Stooge 2	
Amount of speech (% time)	15.2 (.8)	18.2 (.8)	10.5 (.4)	not tested*°
Number of turns	6.5 (.3)	6.3 (.3)	4.3 (.3)	S=S1≠S2

Table 4-5. *Means and standard errors for the amount of speech and number of turns for each speaker per first 5 session minutes. The amount of speech is given as a percentage of that absolute time frame.*

Speech Variables

Table 4-4 shows the data summary for the amount of subject speech and the amount of simultaneous speech during the first five session minutes. Differences between conditions in the amount of speech by the subject were not tested as variances were not homogeneous (Levene test $(2, 53)=4.56$, $p<.02$ 2-tailed). Analysis of variance found no significant differences between conditions with regard to the amount of simultaneous speech ($F(2, 53)=.88$, $p=.42$).

Table 4-5 presents the data summary for the distribution of speech and turns among speakers during the first five session minutes. Differences between speakers in the amount of speech were not tested, as variances were not homogeneous (Levene test $(2, 165)=13.3$, $p<.001$ 2-tailed) and samples might be considered dependent°. Analysis of variance showed differences in the number of turns between speakers to be significant ($F(2, 165)=16.72$, $p<.0001$). Post-hoc comparisons showed the difference lied in the number of turns by Stooge 2 (SNK, $p<.05$).

---

* Failed Levene test for homogeneity of variance.

° Although the total percentage of speech per session stayed well below the 100% level, one might argue that the percentage of time left for stooge 2 depended on the percentage of time used by the subject and Stooge 1. Therefore, analysis of variance was not appropriate as samples might not be independent.

Variable	Speaker Turns *mean (s.e.)*			Results
	Motion	Motion+Gaze	Still+Gaze	
Number of speaker switches	14.7 (1.0)	15.1 (1.1)	18.9 (1.5)	M=MG≠SG
Subject number of turns	5.9 (.4)	6.3 (.5)	7.7 (.7)	M≠ SG
Subject turn duration (s)	18.9 (1.8)	18.0 (1.9)	14.8 (1.7)	not sign.
Stooge 1 number of turns	6.1 (.5)	5.3 (.4)	7.8 (.7)	M=MG≠SG
Stooge 2 number of turns	3.8 (.4)	4.6 (.5)	4.5 (.5)	not sign.

Table 4-6. *Means and standard errors for the number and duration of speaker turns per first 5 session minutes.*

Turntaking Behaviour

Table 4-6 shows the data summary for the number of speaker switches and the number and duration of individual turns during the first five session minutes. Numbers were obtained by automatic analysis of the sound tracks of individual speakers. The concurrent validity of the analysis algorithm was checked by correlation with a human expert performing the same task (φ=.64, p<.001).

Analysis of variance showed the number of speaker switches to differ significantly across conditions ($F(2, 53)=3.75$, p<.03). Post-hoc comparisons showed this difference lied in the condition conveying still images with gaze direction (SNK, p<.05). There were over 25% more speaker switches in this condition.

Differences across conditions in the number of individual turns by subjects showed a similar trend ($F(2, 53)=3.17$, p=.05). Post-hoc comparisons suggested the still image with gaze direction condition to be different from the motion video-only condition (SNK, p<.05). Differences across conditions in the mean duration of subject turns were not significant ($F(2, 53)=1.32$, p=.28).

Differences across conditions in the number of turns by Stooge 1 were significant ($F(2, 53)=5.39$, p<.01). Post-hoc comparisons showed the still image with gaze direction condition was different from the other conditions (SNK, p<.05). There was no significant difference across conditions in the number of turns by Stooge 2 ($F(2, 53)=.94$, p=.40). Stooge 2 showed a practice effect over sessions (correlation between session order and number of turns per session r=.46, p<.001).

Variable	Amount of Stooge Gaze *mean (s.e.)*			Results
	Motion	Motion+Gaze	Still+Gaze	
Mean stooge amount of gaze (% time)	13.8 (1.2)	6.6 (.8)	31.6 (1.5)	M≠ MG≠ SG

Table 4-7. *Means and standard errors for the amount of stooge gaze at the subject's facial region per first 5 session minutes, given as a percentage of that absolute time frame.*

Variable	Estimated Means Adjusted for Gaze			Results	Corr. Gaze
	Motion	Motion+Gaze	Still+Gaze		
Number of speaker switches	15.4	17.1	16.3	not sign.	r=.37 p<.01
Subject number of turns	6.2	7.3	6.3	not sign.	r=.34 p<.02

Table 4-8. *Means and standard errors for the number of speaker switches and subject turns, corrected for stooge gaze, per first 5 session minutes.*

Effects of Gaze at the Subjects' Facial Region on Turntaking

Table 4-7 shows the data summary of the percentage of stooge gaze at the subjects' facial region during the first five session minutes. Numbers were obtained from double-blind video analysis by a research assistant (inter-observer reliabilities on a separate pilot sample, averaged across conditions using the Fisher *Z* transformation φ=.94, p<.001 for Stooge 1 and φ=.87, p<.001 for Stooge 2).

Analysis of variance showed differences in the mean percentage of stooge gaze to be significant across conditions ($F(2, 53)=112.05$, p<.0001). Post-hoc comparisons showed all conditions to differ significantly from one another in the percentage of stooge gaze (SNK, p<.05). Differences between conditions amounted to about a factor 2. This means subjects experienced about four times more stooge gaze in the still image with gaze direction condition than in the motion video with gaze direction condition. In the still image condition, whenever the stooges looked at the subject, they would *always* seem to look her in the eyes. In the other conditions, no such synchronization was present.

As we expected the amount of gaze at the subject's facial region to be a confounding variable we performed a covariance analysis, adjusting the mean number of speaker switches (assuming differences in Stooge 1's turntaking behaviour were due to differences in turntaking behaviour of the subjects) and subject turns for differences between conditions in the mean percentage of stooge gaze. All assumptions for covariance analysis were met.

Table 4-8 shows the data summary for the estimated number of speaker switches and subject turns during the first five session minutes, adjusted for percentage of stooge gaze at the subjects' facial region. With the effect of gaze at the subjects' facial region removed, analysis of covariance no longer showed significant differences across conditions with regard to the number of speaker switches (Roy Bargman Stepdown $F(2, 52)=.56$, $p=.58$) and subject turns (Roy Bargman Stepdown $F(2, 52)=.92$, $p=.41$).

There was a modest, but significant linear relationship between the percentage of stooge gaze at the facial region of subjects and the observed number of speaker switches (Pearson's $r=.37$, $p<.01$ 2-tailed) and between the percentage of stooge gaze at the facial region of subjects and the observed number of subject turns (Pearson's $r=.34$, $p<.02$ 2-tailed).

Question	Number of Positive Answers %			Anova
	Motion	Motion+Gaze	Still+Gaze	
It was easy to create those sentences with the three of us.	*35*	*32*	*47*	*not sign.*
The collaboration with the two partners was pleasant.	*80*	*84*	*100*	*not sign.*
This way of communication is pleasant.	*55*	*68*	*65*	*not sign.*
This communication system is easy to work with.	*80*	*79*	*82*	*not sign.*
It was always clear whom my partners were talking to.	*40*	*47*	*88*	*p<.005*
I could easily see what my partners were looking at.	*30*	*37*	*24*	*not sign.*

Table 4-9. *The number of positive answers per question per condition, in percentages of the total number of subjects per condition.*

Questionnaire

Table 4-9 shows the data summary for answers to the questionnaire. Numbers indicate the percentage of subjects who agreed with the stated question. Numbers were obtained by calculating the percentage of subjects who scored higher than 3 on a scale of 1 to 5 for each question (to avoid response set actual questions were formulated differently than those presented in Table 4-9, see Appendix B).

Analysis of variance (one-way Kruskal-Wallis) on the ranked response categories showed answers to only one question to be significantly different across conditions. Subjects rated the still image with gaze direction condition as superior to the other conditions with regard to the clarity with which they could observe whom their conversational partners were talking to ($\chi^2(2)=10.8$, $p<.005$).

DISCUSSION

In this section, possible explanations for our findings will be discussed for each dependent variable, relating results to our experimental predictions and to findings in literature. First, we will discuss the possibility of a confounding effect of stooge behaviour other than treatment. Next, we will discuss results with regard to the number of deictic verbal references, turntaking variables and task performance. We end this discussion by contrasting our objective measures with subjects' responses to questionnaires, and by discussing implications of our findings for the design of mediated systems.

Confounding Effects of Stooge Behaviour

For each dependent variable, we will now discuss to what extent results may have been due to differences between conditions in stooge behaviour other than the variables controlled for.

Deictic Verbal References

With regard to their use of deictic verbal references, we made no attempt to control stooge behaviour across conditions. One might have expected less deixis by stooges in the *motion video-only* condition, as it was less effective in this condition. However, on average, we found no significant differences between conditions in the number of deictic verbal references made by the stooges. This may have been due to two reasons:

1) Stooges did not see each other's video image. Therefore, they did not know whether they were looking at each other, making deixis between stooges equally difficult in all conditions. Although it may have been possible for them to observe each other's gaze at the reference point, the distance between them (of about 2 m) was inhibitive in this respect.
2) Qualitative analysis of video recordings shows the interaction pattern was typically V-shaped, with the subject in the role of mediator. Stooges were therefore more likely to depend on internal context for deixis, which did not differ between conditions.

We therefore believe it is likely subject behaviour with respect to this variable was caused directly by the experimental treatment.

Turntaking Variables

With regard to turntaking variables, subjects were, by definition, not completely independent of stooge turntaking behaviour. Although, in theory, each participant could have had an infinite number of turns, this is only possible if at least one other participant also had an infinite number of turns. This is because turns always involve a speaker switch. If we look at the distribution of turns between speakers in Table 4-5 on page 107, we see that most of the speaker switches occurred between subjects and Stooge 1. Although Stooge 2 demonstrated no treatment effect, like the subjects, Stooge 1 did have more turns in the *still image* condition (see Table 4-6 on page 108). One might

therefore suspect that treatment effects with regard to turntaking variables were in fact due to differences between conditions in the turntaking behaviour of Stooge 1. There are three reasons why we do not believe this was the case:

1) Qualitative analysis of video recordings seems to confirm that stooges left the initiative to the subjects as much as possible. Whenever subjects took the floor, stooges rewarded this by responding. This makes it likely that subjects influenced stooge turntaking behaviour, rather than vice versa. Since Stooge 1's role was to enter any answer given by subjects, he was more likely to be addressed by subjects than Stooge 2. This is in line with the opinion of one of our subjects: *"I watched the person who typed more than the other one."*
2) If stooges would have altered their turntaking behaviour across conditions, one would have expected the *motion video-only* condition to be the outlier, with a lower number of turns. Trends in the number of turns by Stooge 2, who was much less affected by subject behaviour, are in line with this rationale.
3) We found a linear relationship between the amount of stooge gaze at the subjects' facial region and the number of speaker switches and subject turns. Stooge 1's turntaking behaviour cannot have been *directly* affected by this confounding variable. One might argue that stooges may have had eye-contact in the still image condition. However, if this would have had a confounding effect, one would have expected a more equal distribution of turns between the two stooges in the *still image* condition. The number of individual turns per condition as listed in Table 4-6 on page 108 confirms the trend was in fact opposite.

We therefore believe it is likely that treatment effects with regard to the number of speaker switches and individual turns were in fact due to differences between conditions in the turntaking behaviour of the subjects.

Task Performance

With regard to task performance, we cannot exclude a potential dampening effect of the stooges on differences between conditions in the problem solving process. However, such dampening effect should also have reduced within-group variance on this variable. Qualitative analysis of video recordings seems to at least confirm this positive effect. When necessary, stooges prevented the subjects from misunderstanding the task situation, and their presence probably helped to reduce misconduct by subjects. Stooges did adhere strictly to plan in giving correct and incorrect answers themselves. We found no real evidence of any negative effects. However, it remains difficult to weigh the pros and cons of their presence with regard to task performance.

Deictic Verbal References and Amount of Speech

Results regarding the number of deictic verbal references to persons were in line with expectations. Subjects used twice as many references when gaze direction was conveyed. However, due to the availability of internal context and the relatively small size of the group, deixis did not disappear completely in the *motion video-only* condition. We did not correct our measurement for the amount of subject speech, as differences between conditions with regard to this variable were very small.

Our hypothesis was confirmed, stating that the presence of gaze directional cues in the form of head orientation causes the number of personal deictic verbal references used to rise significantly. The *actual* ability of subjects to use deixis towards the stooges did not differ between conditions (remember, in all conditions, the stooges saw a motion video image of the subject *with* head orientation). We therefore believe effects were due to differences in the subjects' own estimate of the effectiveness of head pointing. It is very likely that the subjects inferred this from treatment behaviour by the stooges, as subjects could not see their own image and believed the stooges used the same system.

Monk et. al. [99] suggest that the use of first-person and second-person pronouns is associated with the social context of the interaction at a semantical level of conversation. Reduced deixis might cause people to use more selfcentric expressions ("*I* think") instead of sociocentric expressions ("What do *you* think?"). Qualitative analysis of video recordings seems to confirm this. Subjects seemed less inclined to reach out towards the stooges when deixis was difficult. At a semantical level, this may have rendered conversations less sociable when gaze directional cues were not available.

Turntaking Variables and Simultaneous Speech

Results with regard to the turntaking variables ran contrary to expectations. We would have expected the *motion video-only* condition to be the outlier, showing less speaker switches and consequently less turns than the conditions in which gaze direction was conveyed. Instead, the *still image* condition was the outlier, with over 25% more speaker switches than both *motion video* conditions. Differences across conditions in the number of individual turns by subjects showed a similar trend. It is evident that the explanation for these results cannot lie in the treatment variable: the *absence* of nonverbal cues in the *still image* condition. Both literature and arguments presented on page 84 suggest that any *potential* effect of this treatment variable should have gone in the opposite direction, yielding less speaker switches when there are less nonverbal visual cues [13]. Although anonymity may have had a positive effect on turntaking in the *still image* condition, only one subject stated this in the questionnaire.

People are indeed capable of determining whom others are looking at with good accuracy. Von Cranach and Ellgring [157] reported that observers, located 1.5 m away and at right angles of the axis between two interactors, correctly identified more than 60% of the fixations by one interactor at the nose bridge of the other interactor as being inside the facial region. Given the extreme angle, they found observers relied mostly on head position rather than eye position. According to Argyle and Cook [8], when the observer is the other interactor, the accuracy is much greater, as observers can rely on eye positional information. Subjects, at a distance of 2 m from another person facing them, have been reported to judge 84% of fixations by that person at their nose bridge correctly as 'looking directly at me' [60]. Under very similar circumstances, Kruger and Hückstedt [89] found nearly all the gazes inside the facial region (at 6 cm from the nose bridge) were correctly identified as face-directed, while only 5% of gazes just outside the facial region (at 16 cm from the nose bridge) were misclassified as being inside. Jaspars et al. [82] and Cline [31] suggest that from a distance of about 1 m, people are able to discriminate the gaze position of someone facing them with an accuracy of approximately 1 cm in their facial plane (which relates to .6 degrees). We may conclude that people can determine quite well whom others are looking at, and that they are even better in determining whether they themselves are being looked at.

Box 4-2. *You can tell quite well who's face others look at, particularly when it is your face they look at.*

As the analysis of covariance demonstrates, a much more satisfactory explanation for our findings is the confounding influence of the large differences between conditions in the amount of stooge gaze at the facial region of subjects. The linear relationship between the amount of gaze at the facial region of the subject and turntaking variables was sufficiently strong to account for differences between conditions. We will now discuss possible explanations for this finding.

In the previous chapter, we saw that during multiparty face-to-face communication, visual attention for another person is usually designated at that person's facial region, particularly the eyes and mouth (this is supported by Argyle & Cook [8]). Even during periods of heavily gesticulation by a speaker, we observed that visual attention of the listener would mostly be aimed at that speaker's facial region, rather than his hands. If people use information about each other's gaze direction in the turntaking process, it therefore seems likely that gaze at each other's facial region is the parameter of interest. Box 4-2 shows that one is able to tell who's face others are looking at with considerable accuracy. However, the accuracy with which one can determine whether another person is looking at one's *own* facial region is even greater. People seem highly sensitive to gaze at their facial region. If viewing conditions are good, they can even discriminate which feature of their face is being looked at (i.e., the left eye, right eye, nose or mouth). It therefore seems likely we should regard our covariate not just as a measure for the amount visual attention received by subjects, but also as a measure for the ability of subjects to discriminate whether they *themselves* were being looked at. This yields two,

possibly complementary, explanations as to why there were more speaker switches in the *still image* condition:

E1 Improved Selfcentric Conversational Awareness. Rather than using head orientational cues to observe the dialogic attention of others towards others, subjects seemed interested mostly in determining whether others directed their auditory or articulatory attention towards *them*. This way, they could ascertain whether they *themselves* might be addressed or expected to speak. Difficulties in conveying gaze at the facial region therefore caused deficiencies in subjects obtaining and yielding the floor in both *motion video* conditions. This explanation is consistent with Kendon's findings [85] with regard to the function of gaze in (dyadic) turn synchronization and our findings from the previous chapter.

E2 Higher Level of Social Intimacy. As subjects received more visual attention in the *still image* condition, they felt more involved in the communication process due to a higher level of social intimacy. This is consistent with our findings from the previous chapter, where we discussed Argyle's Equilibrium of Intimacy. According to this theory, people use gaze at the facial region to regulate the level of intimacy, compensating for proximity, group size and other factors. When the level of intimacy is disturbed (either too high or too low), people feel uncomfortable [8]. In the *still image* condition, the average percentage of gaze at the subjects' facial region was almost exactly the percentage found by Exline [50] in triadic face-to-face conditions (given, on average, an approximately equal distribution of females and males over triads). In the other conditions, the percentages were much lower, yielding a much lower level of intimacy. As this lower level of intimacy could not easily be compensated by other means, subjects were less inclined to take the floor in these conditions.

Our findings with regard to the amount of simultaneous speech do not contradict the above explanations. The fact that we found no significant differences between conditions on this variable indicates either that our measurements were imprecise, or that interruptiveness was not significantly different across conditions. Assuming E1 and equal interruptiveness, this might indicate that in the *still image* condition, speaker switching was eased particularly during pauses. It seems likely that during pauses, gaze at the facial region is interpreted as an inclination to listen, effectively functioning as an offer to take the floor [85].

Although there were no significant differences between conditions in the duration of subject turns, the observed trends do not contradict our findings with regard to turn frequency. We believe that turn duration is simply a less accurate measure than turn frequency. Turn duration as a measure is sensitive not only to variance in the length of utterances themselves, but also to variance in the length of pauses between speakers.

Task Performance

We found no significant differences in task performance between conditions. With regard to this variable, the presence of gaze directional cues appeared to be as redundant as the presence of other nonverbal cues. Monk et al. [99] already suggested that measures of task performance are typically sensitive only to gross manipulations of experimental treatment. Our task may simply not have been as sensitive to experimental treatment as we had expected. Since we used a small group size, with each member having a clearly defined role, the availability of speech may simply have been enough to complete the task successfully. This does, however, suggest that effects with regard to other dependent variables were in fact due to differences between conditions in the communication process itself, rather than to differences in the nature of the experimental task. This effectively allows us to generalize our findings to any task situation in which the efficiency of the turntaking *itself* is the parameter of interest.

Qualitative Observations

With regard to our explanations for the mechanism behind the effect of gaze on turntaking, we found clear support for Explanation 1 (**E1**) in the questionnaire. Subjects found it easier to observe who was talking to whom in the *still image* condition. Although the reduced density of information in this condition may have been a contributing factor, it seems evident this was mostly due to the ease with which they could ascertain who was looking at *them* in this condition.

With respect to our *social intimacy* explanation (**E2**), questionnaire data was less convincing. Even so, trends in subject answers regarding the pleasantness of collaboration with the two partners do not seem to contradict this explanation. Although differences between conditions were not significant, subjects tended to favour the *still image* condition in this respect.

In general, suggestions made by subjects seemed in line with our findings. In their remarks, subjects characterized the *motion video with gaze direction* condition as "less sense of contact than face-to-face communication", while they tended to characterize the *motion video-only* condition as "better than telephony". The *still image* condition fell somewhere in-between, in this respect. Surprisingly, about 25% of respondents regarded the *motion video with gaze direction* condition as "just as if you were there." In general, however, people regarded all systems as a substitute to face-to-face meetings in case they could not travel. About 25% of respondents in the *motion-video only* condition stressed the value of face-to-face meetings, with only about 10% of respondents saying this in the *motion video with gaze direction* condition. Here too, the still image condition fell somewhere in-between.

About 20% of subjects in the *motion video-only* condition complained that "you do not see they are looking at you" or that "it was difficult to see when someone started talking". One subject from this condition said: "It was tricky to get used to using eye-contact — so you know who's looking at you."

Almost 30% of subjects in the *still image* condition complained either of a "lack of continuity in the images" or a "lack of facial expressions". One person in the *still image* condition remarked: "The most important missing element was a continuous image of the conversational partners. However, this wasn't strictly necessary for communicating in an acceptable manner, although it was difficult to combine body language and speech to get an impression of my partners."

About 10% of all subjects complained about the large angle between sets making it difficult to get an overview of both partners at once.

Qualitative observations of video recordings reveal surprisingly few breakdowns of conversation. However, the two observed examples of breakdown due to ambiguity of deixis were both in the *motion video-only* condition. Also, interaction seemed less lively in this condition.

Figure 4-5. *How reciprocal video tunnels allow participants to look into the camera while looking at each other.*

Implications for Design

We believe that a higher turn frequency is an indication of improved turn efficiency. As we have seen, most empirical studies seem to confirm this rationale. Although the effects of a higher turn efficiency may be highly dependent on the task situation, we believe that with regard to synchronous interactive multiparty communication one *can* generalize that mediated systems should preserve gaze at the facial region. The picture with regard to the conveyance of motion video is less clear.

Implications of Conveying Gaze at the Facial Region

We will first discuss the implications of the requirement to convey gaze at the facial region for the design of multiparty mediated systems. With regard to motion video systems, this requirement can be realized using *video tunnels*, which allow a camera to be placed behind the video image of a conversational partner [2, 24]. In Figure 4-5, we see the principal of operation. A half-silvered mirror is placed at a 45° angle between the camera and the video screen. This superimposes the image of the screen onto the image of the camera. Since the screen emits light, the camera is effectively hidden from view. Therefore, when video tunnels are used, gaze and mutual gaze at the facial region may be preserved. However, successful preservation of gaze in a mediated setting may depend on the following two design considerations:

1) The number of cameras used.
2) The number of video tunnels used.

We will start by discussing the first consideration. Our results suggest that with the effect of gaze removed, subjects in the *motion video-only* condition (which simulated a single camera setup) would not have had significantly less turns than subjects in the *motion video with gaze direction* condition (which simulated a multiple camera setup). However, there are two reasons why this observation cannot be generalized to a real-world situation:

1) As we have seen, people seem highly sensitive to the angle of gaze at their facial region. It is likely that in a single camera system, the angle between the camera lens and the representation of *each* participant on the screen should be prohibitively small in order to allow gaze at the facial region to be discerned by all participants*. This was not an issue in the *motion video-only* condition because stooges explicitly looked into the lens of the camera, and had no on-screen representation of the other stooge. One cannot expect the same in real-life situations.
2) In a single-camera system, participants cannot gaze at the facial region of a single *individual*. Any eye-contact will therefore always be with all participants at once. This is not prohibitive with regard to our second explanation for the effect of gaze on turntaking. If participants simply enjoy having the visual attention of others, it need not necessarily matter whether this attention is real or not. However, with regard to our first explanation (lack of Selfcentric Conversational Awareness), the use of a single camera system seems prohibitive. This may have been an issue in the *motion video-only* condition, although we found no *significant* effect. This was probably because our setup was not symmetrical, and because we used very small groups. However, in real-world situations it is very likely that the use of a single camera system *would* significantly affect turntaking efficiency due to a lack of Selfcentric Conversational Awareness.

The above rationale suggests that a multiple camera setup, in which each participant has a camera for each other participant, is required to effectively preserve gaze at the facial region.

With regard to our second consideration — the number of video tunnels required — the picture is less clear. It is of course possible to put multiple cameras inside a single video tunnel. What is required for this is that each window with a video image of a participant has a camera put behind it. By arranging windows and cameras such that a round table meeting is simulated, spatial orientation can even be preserved [26]. Whether or not multiple tunnels are required depends on whether head orientation needs to be conveyed, and if so, the angle at which such head orientation becomes discernable. Although we have not investigated these issues, we did find that the presence of head orientational cues has a significant effect on deixis. Although it may be possible that the presence of gaze at the facial region would suffice in providing external context for deixis using second-person pronouns, it seems reasonable to assume that in general, the presence of head orientational cues can be considered beneficial for achieving deixis. Thus, whether or not a single video tunnel would suffice depends on the minimum tunnel width required for head

* Note that this also means that even with video tunnels, gaze at their facial region may not be preserved if participants move their heads too much. Whether or not this is problematic may depend on the size of projection of participant on-screen representations, and their distance to screen. This problem will be further addressed in the *Future Directions* section, Chapter 6, page 163.

orientation to be used and clearly discernable, and the maximum tunnel width that can be achieved. When multiple video tunnels are used, the maximum width of the tunnel becomes irrelevant in this equation. If each participant uses one video tunnel for each other participant in the group (with each tunnel containing one camera and one video image), video tunnels can easily be arranged in any required angular setting [26]. Video tunnels are not a requirement when still pictures are used. As we have seen in the *still image* condition, a combination of participant snapshots at different angles of gaze can provide an effective simulation of gaze directional cues, including gaze at the facial region. The reason the use of video tunnels is not required here is because participants only need to look into the camera lens once, when their frontal picture is being taken. A great advantage of this is that one can easily ensure gaze at the facial region is *always* effectively preserved (see footnote on page 120). However, the use of a keyboard to select between images seems prohibitive. In Chapter 5, we will present a candidate solution based on eyetracking technology. When the mediated system can gauge whom you look at, it can automatically present the correct snapshot to the correct participant [152].

Implications for Conveying Motion Video

The situation with regard to motion video in multiparty mediated communication is rather more complex. We did not find a significant effect of the presence of visual cues typically conveyed by motion video on turntaking efficiency, or any other measure. However, because of other, more qualitative aspects, one cannot simply conclude one should therefore not convey motion video [129]. Although the presence of motion video does not seem a requirement in task situations which are not highly personal, this may be different in cases involving conflict, negotiation or other highly personal matters [99]. In such situations, most people would probably prefer face-to-face communication. But there are circumstances in which face-to-face contact is simply not possible. We therefore believe the choice for motion video should depend on the possibility of travel, user preference, task situation, and the availability of network bandwidth. However, one should realize that if gaze at the facial region is to be preserved, a single video stream per participant is insufficient. The requirement to convey gaze at the facial region may therefore be prohibitive with respect to the use of motion video, as it results in a number of *unique* video streams that increases with almost the square of the group size ($n^2 - n$, in which n is the number of participants). As will be further discussed in Chapter 5, this effectively eliminates any positive effect of compression techniques such as *Multicasting* on network bandwidth consumption [49, 154]. However, if bandwidth consumption is not an issue, one may simply choose to always provide motion video, as long as this does not mean gaze directional information is lost.

CONCLUSIONS AND DESIGN RECOMMENDATIONS

We will now present our conclusions as to the effect of a representation of gaze directional and other upper-torso visual cues on multiparty mediated communication. First, we will discuss the empirical conclusions, after which we will outline our recommendations for the design of multiparty mediated systems.

Empirical Conclusions

We found no significant effect of the presence of gaze directional cues on task performance. Similarly, we found no significant effect of the presence of other nonverbal upper-torso visual cues on task performance. Although this may indicate our task was not sensitive enough to experimental treatment, it does allow us to conclude that findings were not due to inherent differences between conditions in the nature of the experimental task.

We accept our hypothesis that the presence of gaze directional cues in the form of head orientation causes the number of deictic verbal references towards persons to rise significantly. Subjects used twice as many deictic second-person pronouns when head orientation was conveyed. We believe this was due to differences between conditions in the subjects' estimate of the effectiveness of head pointing in disambiguating deixis. Although we found no conclusive evidence for this, our qualitative observations suggest reduced deixis towards persons might cause conversations to become more selfcentric at a semantical level.

With regard to turntaking characteristics, we found that the presence of a representation of the visual attention of others towards *oneself,* in the form of gaze at the facial region, has a significant positive effect on turn frequency in multiparty mediated communication. We did not find a similar effect of the presence of other nonverbal upper-torso visual cues. There is a modest, but significant, linear relationship between the amount of gaze at the facial region of subjects and the number of subject turns (r=.34) and speaker switches (r=.37). Gaze at the facial region accounted for at least 12% of the variance in multiparty turntaking. Although our evidence is not fully conclusive in this respect, the potential increase in turn frequency may be in the order of 25% when gaze at the facial region is conveyed in a manner that preserves its temporal and spatial characteristics as observed in multiparty face-to-face communication. We believe such increased turn frequency is an indication of a more natural, and perhaps more efficient turntaking process. With regard to the effect of gaze at the facial region on this process, we have two, possibly complementary, explanations:

E1. Gaze at the facial region may improve *Selfcentric Conversational Awareness*. Subjects may have used gaze at their facial region to determine whether others might direct their auditory or articulatory attention towards them. This way, they could ascertain whether they might be addressed or expected to speak, for example, during pauses. We found clear support for this

explanation in the questionnaires. Subjects found it easier to observe who was talking to whom in the condition in which gaze at the facial region was best preserved.

E2. Gaze at the facial region may increase *Social Intimacy*. According to the Equilibrium of Intimacy theory, subjects in conditions with less gaze at their facial region may have felt uncomfortable and therefore less inclined to take the floor. Although trends in the questionnaires do not contradict this, we found no conclusive evidence for this explanation.

In general, we believe the signalling function of gaze at the facial region can best be described as a filter describing the identity (by means of its spatial characteristics) and extent (by means of its temporal characteristics) of activation of communication channels between conversational partners. This type of model will be further elaborated upon in the next chapter.

Design Recommendations

With respect to the design of mediated systems for multiparty communication, we formulate the following incremental requirements:

1) *Preservation of relative position.* Relative viewpoints of the participants should be based on a common reference point (e.g., around a shared workspace), providing basic support for the use of a common external context in deictic references.
2) *Preservation of head orientation.* Its representation eases the use of deictic references and may play a role in determining who is speaking to whom.
3) *Preservation of gaze at the facial region.* Allowing participants to gaze at each other's facial region eases turntaking: participants may find it easier to determine when they are addressed, or expected to speak. In addition, participants may find it easier to regulate the level of social intimacy. As such, gaze at the facial region may aid in providing a greater sense of telepresence.

Only with respect to the third requirement, is our evidence fully conclusive. The other two requirements should be regarded as preconditions for an optimal implementation of the third requirement.

If motion video is conveyed, we recommend the use of a multiple camera setup, in which each participant has a camera for each other participant. By putting each camera inside a video tunnel displaying the image of that participant, the above requirements can be implemented. This does, however, lead to a number of *unique* video streams that increases with almost the square of the group size (n^2-n, in which n is the number of participants). When still images are used, the above requirements can be implemented without the need for multiple video streams or video tunnels. When a mediated system can measure whom you look at, it can manipulate a pictorial representation of each participant such that the

above requirements are met. In the next chapter, we will present our candidate solution based on eyetracking technology.

With respect to multiparty turntaking efficiency, or any of our other dependent variables, the use of motion video does not seem a *requirement* in task situations which are not highly personal. In situations that *are*, it seems likely most people would opt for face-to-face communication instead of mediated communication. Even so, we believe the use of motion video should be at least optional. The choice for motion video may depend on the possibility of travel, individual preference, task situation and the availability of network bandwidth. If the latter is not an issue, one may simply choose to always provide it, as long as this does not mean gaze directional information is lost.

Chapter 5

GAZE: Mediating Attention in Multiparty Communication and Collaboration Tools

ABSTRACT

In this chapter, we show how functions of attention in communication and collaboration could be generalized into a comprehensive model of awareness functionality for multiparty mediated systems. We do this by defining different elements of awareness information in terms of attentive states of participants. Different kinds of awareness are distinguished: at Macro- and Micro-level, the latter consisting of Conversational Awareness and Workspace Awareness. We discuss a design rationale for conveying Micro-level attentive state information in groupware systems. As a network-friendly prototype implementation, we present the GAZE Groupware System, which uses advanced, desk-mounted eyetrackers to metaphorically convey the visual attention of participants in a 3D virtual meeting room, and within shared documents.

INTRODUCTION

As we have seen in the previous chapter, we defined *Conversational Awareness* as the knowledge conversational partners have about each other's dialogic attention during mediated multiparty communication (i.e., knowing who is talking or listening to whom). Given the evidence presented in those chapters, it does seem likely the process of turntaking in multiparty situations may benefit from the availability of such knowledge. There seems to be an interesting multi-modal relationship between different forms of attention in multiparty dialogue: if turntaking optimizes auditory attention (by, as a rule, allowing only a single person to speak [123]), and representations of visual attention optimize turntaking without interfering with dialogic attention, then the state of the attentive filter of others is used to fine-tune the attentive filter of oneself [105]. Indeed, participants seem particularly interested in Selfcentric Conversational Awareness: knowing whether others direct their dialogic attention towards *them*. They also seem to depend on the availability of such knowledge when referring to other participants (deixis). We have seen that Conversational Awareness information may be effectively conveyed by means of gaze directional cues, and that such information does not seem *highly* redundantly

coded by other cues*. In Chapter 2, we discussed how gaze directional cues may also be

* We should note that we did not investigate the use of gestural cues. However, as a generic indicator of Conversational Awareness they seem far less ubiquitous than gaze directional cues.

	Same Place (co–located)	***Different Place*** (distributed)
Same Time (synchronous)	face-to-face	telephone
Different Time (asynchronous)	post-it note	letter

Table 5-1. *The Time/Space matrix with sample media (adapted from Dix [40]).*

used to make participants aware who is working on what in a shared workspace (i.e., *Workspace Awareness* [68]). If provided with sufficient detail, gaze-related information may aid participants in referring to shared objects, sometimes as efficiently as gestural information [151]. More generally, we recognize the potential of representations of visual attention as a transparent and ubiquitous means for mediating awareness about other participants' attention for:

1) Persons;
2) Objects in a workspace;
3) The relation between these entities.

We feel *groupware* applications, which mediate the communicative and collaborative needs of individuals during computer supported group work, should provide support for awareness about other participants' state of attention which is broader and more integrated than suggested by the design recommendations presented in the previous chapter. Therefore, before implementing a system based on those recommendations, we believed it beneficial to first generalize the coding of attentional information into a broader conceptual framework for the design of awareness information in groupware applications.

In the remainder of this chapter, we will propose a design rationale for the systematic implementation of awareness features in groupware systems based on conveying the attention of others. We will then discuss our candidate implementation of a network-friendly groupware system which provides integrated support for awareness about conversational and workspace activities of others by monitoring and metaphorically representing the visual attention of participants.

DESIGN RATIONALE

Our aim was to design a groupware system with integrated and transparent support for awareness features. According to Ellis et al. [45], groupware systems are *"computer based systems that support groups of people engaged in a common task, providing an interface to a shared environment"*. A traditional classification of groupware environments is the time/space matrix shown in Table 5-1 [40]. In the time dimension, participants can either be working or communicating with each other at the same moment in time (synchronous), or at different moments in time (asynchronous). In the space dimension, participants can be working or communicating at the same place (co-located), or at different places throughout the world (distributed). In the latter case, a computer network (such as Internet) mediates and distributes their collaborative and communicative needs. In this project, we restricted ourselves to awareness support in *synchronous* use of *distributed* groupware systems.

Our design strategy was motivated by the following themes:

1) *Integrated Support for Conversational and Workspace Awareness.* As a main functional requirement, our system should provide a seamless integration between awareness about other participant's work activities and awareness about participant's communication activities [25]. This in order to prevent a plethora of user interface widgets for awareness support, each using its own metaphoric representation of awareness information [64, 65, 69, 132]. Our design recommendations regarding support for gaze directional cues (conveying relative position; head orientation and gaze at the facial region) to convey Conversational Awareness in video-mediated communication systems provide a paradigm on which such integration could be based.

2) *Implicit Collection of Awareness Information.* Rather than asking users to make explicit verbally or otherwise whom or what they are attending to, a clever monitoring of the spatial properties and timing of normal user behaviour (e.g., their system input) can provide a wealth of implicit information about their activities. We thus take a *noncommand* approach to providing awareness information, as discussed by Nielsen [107]. This should lead to a more transparent and efficient interface, with lower mental load and less interruption of task-oriented activities [128]. In order to accomplish this in a mediated setting we do not necessarily need intelligent systems. All that is required is a paradigm for monitoring input activity of individual users, and presenting this as awareness information to users on the other side of a network.

3) *Scalability of Networked Awareness Information.* As we have seen in the previous chapter, the use of gaze directional cues for conveying Conversational Awareness in video-mediated systems may lead to inefficient use of network resources, if motion video is conveyed. Since the purpose of groupware systems is to support many users, typically across a

computer network, scalability of awareness information should be seen as an essential technical requirement. With the exponential growth in Internet use came an exponential growth in traffic, worsened by an exponential growth in individual bandwidth requirements [29]. Although network paradigms such as Asynchronous Transfer Mode (ATM) [155] might solve network bandwidth problems in the future, we base our skepticism on the failure of ISDN to do so [154]. If we take bandwidth considerations seriously, the use of cameras to supply Conversational Awareness information seems limited, certainly with regard to bandwidth constraints of the current Internet.

4) *Representing Awareness Information With Natural Affordances.* According to Sohlenkamp and Chwelos [132], the design of the *system image* (i.e., the perceived aspects of a user interface [148]) of groupware applications should, where possible, be based on intuitions, knowledge and skills that people have acquired through years of shared work in the real world. Such knowledge may include the current Graphical User Interface (GUI) *Desktop Metaphor*, with its direct manipulation character [128]. As discussed, gaze directional cues may provide us with a suitable metaphor for conveying Conversational Awareness information. All that is required is an extension of this metaphor to the workspace, providing information about other users' relations to shared objects.

Integrating Support for Conversational and Workspace Awareness: Modelling the Attention of Others

According to McDaniel [97], awareness in groupware systems was traditionally defined as any information that answers the questions posed by the six "W" words: Who, What, When, Where, Why, and hoW. Research on awareness issues has been plagued by such fuzzy and high-level definitions. They led to confusion about what actually *constitutes* awareness information. We will therefore start by simply defining awareness as *"knowledge about the attention of other participants"*. On the basis of this definition, we will attempt to redefine most of the above elements of awareness information in terms of the time and place of the attention of other participants. Thus, we regard awareness in mediated communication and collaboration — we will use the term *communilaboration* for the intersection of these two — as constituted by a network of joint attention observations. In the following sections, we will outline a systematic mapping between awareness functionality and information about the attentive states of other participants. We will start by narrowing our focus to two complementary levels of awareness support: *Macro-level Awareness* dealing with aspects of the world *outside* synchronous distributed communilaboration, and *Micro-level Awareness* dealing with awareness aspects *during* synchronous distributed communilaboration. We note that these levels should not necessarily be regarded as a strict dichotomy.

Macro-Level Awareness

Macro-level Awareness are forms of awareness which convey background information about the activities of others prior to or outside of a meeting*. This relates to Greenberg's Informal Awareness [65] and Gaver's General Awareness [58]. Both are defined as *"…the general sense of who is around and what others are up to"*. Who is available for a meeting, what will the meeting be about, where, why and when will it take place and what tools will be used? Most of this information is rather discrete by nature. Often, small, low-frequency images [42] or activity indicators [65] can be used to sense the availability of persons for remote communilaboration. In providing Macro-level Awareness information, we stress the importance of using real snapshots for identification purposes. In addition, we suggest real names are provided, perhaps with pointers to web pages providing background information. Although we believe Macro-level Awareness *can* be seen as a discrete form of attentive state information, given the emphasis of this thesis on synchronous interactive communilaborative meetings we will limit our framework to concepts relating to its Micro-level counterpart only.

Micro-Level Awareness: Conversations in a Workspace

Micro-level Awareness is a form of awareness which gives online information about the activities of others *during* synchronous distributed communilaboration. This relates to the concept of Focused Collaboration Awareness discussed by Gaver [58]. Micro-level Awareness usually has a more continuous nature than its Macro-level counterpart. It consists of two categories: *Conversational Awareness* and *Workspace Awareness*. Conversational Awareness provides information about who is communicating with whom, Workspace Awareness provides information about who is working on what. Both imply a notion of space: in order to constitute these forms of awareness, one needs to know where 'who' is and where 'what' is. Together, they can provide information about who is talking to whom about what, thus providing an external context for deixis [40].

Defining Micro-Level Awareness in Terms of Attentive States

Gutwin and Greenberg [68] proposed a framework for Workspace Awareness according to a number of elements that play a role in this form of awareness. For each element, they considered the mechanisms people use to gather awareness information. We have adapted their framework to include Conversational Awareness, adding the element People (see Table 5-2). We also defined the different elements in terms of their relation to the *attentive states* of others. We define an attentive state as a description of someone's focus of attention during an activity. At a syntactical level this involves

* Note that this is not necessarily an asynchronous activity. Monitoring behaviour of others prior to a meeting can very well be implemented in a synchronous way.

		Attentive State	Element	Awareness Functionality	
				Workspace	Conversational
Syntax		Locus of Attention *(Spatial)*	Location	*Where are they working?*	*Where are the people they communicate with?*
		Attention Span *(Temporal)*	Presence	*Who is participating?*	
			Activity	*How actively are they working?*	*How actively are they communicating?*
Semantics	*Entity*	Attending to Objects	Objects	*What object are they using or referring to?*	
		Attending to People	People	*Whom do they work or communicate with?*	
	Action	Attending to Actions	Action	*What action are they performing or referring to?*	
Pragmatics		Attention Range	Extents	*What can they see?*	*What channels can they use?*
			Abilities	*What can they do?*	*Whom can they communicate with?*
			Influence	*Where can they make changes?*	*Where can they be?*
		Future Attention	Intention (them)	*What will they do next?*	*Whom will they communicate with next?*
			Expectations (me)	*What do they need me to do next?*	*Who wants to communicate with me next?*

Table 5-2. *Organizing elements of Micro-level awareness according to attentive state.*

describing the spatial and temporal properties of someone's (visual) attention, at a semantical level which actions, objects or people someone is attending to. For each element of Workspace Awareness, Gutwin and Greenberg described its functionality by listing questions that participants might ask themselves during shared activities. In Table 5-2, we did the same for Conversational Awareness. Some elements have shared functionality between Workspace and Conversational Awareness. These are represented in joint cells.

Our model is hierarchically organized in three levels: the ***syntax***, ***semantics*** and ***pragmatics*** of conveying awareness information in terms of attentive states. Each category of attentive state is attributed to one of these levels, and each element of awareness is attributed to a category of attentive state. At the ***syntax*** level there are two categories, the basic building blocks of our model. *Locus of Attention* describes the spatial aspects of attention, while *Attention Span* describes the temporal aspects of attention. All higher-level categories in our model can be expressed in terms of these space/time coordinates. The next, ***semantical*** level, is functionally the most important. It describes what *Actions*,

Objects and *People* other participants are attending to. It is subdivided into *Entity* and *Action*. Entity identifies which objects or persons users are attending to at a given time. Action describes how this relationship varies over time. Thus, actions are described by the dynamics of attending to entities.

Categories at the ***pragmatics*** level heuristically describe expectations about the spatial and temporal behaviour of others based on their history of attending to actions, objects and people. *Attention Range* relates to expectations in the spatial domain, while *Future Attention* relates to expectations in the temporal domain. Someone's Attention Range can be described by the spatial range of their history of attention to actions, objects and people, i.e., the space occupied by their behaviour. Someone's Future Attention can be described by the rhythms of their behaviour, based on a history of switching attention between actions, objects and people (e.g., their turntaking behaviour).

The above framework should be seen as a design model, or language, for conveying awareness in groupware applications. Our syntax, semantics and pragmatics are levels of this language, not of the actual communication process. Such a language would also be of use in the analysis of existing task situations. By monitoring the participant's locus of attention — the syntax of our language — one can determine which objects (or other participants) they are attending to. This allows one to make higher-level inferences about the semantics and pragmatics of their (joint) activities, such as what actions they actually perform.

In order to communicate awareness information, groupware systems should be able to collect information from individual users, and represent this information to other users across a network. We will now discuss how current and new input devices could be used to implement input of awareness information by means of a noncommand interface.

Implicit Collection of Awareness Information: Measuring Attention

We agree with Dourish and Belotti [41] that awareness information should be collected in a passive fashion, rather than being provided explicitly by participants. Nielsen [107] describes a completely new user interface paradigm based on this principle: *noncommand interfaces*. According to Nielsen, noncommand interfaces, like face-to-face conversations, rely on a more fuzzy dialogue between users and user interfaces than is the case with current user interface paradigms. In the noncommand paradigm, instead of a user issuing commands (by means of a command line syntax or by clicking menus or icons with a mouse), the computer *observes* user activity. The system then tries to make sense of available human input using a set of heuristics or a disambiguation process which could be similar to *grounding* in human dialogue [30]. Thus, computers would only need to query the user when certain information, required to understand what action should be taken, is deemed missing. We believe noncommand interfaces, if applied appropriately, can lead to a more transparent and efficient interface, with lower mental requirements and less interruption of task-oriented activities. By means of anticipation and

estimation, noncommand input may take us a step further towards the original goal of direct manipulation interfaces: the shifting of user attention from tool to task [128]. In order to accomplish this in a mediated setting, we do not necessarily need intelligent systems. All that is needed is a specification of what individual user activity should be monitored, and how this should be presented as awareness information to users on the other side of a network. If we follow our attentive state model, what we should monitor is the locus and temporal pattern of individual users' attention. Depending on the application, there are a number of ways in which such monitoring might be accomplished:

1) *Using Video Cameras.* A great benefit of video data for Micro-level Awareness purposes is its real-world and temporal nature. For example, video data may be very useful for conveying the attention span of others by means of their body movements, or the dynamic identity of real-world objects in the focus of other people's attention. A problem with video is that it can be difficult to achieve a seamless integration of spatial Conversational and Workspace Awareness properties [25, 110]. If the shared workspace is displayed on computer screens, then depending on the positions of computer screens and the representation of work spaces on those screens, angles of looking or gesturing may easily become incoherent with actual participant attention. The problem of achieving eye-contact using camera/display units, as discussed in the previous chapter, is an example of this problem. Another problem with video input is that the conversion of *generic* video images into a machine-readable format is still problematic. This problem may, for now, inhibit the use of such information by a noncommand interface for resolving decisions, e.g., about what awareness information to convey. A third problem with use of video data may be the heavy network bandwidth requirements, which we will discuss later.

2) *Using Microphones.* As we have seen in the previous chapters, speech activity can be an excellent predictor of turntaking patterns. As such, data from individual users' microphones might be used to gauge Conversational Awareness information. However, microphone data may need disambiguation before being useful as a provider of awareness information, or as input data in a noncommand decision process. Too literal an interpretation of such information, for example, when determining the locus of auditory attention in multiparty communication, may actually be detrimental to user performance (e.g., see Buxton et al. [26] for a discussion of problems with LiveWire voice-activated switching). Again, the temporal properties of audio data seem the most relevant. Microphone input could, for example, be used to monitor user presence or activity. Microphone input seems less appropriate for providing Workspace Awareness information. As for network constraints, audio data requires far less bandwidth than video data. In addition, we believe the availability of speech should be regarded a minimum requirement during synchronous mediated communi-laboration anyway [27].

3) *Using Manual Input Devices.* Manual input devices such as the mouse and keyboard are important means for gauging Micro-level Awareness information. In text-based environments, the duration and aim of keyboard input may provide Conversational Awareness information in ways similar to the above use of microphone input. Dialogic attention could then be represented by, e.g., font size of textual communication. In graphical user interfaces, a representation of the location of pointing devices within a shared workspace may be used to convey Workspace Awareness information. Many current-day groupware systems already provide such *telepointers* as an indication of the locus of participant activity [70]. Advantages of the use of manual input devices for providing awareness information include: they are low-cost and ubiquitous; they already are the main means of manipulating objects in shared workspaces; their data is machine-readable and low-bandwidth by nature. A disadvantage of manual pointing devices may be that they often do not return to a zero state [23]. If a participant leaves her mouse pointer at a position within a shared workspace, the telepointer representation may falsely indicate her attention to that part of the workspace. In the future, such problems might be circumvented by basing the decision to represent a telepointer on a fuzzy assessment of data from different input devices. We believe a more important restriction in the use of manual pointing devices is that they typically require an explicit manipulative action. Hence, they seem suitable mostly for gauging action-related awareness information, such as conveying the direct manipulation of shared objects. The use of manual pointing in providing Conversational Awareness information seems limited to manual deixis towards other participants.

4) *Using the Real World as an Input Device.* A recent development is the use of real-world objects, rather than software objects, as a user interface to software processes (so-called Tangible Media, see Ishii and Ullmer [80]). In this approach, the orientation and position of objects in the real world, e.g., on a desk, is gauged by means of sensors or simple image recognition techniques (for example, by recognizing barcode stickers on objects [145]). Attributes of real objects could thus provide low-bandwidth Workspace Awareness information to participants on the other side of a network, where they could be re-synthesized by projection onto their desk. The biggest advantage of this approach is the richness and transparency of the interface for single users. For now, the biggest drawback is that software manipulation of real-world artifacts is still limited. Thus, the joint manipulation of real objects may be problematic. Although we recognize the potential of this technique, we consider it beyond the scope of this thesis. In a related approach, data-suits and other forms of sensor technology may gauge a wide range of parameters of human behaviour in various forms of transparency, such as head or body orientation [22, 115, 167]. Eye and head orientation tracking are examples of such technology,

and given our findings in previous chapters, these techniques would seem the most relevant in this category for comprehensive gauging of attentive state information in general, and Conversational Awareness information in particular.

5) *Using Eye and Head Tracking Devices.* The orientation of the human eye or head can be gauged by tracking devices. Although at the moment, eye tracking technology is not yet used for generic input purposes, this is changing rapidly [83, 107]. As we have seen in Chapter 3, capturing the actual focus and span of visual attention by means of an eyetracking system may provide a relatively direct and high-resolution means of capturing information about participants' attention for actions, objects and people alike. Eye input may thus provide an integrated approach for gauging Conversational and Workspace Awareness information. In addition, eyetracker information is noncommand, machine-readable and low-bandwidth by nature [107]. Many problems with the application of eyetracking in user interfaces were in fact *due* to inadvertent use of eye fixation information for issuing system commands (the *"Midas Touch"* problem, see Velichkovksy et al. [150]). A clear disadvantage of eye input is that eyetracking devices are still rather expensive. However, this seems mostly due to the low production volume. Indeed, low-resolution eyetrackers are already becoming available for less than $1500 [158]. Unfortunately, eyetracking still has an undeserved negative reputation in terms of usability. Archaic requirements such as bulky head attachments or fixation of the user's head* need no longer apply. With up to 900 cm^3 of head movement tolerance, the transparent application of desk mounted eyetrackers for desktop computer input purposes has recently become a realistic option [5, 91]. It is with ranges larger than these that head orientation sensors become a good alternative, at least for gauging Conversational Awareness information [100, 115]. The inaccuracy of head orientation information would probably require an alternative source of input for the measurement of Workspace Awareness information.

Next, we will discuss the impact of the selection of input modality on network bandwidth requirements of a groupware system.

* The use of a headrest in the study presented in Chapter 3 was due to the accuracy requirements of scientific measurement.

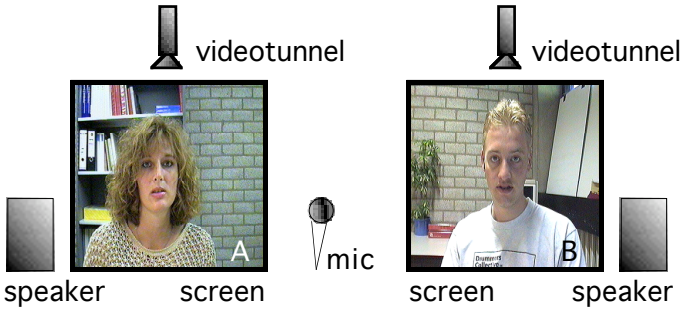


Figure 5-1. *Video tunnel setup for conveying Conversational Awareness with three participants. Each participant is represented by a camera/display unit.*

Scalability of Networked Awareness Information

Since the purpose of groupware systems is to support many users, in our case across a computer network, scalability of the network bandwidth consumed by awareness functionality is a technical design constraint that should be taken seriously [66, 154]. Since the focus of this thesis has been mostly on the provision of Conversational Awareness information, we will limit our discussion to a simple comparison between the impact on network resources of methods of input for capturing such information. From the above discussion, it becomes apparent that currently, the best candidates for gauging Conversational Awareness information are *1) video cameras* and *5) eye or head tracking devices*.

System 1: Using Video Cameras

As we have seen in the previous chapters, Conversational Awareness information should be effectively conveyed if information about the relative position, head orientation and gaze of individual users is mediated. When motion video is used, the design recommendations presented on page 123 advise the use of a multiple camera setup, such as the one depicted in Figure 5-1. Each participant has a camera/display setup for each other participant using the system. In between the camera and display of each unit, a half-silvered mirror is placed at an angle of 45 degrees. This *video tunnel* principle allows gaze at the facial region to be conveyed (see page 119 for a discussion) [2]. A good example of a mediated system using such setups is *MAJIC* [110].

In normal packet-switched networks (such as the Internet), the video from each camera and audio data from the microphone in the above system would need to be broadcast individually to each other participant in a meeting (rather like in a Cable TV network). *Multicasting* is a new Internet technique which prevents the inefficient use of the network bandwidth caused by such individual broadcasting techniques [49, 154]. In Multicasting, each unique stream of video data is put on the net only once, and is then picked up by the system of each other participant in the meeting (rather like a standard TV broadcast is picked

up by the TV antenna of viewers that are tuned in). Thus, using Multicasting, the total bandwidth consumption for System 1 would be equal to:

$$B = nA + n(n-1)V \qquad \textit{(Equation 5-1)}$$

In this equation, B is the total amount of bandwidth used, n is the number of participants, A is the amount of bandwidth per audio input, and V the amount of bandwidth used per unique video stream. It is clear that System 1 does not scale linearly with the number of participants. With four participants, 12 units of video bandwidth are required. With six participants, this rises to 30 units of video bandwidth.

System 2: Using Eye or Head Tracking Devices

When eye or head tracking devices are used, the manipulation of images of individual users would suffice to convey their visual attention towards other participants, as evidenced by the empirical study presented in Chapter 4. In such system, pictorial representations of users would be manipulated such that their relative positioning, head orientation and gaze would be preserved. This manipulation could occur according the measured locus of visual attention. The *still image* condition in our empirical study, discussed on page 90, is one example of such a system. The *Talking Heads* system by Negroponte [103, 104] is another example (see page 14). As the latter system evidenced, it may well be possible to convey motion video images using this type of system (see the *Future Directions* section in Chapter 6 on page 164 for a discussion). If motion video would be conveyed, the total bandwidth consumption B in a Multicast network would equal:

$$B \approx nA + nV \qquad \textit{(Equation 5-2)}$$

Since video is not used to convey visual attention, this system scales linearly with the number of participants. When motion video is not conveyed, V approaches zero, since all that is conveyed is the coordinates of visual attention of the participants. In that case, the amount of bandwidth needed is no more than would be required by audio only.

Concluding Remarks

Video input for conveying Conversational Awareness simply does not scale well with the number of participants. This, and the fact that the availability of a measure of visual attention should ease the integration of Conversational and Workspace Awareness information, led us to choosing System 2 with eyetracker input as a basis for our system design. Given the spatial range of available eyetracking devices, we, for now, limited our design to a desktop computer environment. Our empirical findings in Chapter 4 put the requirement for motion video as a channel for communication into perspective. We decided to initially use the *still image* condition in that experiment as a basis for our design: essentially an audio-mediated environment in which still images of

participants are manipulated in order to visually represent Conversational Awareness information. Next, we will discuss how we filled in our framework with concrete representations of Micro-level awareness information, given the above design constraints.

Representing Awareness Information With Natural Affordances: Designing a Virtual Meeting Room

In our discussion of the design of representations for groupware system functionality, we will concentrate on how Micro-level Awareness information could be represented in an audio-visual desktop computer environment. We will focus on the integral and synchronous provision of Conversational and Workspace Awareness information, rather than on the design of the communication and collaboration tools themselves.

This design theme relates to the *system image* (or "Look and Feel") of awareness functionality, guiding how our framework could be used to render the perceived aspects of a groupware system user interface [148]. In doing so, we wanted to make use of existing knowledge and skills of users as much as possible. We therefore chose a metaphoric design approach, in which elements of the interface and their behaviour would be based on real world equivalents as much as possible. We tried to use Gibsonian affordances [59, 108] to render awareness functionality into the user's perception as directly as possible. Thus, we tried to allow users to rely as much possible on 'knowledge' in the system image, rather than on knowledge in their heads [166]. We agree with Sohlenkamp and Chwelos [132] that a metaphoric design approach should not be followed too rigidly in order to prevent the inadvertent modelling of limitations inherent to the real world. Instead of modelling the real world on a one-to-one basis, we therefore attempted to model the essential bits only (see *Conveying the Right Cues*, Chapter 2, page 6). Finding a basis for our representations in the real world included making use of users' knowledge of current Graphical User Interface *Desktop Metaphor* [131]. In the design of their DIVA groupware system functionality, Sohlenkamp and Chwelos [132] simply expanded the single-user desktop paradigm to include multiple users, adding the elements *people* and *rooms* to elements already present in the desktop paradigm: *documents*, *desks* and *pointers* (with the latter becoming telepointers). Thus, they padded an existing computer metaphor with elements borrowed from the real world. In order to achieve a seamless integration of representations for Conversational and Workspace Awareness information (our main functional requirement), different elements of the system image should have some form of spatial and temporal relation with each other. We therefore followed an approach similar to Sohlenkamp and Chwelos, building a *virtual meeting room* in which the above user interface elements are jointly represented [48]. However, as we will now discuss, a virtual meeting room alone may not be sufficient for supporting focused collaboration.

General Communilaboration versus Focused Collaboration: WYSIWIS?

Stefik et al. [137] proposed the *"What You See Is What I See"* (WYSIWIS) paradigm as a means of providing a consistent and coordinated display of user interface elements to all participants. In strict WYSIWIS, all participants essentially have exactly the same display containing exactly the same information at exactly the same moment in time. In their Colab environment [136, 137], Stefik et al. supported WYSIWIS by maintaining synchronized views, and by offering facilities for telepointing with publicly visible cursors. This would allow participants to have a common understanding of their virtual world, permitting them to rely on the availability of external context in, for example, deixis. However, in [136], Stefik et al. pointed out that a strict application of WYSIWIS throughout user interface elements may be too inflexible. It may, for example, lead to problems in supporting the parallel work on different tasks by subgroups, or the transfer of information between private and public spaces. Instead, they recommended strict WYSIWIS as a foundational abstraction, with a selective easing of compliance (relaxed-WYSIWIS) along four dimensions:

1) *Display space.* Strict WYSIWIS applies to everything on an individual display; applying it only to a subset of visible objects (e.g., windows and cursors) relaxes this constraint.
2) *Time of display.* Strict WYSIWIS requires that images be synchronized; allowing delays in updating or viewing of images relaxes this constraint.
3) *Subgroup population.* Strict WYSIWIS requires shared viewing to apply to everyone in the full meeting groups; allowing sharing to be limited to subgroups relaxes this constraint.
4) *Congruence of view.* Strict WYSIWIS requires that images be identical; allowing alternative views relaxes this constraint.

As a response to this, Gutwin et al. [70] warned that relaxed-WYSIWIS may actually lead to a lack of awareness, since increased individual control reduces the group focus inherent in strict WYSIWIS systems. Thus, there may be a conflict between requirements for general and focused collaboration. We therefore decided to have a more strict environment for general group activities, and a more relaxed environment for focused collaboration activity. The virtual meeting room would provide a place on the display where general group activity is grounded in a rather strict manner. Within it, only congruence of view would be relaxed, and only in that each participant's viewpoint would be strictly located at the position of his representation. This way, we ensured compliance with our design recommendations (preserving relative position; head orientation and gaze), allowing effective use of gaze as a metaphor for conveying Conversational Awareness information. Individual viewpoints would otherwise be fixed such that all awareness information would be within field of view of all participants. On the rest of the display, around the virtual meeting room, focused collaboration could take place using task-specific relaxed-

WYSIWIS document editors* (e.g., those proposed by Greenberg and Gutwin [64, 69]). Our WYSIWIS relaxation requirements for focused document editing were based on recommendations made by Baecker et al. [14], and almost the opposite of those of the virtual meeting room. Different documents may appear at different locations on displays of individual participants; updating of images may depend on where individuals work within the document; document contents need be displayed only to the subgroup working on them; document contents is typically viewed from the same angle by all participants (e.g., during text editing), but position of individual users within documents (i.e., the part of the document that is displayed) is totally relaxed. However, in order to provide a common glue between document editors (focused collaboration tools) and the virtual meeting room (the general communilaboration tool) we introduced one constraint: there should always be at least one WYSIWIS representation of a telepointer linking the attention of a participant in the meeting room to his attention to sections of document content.

Attentional Focus as an Organizational Metaphor

In a larger perspective, the concept of *attentional focus* can be regarded as an organizational metaphor throughout the design, gluing representations of awareness functionality at different levels of refinement together so that they can be recognized as a whole [161]. This becomes apparent when we consider the suggested user interface elements as a ways of representing attention of participants. *Rooms* are ways of organizing the presence of people, signalling the general availability of their attention for a common communilaborative goal. Within rooms, the co-location and orientation of *persons* is a way of organizing the joint attention of sub-groups towards a common communilaborative task [102]. *Desks* are ways of organizing task-specific objects, providing an overview of their availability for collaborative attention. Within desks, *documents* signal the availability of a task as a focus for collaborative attention. Also within desks, *telepointers* signal the actual focus of collaborative attention towards a certain task. Finally, when documents are opened, relaxed WYSIWIS document editors allow participants to focus their attention according to individual interest. As will be discussed, telepointers link this focus within the document to the focus within in the virtual meeting room (as an example, see the Gestalt view of the SASSE environment [14]). Thus, using the above organizational metaphors, the focus of attention of participants may be described and guided from the very general to the very specific.

* The space around the meeting room is, of course, also available for local activities.

Visual Representation of User Interface Elements

We will now discuss how the above discussed user interface elements could be rendered with visual behaviour. To keep the user interface as simple as possible, we chose to represent only five elements of real meeting rooms: *rooms*, *desks*, *persons*, *documents,* and *pointer light spots*. Other attributes include a stationary pad, exit sign, and a trash can.

1) *Rooms.* Rooms contain all people, desks and documents required for a synchronous distributed communilaboration session. Depending on the environment, rooms could be represented by text windows (e.g., in chat environments), 2D surfaces (e.g., DIVA [132]), or 3D worlds (e.g., MASSIVE [66]). As will be discussed later, we wanted to use head orientation as a metaphor for conveying visual attention of participants. If this information was to be used for conveying Conversational Awareness as well as Workspace Awareness, 3D orientation would seem a requirement. We therefore chose a 3D room design, which could function as a container for organizing the attention of participants at the *presence* level. Just like people located in the same real room are able to see and hear each other, so too would people within a our virtual room hear and see each other. As with DIVA [132], the entering of a person into a room could establish audio-visual communication links with people already present. We restricted ourselves to using rooms as a means for organizing private meetings only (i.e., a virtual meeting room [48]).
2) *Desks.* These containers represent a way of organizing attention of participants towards any number of collaborative objects. Depending on the environment, a simple representation of a directory structure might be used for grouping shared files [4]. Like DIVA [132], however, we chose the single-user desktop metaphor as a basis, expanding it into a shared surface onto which iconic representations of shared file objects could be placed and organized by position [131]. However, our representation would function not just as a means for organizing collaborative objects, but also as a means of organizing persons. By placing representations of persons around a 2D desk surface in our 3D meeting room, face-to-face round-table communilaboration could be used as a metaphor for integrating Conversational and Workspace Awareness information.
3) *Persons.* A participant is represented by a *persona*: a metaphoric rendering of real participant behaviour [90, 104]. As we have seen, an important functional requirement for personas is that they represent a participant's visual attention. Although, depending on the environment, personas may be rendered by a name (chat environments), a 3D model (avatar environments), or a video stream (video conferencing), this rendering would need to include a visual representation of real participant attention towards other persons. Gaze may be considered an ideal metaphor for this purpose. Based on our findings, we consider real images of participant gaze as the most

effective way of conveying this. As discussed, we based our design on still images, rather than motion video images. In order to achieve a smooth integration of the persona in the 3D meeting room, we decided against the use of different images for conveying different loci of visual attention, as was the case in the *still image* condition in Chapter 4. Instead, for each participant, we suspended a single frontal snapshot — made while looking into the camera lens — in the 3D meeting room. 3D orientation of this 2D persona would then metaphorically convey the direction of gaze of that participant, as measured by the eyetracking device.

4) *Documents.* These containers represent a way of organizing attention of participants towards a particular task. In standard desktop environments, document icons typically function as a representation of associated document content [131]. Although, depending on the environment, such content may be directly presented as a computer file, we chose to follow the desktop paradigm. Document icons can be placed on a desk in the virtual meeting room as a means of sharing the associated contents. This content can be accessed by opening the document icon (e.g., by double-clicking it), at which moment it is downloaded and displayed in a focused collaboration editor outside the virtual meeting room. Document editors appear only to those participants that opened the document, but the associated document icon on the desk remains visible to all. Documents can be associated with local editors, or editor software could be embedded as part of the document content. In principle, documents can contain any kind of information, as long as an associated editor is available to all parties. As discussed, information display in such editors would typically be based on all participants having the same point of view, but should otherwise follow a relaxed-WYSIWIS paradigm. As discussed, telepointers provide ways of linking the focus of attention of individuals on sections of document content to the document representation in the virtual meeting room.

5) *Telepointer Light Spots.* These represent the actual attention of participants for objects in a shared work space. During presentations by an individual in a group meeting, light spots produced by laser pointing devices are now widely used to communicate the exact focus of attention of a presenter. We used these light spots as a metaphor for telepointing, illuminating objects in a shared workspace according to the attention of individual participants. As a source of information about this attention, we could use mouse position or the actual point of gaze as provided by the eyetracker. With the latter, participants need not take any action other than looking to provide others with Workspace Awareness information. During general communilaboration in the meeting room, when a participant looks at a location on a desk, a light — appearing to be emitted from his persona — illuminates the spot. We thus borrowed a functional metaphor from the helmets used by miners to illuminate their work environment (the *Miner's Helmet* metaphor). The light spot is also associated with the emitting

persona by means of colour coding. Multiple light spots of the same colour are used to represent the same focus of attention at different levels of refinement. During focused collaboration in a document editor, when a collaborator looks at a location within the document content, a light with his colour illuminates the spot. This light spot is visible only to persons working within the document. Therefore, whenever a person looks at document content, the associated document icon in the meeting room should also be illuminated by a light spot of his colour. This light spot is visible to all. If documents contain multiple sections, multiple light spots of the same colour could indicate in a strict-WYSIWIS fashion which section each collaborator is focusing on (see the Gestalt viewer of the SASSE environment [14] as an example of how this might be accomplished). All light spots generated by a single persona should remain tightly associated by movement and colour. This way, light spots may provide a kind of attentional glue between focused collaboration and general communilaboration activities.

Visual Contributions to Micro-Level Awareness

In Table 5-3, the functionality of the above visual representations in providing Micro-level Awareness information is summarized. The orientation of the persona and the location of the corresponding light spot convey the spatial aspects of someone's visual attention. From the movements of personas and light spots, people can see whether their partners are actually present, and if so, how actively they are working and communicating. These spatial and temporal aspects of awareness also provide valuable cues for inferring attentive states at the semantical level. People working together on an object may have their personas oriented towards the location of this object and their light spots hovering around the object, thus conveying Workspace Awareness information. People speaking to each other may have their personas oriented towards each other, thus conveying Conversational Awareness information. Actions can be inferred through the dynamic interactive behaviour of light spots, objects and personas. Attention Range and Future Attention can be inferred through the spatial and temporal patterns found in a history of such behaviour.

We confined ourselves to representing explicitly only the spatial aspects of attentive states at the syntax and entity levels (at any given moment in time). All higher-level inferences about these representations are left to the user's interpretation. This does not mean that our framework would not allow explicit representation of higher-level attentive states. For example, one could implement Attention Range explicitly by translucently colouring parts of space where users have done things, or by altering the beam size of light spots.

		Attentive State	Workspace Awareness	Conversational Awareness
Syntax		Locus of Attention *(Spatial)*	*Location of light spots*	*Location and orientation of Persona*
		Attention Span *(Temporal)*	*Presence of light spots* *Dynamics of light spots*	*Presence of Persona* *Dynamics of orientation*
Semantics	*Entity*	Attending to Objects	*Position of objects* *Position of light spots on objects*	*Orientation towards objects*
		Attending to People	*Joint light spot positions* *Joint orientation towards an object*	*Orientation towards other Persona*
	Action	Attending to Actions	*Dynamics of attending to objects*	*Dynamics of attending to people*
Pragmatics		Attention Range	*Spatial patterns in the dynamics of attending to objects*	*Spatial patterns in the dynamics of attending to people*
		Future Attention	*Temporal patterns in the dynamics of attending to objects*	*Temporal patterns in the dynamics of attending to people*

Table 5-3. *Representing elements of Micro-level Awareness visually according to the attentive state of participants.*

Auditory Contributions to Micro-Level Awareness

As exemplified by the human turntaking mechanism, the presence and level of speech activity are important elements of the attentional relationship between persons. Together with the semantics of speech communication, level and position of speech sources, and their spatial co-location with other sources of information, may be considered as parameters for constituting or conveying dialogic attention [21, 28, 38, 101]. As such, we can regard these parameters as potential providers of awareness information in mediated situations. For example, the spatial co-location of speech and persona could provide Conversational Awareness information about who is speaking. In recent years, we have seen an increase in the use of parameters such as proximity, position and orientation of personas as a means of controlling auditory Conversational Awareness in groupware systems [66, 102]. In their paper on the MASSIVE project, Greenhalgh and Benford [66] suggest a generic spatial model for managing communication of participants in large virtual worlds. They propose the use of regions around personas (so-called *auras*) as a way of determining what connections should be made. In their system, the spatial collision of auras of different participants triggers the connection of audio channels between them. Once communication is established, they suggest a similar spatial model

for managing further participant awareness. Amongst other things, they propose letting participants define a shape, typically a cone, around their persona which describes their *focus* of attention. The amount of overlap between focus regions of participants could then be used to control the mutual volume of the audio connections between them, or the level of detail of a graphical rendering. The MASSIVE system provides a means of controlling aura and focus attributes by simple selection of preset shapes. Given the evidence presented in earlier chapters, however, measures of the actual focus of visual attention may be a much more dynamic and transparent way of providing such information. For example, statistics of joint visual attention, like the percentage of time spent looking at each other's persona, could be used in our virtual meeting room to control the level of audio connections between participants. This might aid participants in maintaining their attention during side conversations, when multiple speakers are active. For example, audio of visually attended speakers could be provided with higher quality, while distracting sounds made by unattended individuals might be attenuated. This *might* be implemented by a gradual low-pass filtering of audio between participants that do not look at each other or the same objects for a certain length or percentage of time. The quality of transmission service for motion video personas could be controlled in a similar manner [66]. Following a model of the retina, the foveated persona could be rendered with more visual detail than other personas [43]. As with audio, video images of attended persons could thus be of higher quality, while visual distraction by information in peripheral vision might be reduced. Following the above scheme, network bandwidth need not be sacrificed to achieve a higher quality of service. Indeed, network as well as computing resources might be allocated more efficiently, guided by individual interest. For example, when Multicasting (see page 136) is not used, the sample rate of audio, and the colour depth or resolution of video images could be decreased on an individual basis, greatly reducing the impact of individual data streams on network traffic. However, little is known as to the effect of the above provisions on usability. We therefore considered the above issues as beyond the scope of this thesis, recommending further investigation instead.

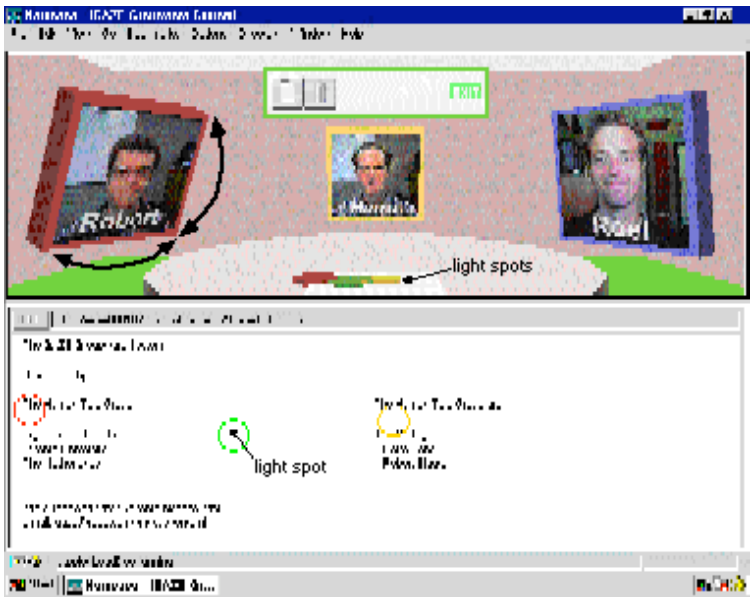


Figure 5-2. *The GAZE virtual meeting room (top) with shared document editor (below). In the meeting room, personas rotate according to where users look. Light spots convey where users look within shared work spaces.*

Figure 5-3. *Participant using the GAZE Groupware System. The camera on top of the monitor is used for updating snapshots of the participant's persona. The black camera below the monitor is the infrared eyetracking camera. Its image is displayed on the small monitor as an example only.*

A PROTOTYPE: THE GAZE GROUPWARE SYSTEM

Based on the above rationale, we developed a prototype groupware system which provides integral support for Conversational and Workspace Awareness by conveying the participants' visual attention. Instead of using multiple streams of video for this purpose, the GAZE Groupware System (GGS) measures directly where each participant looks by means of an advanced desk-mounted eyetracking system. The system represents this information metaphorically in a 3D virtual meeting room and within shared documents. The system does this using the Sony Community Place [133] plug-in, which allows interactive 3D scenes to be shared on a web page using a standard multiplatform browser such as Netscape. In this prototype, we did not yet integrate support for multiparty audio communication. Instead, the GAZE Groupware System can be used in conjunction with any multiparty speech communication facility such as an Internet-based audio conferencing tool, or standard telephony.

A Session in the GAZE Virtual Meeting Room

The GAZE Groupware System simulates a four-way round-table meeting by placing a 2D image (or persona) of each participant around a desk in a virtual room, at a position that would otherwise be held by that remote participant. Using this technique, each person is presented with a unique view of each remote participant, and that view emanates from a distinct location in space. Each persona rotates around its own x and y axes in 3D space, according to where the corresponding participant looks. Figure 5-3 shows the system in use in a four-way situation. When Robert looks at Roel, Roel sees Robert's persona turn to face him. When Robert looks at Harro, Roel sees Robert's persona turn towards Harro. This should effectively convey whom each participant is listening or speaking to. When a participant looks at the shared desk, a light spot is projected onto the surface of the desk, in line with her persona's orientation. The colour of this light spot is identical to the colour of her persona. This allows a participant to see exactly where the others are looking within the shared workspace. By direct manipulation, e.g., with their mouse, participants can put document icons, representing shared files, on the desk. Whenever a participant looks at a document icon or within the associated file, her light spot is projected onto that document icon. This allows people to use deictic references for referring to documents (e.g., *"Here, look at these notes"*). Shared documents are opened by double clicking their icon on the desk. When a document is opened, the associated file contents appears in a separate frame of the web page (see Figure 5-2). In this frame, an editor associated with the file runs as an applet. When a participant looks within a file, all participants looking inside that file can see a light spot with her colour projected over the contents. This light spot shows exactly what this person is reading. Again, this allows people to use deictic references for referring to objects within files (e.g., *"I cannot figure this out"*). We realize, that providing such

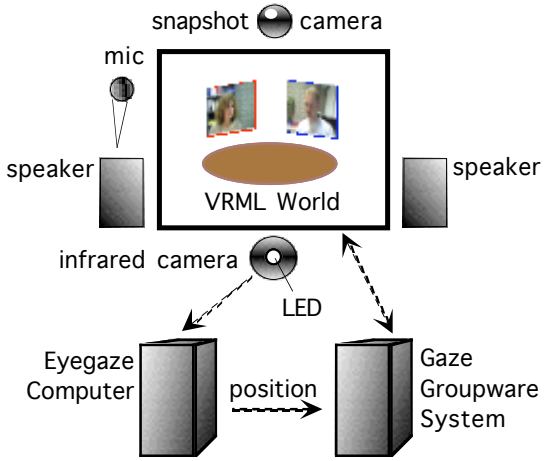


Figure 5-4. *The GAZE Groupware System hardware setup.*

information may invade the privacy of individual users. By (annoyingly) projecting their own gaze position whenever it is shared, we hope to ensure that individuals are aware their gaze position is transferred to others [76]. Although files can be referred to by URL, in the current prototype they are still restricted to ASCII text only, and cannot yet be edited.

Hardware Setup

Each participant has a hardware setup similar to the one shown in Figure 5-4. The GAZE Groupware System consists of two key components: the Eyegaze system, which determines where the participant looks; and the GGS computer, a Windows '95 Pentium running Netscape, the GAZE Groupware System, a web server, frame grabbing software and an Internet-based audio conferencing tool. The Eyegaze system, which is discussed in detail below, reports the gaze position of the user over a serial link to the GGS computer. The GGS computer determines where the participant looks, manipulates her persona and light spot, and conveys this information through a TCP/IP connection via a server to the other GGS computers. The Eyegaze system is not required. Participants can also switch to using their mouse to indicate point of interest. The video conferencing camera on top of the monitor is currently used to make snapshots for the persona (future versions might incorporate motion video). When making a snapshot, it is important that users look into the video conferencing camera lens, as this will allow them to achieve a sense of eye-contact during meetings.

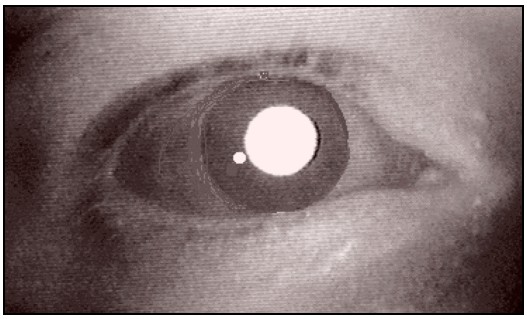


Figure 5-5. *The eyetracker infrared camera image.*

The LC Technologies Eyegaze System

When the eye remains relatively still for more than about 120 milliseconds, we speak of a fixation (see Chapter 3, page 23). For determining where the user is looking, it is these fixation points that we are interested in. Our system measures the eye fixation points of a user by means of the Eyegaze System [91], an advanced, desk-mounted, imaging eyetracker with a spatial resolution of approximately .5 degrees of arc and a temporal resolution of 50-60 Hz. The Eyegaze system consists of a 486 computer processing the images of a high-resolution infrared video camera. This camera unit is mounted underneath the screen of the user (see Figure 5-4), and is aimed at one of his eyes (see Figure 5-3 on page 146). On top of the camera lens, an infrared light source is mounted which projects invisible light into the eye. This infrared light is reflected by the retina, causing a bright pupil effect (the large circle in Figure 5-5) on the camera image. The light is also reflected by the cornea of the eye, causing a small glint to appear on the camera image (the small dot in Figure 5-5). Because the cornea is approximately spherical, when the eye moves, the corneal reflection remains roughly at the same position. However, the bright pupil moves with the eye. By processing the image on the computer unit, the vector between the center of the pupil and the corneal reflection can be determined. In order to correctly translate this vector into screen coordinates, the user needs to calibrate the Eyegaze system once before use. This calibration procedure takes about 15 seconds. When the coordinate remains within a specified range for approximately 120 ms (3 complete camera frames), the Eyegaze system decides that this is a fixation. It then starts reporting the coordinates over a serial link to the GAZE Groupware System running on a separate computer (see Figure 5-4). The GAZE Groupware System uses this coordinate to determine at which object or participant on the screen the user is looking.

Software Implementation

The GAZE Groupware System was implemented using the Virtual Reality Modeling Language 2.0 [130]. This cross-platform standard separates 3D graphic descriptions (rendered natively) from their dynamic behaviour (running on a JAVA Virtual Machine). Sony Community Place [133] is a plug-in for Netscape which implements the VRML 2 standard and adds a client-server architecture for sharing 3D graphics and behaviour over TCP/IP. For each dynamic object a user sees in the virtual meeting room, there is a corresponding JAVA object. Whenever such an object does something, its behaviour is broadcast via the Community Place Server by means of messages to the other systems. This way, all participants' copies of the meeting room are kept in sync. Eyetracker input is obtained from a small native driver application polling the serial port or the mouse. Document editors are JAVA applets running separately from the VRML world, although they do communicate with it to obtain eyetracking data and URLs. All code, graphics, and documents are shared using web servers running on each GGS computer.

Evaluation of the System

Informal sessions with several hundred novice users at ACM Expo'97 indicated our approach to be a promising one. Most participants seemed to easily interpret the underlying metaphors, particularly those related to Conversational Awareness. The eyetracking technology was, in many cases, completely transparent. Users would sit behind the system and immediately start chatting, without calibration or instruction. Although we empirically evaluated some of the underlying assumptions of the system (as discussed earlier), the prototype has not yet been tested for usability. Indeed, we should regard the current system as a first attempt to implement the sketched principle of conveying attention-related information in groupware, rather than as a finished product. Many details have yet to be filled in. As such, we identified a number of unanswered questions with regard to the usability of the current prototype:

- *Skewed projection as a metaphor for gaze direction.* Our metaphor for representing gaze direction is based on skewed projections of a 2D image on which the user is depicted looking into the camera. Although we found no evidence for this in our demonstrations, skewed projections of pictures may evoke a sense of eye-contact when it is not intended as such (rather like watching the TV News from an angle).
- *No spatial separation of audio or visual encoding of speech activity.* Although it is not necessary for audio sources to be exactly co-located with the visual representation of users, spatial separation of their voices may ease selective listening [28, 125, 127]. This feature is not yet integrated into the prototype. Users currently need to depend on auditory discrimination of voices for identifying the source of individual speech activity. Spatial separation of audio and visual encoding of speech activity (using the subtle

animation techniques demonstrated in our video simulation [153] or by using motion video) may solve this issue.

- *No option for motion video.* Although the requirement for motion video seems to largely depend on the task situation and availability of network resources, we hope to include streaming video as an option when VRML allows this.
- *Colour coding and light spot confusion.* Light spots can only be attributed to a persona by colour and synchronized movement. We would like to devise a more redundant coding scheme. When there are many light spots, novices may get confused or distracted [136]. It should at the very least be possible to turn light spots off.
- *Privacy.* Although knowing what others are reading may be beneficial during a joint editing process, there are many task situations where this could be detrimental. Users should always be aware when their gaze is being transmitted, and when not. Currently, we hope to ensure this by (annoyingly) projecting the user's own gaze whenever she looks at shared objects. This is not a satisfactory solution [76].
- *Eyetracker limitations.* Although the eyetracker works well while talking, head motion is still limited to about 5-10 cm in each direction. However, if the eye moves out of range, the eyetracker resumes normal operation as soon as it is back in range. A version of the Eyegaze System which allows 30 cm of head movement in each direction will be released shortly. Other systems already provide such ranges [5]. Although the eyetracker works fine with most glasses and contact lenses, a small percentage of users has problems with calibration. Eyetracking is still expensive, but current developments lead towards eyetrackers which are just another input device: inexpensive and transparent in use.
- *Meeting room restrictions.* Although this is not an intrinsic limitation, the system currently allows only four users in the meeting room. Users are currently not allowed to move freely through space, or control their point of view. We are not sure to what extent this should be allowed.

CONCLUSIONS

In this chapter, we have shown how many of the interpersonal awareness aspects in synchronous interactive distributed communilaboration, particularly those on a Micro-level, can be described in terms of conveying the attentive state of others. We defined an attentive state as a description of someone's focus of attention during an activity. At a syntactical level this involves describing the spatial and temporal properties of someone's (visual) attention, at a semantical level which actions, objects or people someone is attending to. Our Attentive State model of awareness allows groupware designers to conceptualize in a more structured way the kinds of features they need to convey. It provides a way of thinking about capturing awareness information using direct yet transparent means and representing it across modalities using attention-based affordances. Our model is by no means exhaustive or complete. We consider it a simple reference framework which can be applied to a wide variety of situated communilaboration. As for our application, the GAZE Groupware System, we demonstrated how our framework may lead to improved awareness features without requiring any explicit additional input from the participants. The system measures directly where each participant looks using a desk-mounted eyetracker. It represents this information metaphorically in a 3D virtual meeting room and within shared documents. The system not only shows how careful modelling of awareness features might improve distributed communilaboration, but also how it could be combined with a scalable and flexible use of network resources. As such, we feel attention-based groupware systems have the potential of becoming an important and generic awareness supplement to multiparty speech communication over telephone systems and Internet alike.

Chapter 6

Conclusions and Directions

INTRODUCTION

In this chapter, we will provide an overview and integration of the main empirical and practical conclusions of the studies presented in this thesis. The thesis investigated the functions, effects and design implications of conveying the visual attention of others — in the form of their gaze direction — in synchronous, interactive, multiparty mediated communication (and collaboration). Our studies were instigated by reports of problems in multiparty communication using (video) mediated systems that do not preserve gaze directional information. When using mediated systems, problems may occur with, amongst others, the regulation of turntaking and the referencing of other individuals. Our assumption was that this is directly caused by the lack of attention-related information in such systems. As a result of this absence, it may be difficult to establish the dialogic attention of others — whom others are talking or listening to — in a nonverbal fashion. Gaze-related cues would function as the main means for coding the dialogic attention of others in a way that would not interfere with verbal communication of content-related information. Perhaps apart from explicit hand gestures, other nonverbal visual cues would not contribute significantly towards this goal. We will first discuss how our empirical findings largely confirmed the above contentions. We will then discuss the implications of these findings for the design of multiparty mediated systems. Finally, we will explore directions in which the present research might be expanded in the future.

EMPIRICAL CONCLUSIONS

We will now summarize conclusions as to the function and isolated effect of a representation of the visual attention of others — in the form of their gaze — in multiparty communication, with a special focus on its role in conveying dialogic attention.

Visual Attention: An Effective Indicator of Dialogic Attention

In Chapter 3, we investigated to what extent the focus of visual attention of others *might* function as an effective indicator of their focus of dialogic attention. We examined this by measuring the amount of time subjects spent looking at the facial region of conversational partners listened or spoken to during four-way face-to-face discussions. We compared those findings with the amount of time subjects spent looking at others than the individual listened or spoken to. We found that gaze at the facial region may indeed be considered an excellent indicator of dialogic attention towards individuals in multiparty conversations. When someone is listening to an individual, there is an 88% chance that the person gazed at is the person listened to. When someone is addressing a single individual, there is a 77% chance that the person gazed at is the addressed individual. In this more or less dyadic condition, we found about 1.6 times more gaze while listening (62%) than while speaking (40%). When a speaker addresses more than a single individual, it seems likely that gaze may still be considered an effective indicator of his dialogic attention. When addressing a triad, speaker gaze typically seems to be distributed evenly across listeners. However, the total amount of speaker gaze rises significantly to about 59% of time. In such situations, the amount of gaze received by individual listeners (20%) is therefore still significantly more than the amount of gaze they would have received when not addressed (12%). In order to gain some insight into the impact of individual differences on these findings, we studied the effect of the personality variables extraversion and autonomy on gaze behaviour of subjects. We found that introverts have a tendency to gaze relatively more than extraverts at other individuals than the speaker listened to. Autonomous individuals tend to look relatively less at the individual in the focus of their dialogic attention. We conclude that our estimates of the effectiveness of gaze as a indicator of dialogic attention may not be considered free of individual differences. In addition, our estimates may be generalized only to situations where there is no requirement to look at task objects.

Representing Visual Attention as Gaze Causes Increased Speaker Switching and Deixis

In Chapter 4, we investigated whether the presence of a representation of visual attention — in the form of gaze directional cues — would have an isolated effect on multiparty mediated communication. Our second factor was the availability of other nonverbal upper-torso visual cues, as typically conveyed by video-mediated systems. We studied this by gauging parameters of the communication process during interaction of groups of three participants (one subject and two actors) solving language puzzles under three mediated conditions (all of which conveyed audio):

1) Simulated the use of a standard single-camera full-motion video-mediated system which effectively conveyed all nonverbal upper-torso visual cues of the actors other than head orientation and gaze at the facial region of subjects.
2) Simulated the use of a multiple camera full-motion video-mediated system which effectively conveyed all upper-torso nonverbal visual cues of the actors (including head orientation), except gaze at the facial region of subjects.
3) Simulated the use of a system with still images, manually selected by the actors, displaying head orientation and gaze at the facial region of subjects, but no other nonverbal upper-torso visual cues apart from physical appearance.

The presence of gaze directional cues in the form of head orientation caused the number of deictic verbal references to persons (deictic use of second-person pronouns) to increase significantly by a factor two. We believe this was due to differences between conditions in the subjects' estimate of the effectiveness of head pointing in disambiguating verbal deixis. We found no effects of the presence of nonverbal upper-torso visual cues other than gaze direction on any of our dependent variables. We did find a significant positive linear relationship between the amount of actor gaze at the facial region of subjects and the number of speaker switches ($r=.37$) and subject turns ($r=.34$). As such, the presence of a representation of the visual attention of others increased turn frequency, but only if it could be recognized by subjects as being aimed at themselves. As demonstrated by subject behaviour in the still image condition, the potential increase in turn frequency may be in the order of 25% when gaze at the facial region is conveyed in a manner that preserves its temporal and spatial characteristics as observed in multiparty face-to-face communication. We believe such increase is an indication of a more natural, and perhaps more efficient turntaking process. Since we found no significant effect of the presence of gaze directional or other nonverbal upper-torso visual cues on task performance, we believe the above findings are generalizable to other pure communication situations, at least where the presence of gaze directional cues is concerned.

Finding Explanations: Knowledge About the Dialogic Attention of Others Might Affect Turntaking

So to what extent may we attribute the effect of gaze on turntaking to its conveyance of dialogic attention information? Throughout our empirical chapters, we found three predominant explanations relating to perhaps the most *basic* functions of gaze at the facial region in multiparty conversation. The first two provide a *communication of interest*, the third provides *perception* of this and other visual information:

1) *Communication of (Dialogic) Attention*. Gaze provides a transparent and nonverbal visual signal of attention towards other persons. More specifically, it provides an excellent nonverbal means of conveying one's dialogic attention towards other persons. Firstly, although our evidence is indirect at best, it seems likely this function contributed to our finding of increased turn frequency in Chapter 4. Subjects could use gaze at their facial region to determine whether others might direct their auditory or articulatory attention towards them. As such, gaze at the facial region may improve Selfcentric Conversational Awareness, allowing subjects to ascertain whether they might be addressed or expected to speak. We found clear support for this explanation in our questionnaires. Subjects found it easier to observe who was talking to whom in the condition in which gaze at the facial region was best preserved. Secondly, although evidence is inconclusive, this function provides an explanation for our finding in Chapter 3 of increased speaker gaze with larger extents of articulatory attention. Since, when addressing more than a single individual, there is less time per individual to signal they are being addressed, speakers need to gaze more. This explanation is supported by Kendon [85].
2) *Regulation of Social Intimacy.* Gaze seems to have a direct effect on the state of arousal of the person receiving it. As such, gaze is one of the most intimate acts that can be performed at a distance. By avoiding or seeking gaze, conversational partners may attempt to keep the arousal state of themselves, as well as others, at a mutually satisfactory level (seeking an Equilibrium of Intimacy). Firstly, this function may have contributed to our finding of increased turn frequency in Chapter 4. Subjects in conditions with less gaze at their facial region may have felt uncomfortable and therefore less inclined to take the floor. Although trends in our questionnaires do not contradict this explanation, we found no conclusive evidence. Secondly, this function provides an explanation for our finding in Chapter 3 of increased speaker gaze with larger extents of articulatory attention. Since, when addressing triads, there is less time for speakers to gaze at each addressed individual, speakers would need to gaze more in order to maintain as satisfactory a level of intimacy with their audience as possible. Again, we did not find conclusive evidence regarding this explanation.

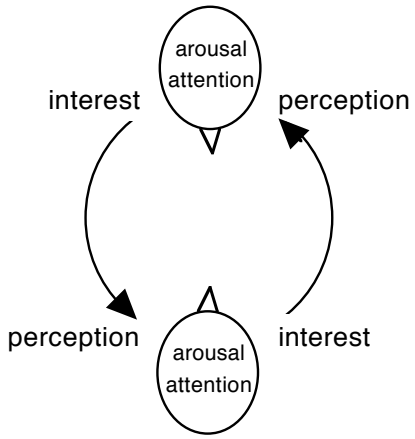


Figure 6-1. *The feedback loop of joint communication of interest.*

3) *Selective (Visual) Attention.* Visual attention should be regarded as a way of serving the requirements of cognitive processes with limited capacity. As such, the selection of relevant visual information may be considered a predominant, if not the most important, reason why people fixate, or look. Since the facial region is one of the richest and most relevant sources of visual information provided by the human body during communication, it is this part that is fixated upon most often. By doing so, the facial-visual and vocal-auditory channels are aligned, which may be regarded as a way of optimizing attentive resources across sensory modalities. The observation of the interest of others may play an important role in this optimization process. As such, we should regard visual attention as a way of closing the feedback loop of joint communication of interest (see Figure 6-1). Visual feedback of stooges could not have contributed to our finding of increased turn frequency in Chapter 4. However, it does provide an explanation for our finding in Chapter 3 of increased speaker gaze with larger extents of articulatory attention. Since, when addressing triads, there is less time to collect information on the nonverbal responses of each addressed individual, speakers might gaze more in order to satisfy their visual feedback requirements. Argyle et al. [12] provided strong evidence for the visual feedback function of gaze.

We believe the above functions are so inextricably synchronized that it is difficult, if not impossible, to investigate the exact causality of their individual contributions to effects of gaze on the multiparty conversational process. Although the communication of dialogic attention may be a parameter, within the scope of this thesis it should suffice that the conveyance of gaze may be considered to have a positive effect on the multiparty turntaking process.

PRACTICAL IMPLICATIONS

We agree with the traditional view that nonverbal upper-torso visual cues are an important complement to speech in support of human communication using mediated systems. However, it seems that the importance of one such cue, gaze at the facial region, has been underestimated. Contrary to other nonverbal upper-torso visual cues, gaze at the facial region does not seem to be highly redundantly coded by speech. This becomes particularly apparent when attempting to support group communication with mediated systems. If the design of multiparty mediated systems is a problem of conveying the least redundant cues first, we believe gaze at the facial region should be the first candidate to augment speech communication. In this section, we will first summarize our design recommendations with regard to the preservation of visual attention — in the form of gaze directional cues — in multiparty communication systems. We will then review how we can gauge, convey, and represent information about the attention of participants for communication as well as collaboration in an integrated and transparent fashion, while allowing a scalable and efficient use of network resources.

Preserving Visual Attention in Multiparty Mediated Communication Systems

With respect to the preservation of visual attention — in the form of gaze directional cues — in the design of mediated systems for multiparty communication, we formulate the following incremental requirements:

1) *Preservation of relative position.* Relative viewpoints of the participants should be based on a common reference point (e.g., around a shared workspace), providing basic support for the use of a common external context in deictic references.
2) *Preservation of head orientation.* Its representation eases the use of deictic references and may play a role in determining who is speaking to whom.
3) *Preservation of gaze at the facial region.* Allowing participants to gaze at each other's facial region eases turntaking: participants may find it easier to determine when they are addressed, or expected to speak. In addition, participants may find it easier to regulate the level of social intimacy. As such, gaze at the facial region may aid in providing a greater sense of telepresence.

If motion video is conveyed, we recommend the use of a multiple camera setup, in which each participant has a camera for each other participant. By putting each camera inside a video tunnel displaying the image of that participant, the above requirements can be implemented. This does, however, lead to a number of *unique* video streams that increases with almost the square of the group size (n^2-n, in which n is the number of participants. This means that Multicasting is not effective with such setup. As a consequence, the conveyance of gaze at the

facial region using motion video may not scale well with the number of participants in terms of network efficiency.

With respect to multiparty turntaking efficiency, or any of our other dependent variables, the use of motion video to convey nonverbal upper-torso visual cues other than gaze direction or physical appearance does not seem a *requirement* in task situations which are not highly personal*. In situations that *are*, it seems likely most people would opt for face-to-face communication instead of mediated communication. Even so, we believe the use of motion video should be at least optional. The choice for motion video may depend on the possibility of travel, individual preference, task situation and the availability of network bandwidth. If the latter is not an issue, one may simply choose to provide it, as long as this does not mean gaze directional information is lost.

Using (Visual) Attention to Integrally Mediate Awareness Information in Multiparty Communication and Collaboration

As we have seen, conveyance of visual attention may be considered a fundamental requirement for conveying interest towards persons in multiparty mediated *communication* systems. However, as we have seen in Chapter 2, it may also function as a means of providing information about objects of interest in multiparty mediated *collaboration* systems [151]. In Chapter 5, we therefore suggested that representations of (visual) attention be used as an integral way of making participants aware about who is talking or listening to whom (Conversational Awareness) as well as who is working on what (Workspace Awareness). Together, these forms provide Micro-level Awareness information about the activities of others during synchronous distributed communication and collaboration. We have shown how Micro-level Awareness information can be elegantly modelled in terms of conveying the attentive state of others. We defined an attentive state as a description of someone's focus of attention during an activity. At a syntactical level this involves describing the spatial and temporal properties of someone's (visual) attention, at a semantical level which actions, objects or people someone is attending to. Using this Attentive State model, we designed a prototype multiparty mediated communication and collaboration system, the GAZE Groupware System, in which Conversational and Workspace Awareness information would be conveyed in an integrated, transparent, and hopefully augmentative fashion. In order to achieve this, we applied the following design rationale:

- *Integrated Support for Conversational and Workspace Awareness.* As a main functional requirement, our system should provide a seamless integration between awareness about other participants' work activities

* Note that we do not consider hand gesturing an upper-torso cue.

(Workspace Awareness) and awareness about other participants' communication activities (Conversational Awareness). Based on our Attentive State model, we implemented this by using visual attention as an integral paradigm for gauging, conveying and representing Micro-level Awareness information. As such, representations of visual attention may function as a glue between knowing who is talking or listening to whom and knowing who is talking about what. The same attentional glue could be used to synchronize awareness information across focused and more general collaboration and communication tools.

- *Implicit Collection of Awareness Information.* We took a noncommand approach to providing awareness information. As the use of nonverbal gaze directional information in conversational turntaking demonstrates, this may lead to a more transparent and efficient interface, with lower mental load and less interruption of task-oriented activities. Eyetracking is one of the most direct, precise, noncommand, machine-readable and low-bandwidth means of gauging information about human (visual) attention. We therefore applied eyetrackers for gauging what persons or objects participants might be attending to.
- *Scalability of Networked Awareness Information.* As we have seen, the use of motion video to convey gaze directional cues for Conversational Awareness purposes may lead to problems of scalability in the use of Multicast network resources. Since the purpose of groupware systems is to support many users, typically across a computer network, scalability should be seen as an essential technical requirement. We tried to circumvent this problem by separating the conveyance of gaze directional cues from the conveyance of other nonverbal visual information. We implemented this by directly transmitting coordinates of visual attention, which can then be used to manipulate a single stream of video, or still image of a participant, such that gaze direction is metaphorically conveyed.
- *Representation of Awareness Information Using Natural Affordances.* The design of the system image should, where possible, be based on intuitions, knowledge and skills that people have acquired through years of shared work in the real world. We extended the single-user *Desktop Metaphor* to include multiple users, adding *meeting rooms* and *persons* to elements already present: *desks*, *documents*, and *(tele)pointers*. Thus, we built a 3D virtual meeting room, in which general communication and collaboration activities of remote participants are organized according to their attentional needs. In the meeting room, 2D snapshots (or persona) of participants are suspended around a desk with shared documents. The 3D orientation of a persona metaphorically conveys the direction of gaze of its owner, as measured by his eyetracking device. Each persona emits a light spot, which is projected onto the desk and within documents according to the measured

focus of attention of its owner (following a *Miner's Helmet* metaphor*). Together, these representations of attention provide integrated Micro-level Awareness information within the virtual meeting room. Focused collaboration editors provide a means of working together on document content, outside the scope of the virtual meeting room. During focused collaboration, light spots of collaborators are projected onto the document contents, mediating their exact focus of visual attention. These light spots are synchronized by colour and movement with light spots and persona in the virtual meeting room. Thus, light spots function as a kind of attentional glue, linking Workspace Awareness in focused collaboration tools with Workspace Awareness in the virtual meeting room.

* This means that, at least within our system, Empedocles' visual ray theory still holds (see page 1).

FUTURE DIRECTIONS

As is so often the case in science and technology, the research presented in this thesis yielded more questions than answers. In human communication, the functions and effects of nonverbal upper-torso visual cues in general, and gaze directional cues in particular, are so extraordinarily complex that we have only been able to scratch the surface. Very little is known about their application in human-computer communication. Since it is impossible to list the many issues that arose during our research project, we will here provide only a tentative selection of possible future research topics. We identified a clear need for a theory of communication behaviour, a theory of communication technology, and a further investigation of the applicability of the ideas demonstrated in the GAZE Groupware System prototype. Regarding the first, a theory of communication behaviour, we would like to highlight the following topics:

- *What is the exact relationship between the amount of gaze at the facial region and the number of turns taken?* By controlled varying of the amount of gaze at the facial region of subjects in a conversational setting, the function between gaze at the facial region and the number of individual turns taken might be described more precisely.
- *In the relationship between the amount of gaze at the facial region and the number of turns taken, what is the contribution of communication of dialogic attention relative to that of increased social intimacy?* It may be possible to further investigate the causality of the observed relationship between gaze at the facial region and the number of turns taken by the person gazed at. If a more optimal social intimacy would account for most of the variance in the number of turns taken, one would expect that a more optimal amount of gaze at the facial region in this respect should yield a higher turn frequency, irrespective of the correlation of that gaze through time with the focus of dialogic attention of the gazer. One would also expect to find a breakpoint in the relation between the amount of gaze at the facial region and the number of turns, indicating an optimum level of intimacy. If, however, the specification of dialogic attention would account for most of the variance in the number of turns taken, one would expect to find a positive relationship between the correctness of this specification and the number of turns taken.
- *Is orientation of thought looking in dyadic communication caused by an interaction between (visual) attention and communication of interest functions of gaze?* In dyadic situations, speakers tend to gaze less than listeners. Gaze avoidance by speakers was traditionally attributed to an interference of visual input with their preparation of verbal utterances. However, we observed that when addressing a triad, the total amount of speaker gaze rises significantly, to roughly the same level as while listening to a single individual. Is this because the need to avoid gaze during preparation of utterances is overridden by a need to communicate interest?

Or does our theory hold that gaze avoidance by speakers is due to an interaction effect between (visual) attention, intimacy and dialogic attention functions of gaze, which disappears when the extent of articulatory attention is large?

- *Are effects of personality on gaze behaviour due to differences in cortical arousal of individuals?* Individual differences in social intimacy requirements might form a primary explanation for the observed effects of personality on the amount of gaze at the facial region during face-to-face communication. By studying patterns of arousal in communicating individuals, perhaps using future brain imaging techniques [149], it may be possible to examine the potential relationship between cortical arousal, personality and gaze in a more detailed fashion.

We also identified a need for a theory of communication technology. Regarding this, we recommend that, amongst others, the following research topics be investigated:

- *When motion video is conveyed, what is the maximum angle of parallax between the location of a video conferencing camera lens and the location of participant eye region representation(s), if gaze at the facial region is to be preserved?* Humans are very sensitive to the exact location within their facial region of eye fixations of others facing them. When using video tunnels, the lens of the video camera should therefore be co-located with the position of the eyes in the video representation of the corresponding participant. We do not know how exact this co-location should be. Our own qualitative observations suggest that an angle of parallax of 3 degrees at normal viewing distance may already inhibit correct perception of gaze at the facial region. This is not only relevant to the correct operation of video tunnels in general, but also to the application of motion video in our GAZE Groupware System. Since gaze direction is conveyed separately in this system, a multiple camera setup would not be required as long as gaze at the facial region remains recognizable for all participants. If the maximum angle of parallax is inhibitively small, a multiple camera setup may be required even with the GAZE Groupware System to allow correct capturing of frontal gaze. The maximum angle of parallax could be established by means of subject evaluation of gaze at their facial region in an on-screen image of another person facing them. This person would look towards the camera capturing the image, and would do so at controlled angles from the center of the camera lens. Such experiment would, however, need to take into account other variables such as distance between subject and screen, on-screen image size and the zoom factor of the camera lens.
- *Attentive Filtering of Audio and Video Signals.* The quality of service of audio as well as video transmissions could be based on the locus and span of visual attention of individual users, as discussed on page 145 of this thesis. Regarding audio, we need to investigate whether attenuation of

auditory information from participants outside the focus of visual attention would aid selective auditory attention, and if so, what parameters would need to be controlled in this process. Regarding video, early results by Duchowski and McCormick [43] show that the presentation of lower-resolution motion video to peripheral vision using a simple linear spatial degradation function centered around the on-screen point of gaze may indeed be imperceptible. Problems with network delays might, however, impair practical application. It is also still unclear to what extent such functionality might *aid* users (e.g., that are field-dependent [163]) in maintaining selective visual attention. The above attentive filtering techniques may also provide a means of compressing audio and video signals, but only when standard broadcasting, rather than Multicasting, is used as a network paradigm.

Of course, we are also aware of the need to investigate further the applicability of the ideas demonstrated in the GAZE Groupware System prototype. Regarding this, we recommend that, amongst others, the following research topics be investigated:

- *Usability evaluation of the GAZE Groupware System.* In this thesis, we focused on the empirical basis for the conveyance of gaze directional cues in mediated systems. However, we need to also carefully evaluate the real-world usability of our prototype design, based on the issues listed on page 150 of this thesis.
- *The application of motion video in the GAZE Groupware System.* We hope that extensions to the VRML 2 standard will, in the nearby future, allow use of streaming video for personas in our 3D virtual meeting room. When this becomes feasible, we will need to use video tunnels (one per user) to allow capturing of gaze at the facial region. Depending on the above discussed maximum angle of parallax, multiple cameras should be put inside each video tunnel at the position of personas. Since the user's on-screen point of gaze is known, software could automatically determine what camera captures the user's frontal gaze most accurately. The video signal of this camera could then be multicast to other users as a single stream. Alternatively, a single miniature camera could be moved mechanically within the video tunnel, following the horizontal component of the user's on-screen point of gaze. This way, it would be possible to have a more flexible setup, while still capturing images of the user looking exactly into the camera lens.
- *Further applications of eyetracking in human-computer communication.* One can conceive of many more applications of eyetrackers in (noncommand) user interfaces. To name but a few, we would like to investigate their application in transparent control of focus in fisheye-view representations, their use in specifying attention towards intelligent interface

agents, and their more general utilization in supplying external context information for deixis in speech interfaces.

References

1. *NEN-EN 894-2. Veiligheid van machines - Ergonomische eisen voor het ontwerpen van informatie- en bedieningsmiddelen, Deel 2 Informatiemiddelen*, 1993.
2. Acker, S. and Levitt, S. Designing videoconference facilities for improved eye contact. *Journal of Broadcasting & Electronic Media* 31(2), 1987, pp. 181-191.
3. Aiello, J.R. A test of equilibrium theory: visual interaction in relation to orientation, distance and sex of interactants. *Psychonomic Science* 27, 1972, pp. 335-336.
4. Appelt, W. and Busbach, U. The BSCW system: A WWW-based application to support cooperation of distributed groups. In *IEEE Proc. of the Fifth Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises.* Stanford, CA: Los Alamitos, CA USA: IEEE Computer Society Press, 1996, pp. 304-309.
5. Applied Science Laboratories. *ASL Eyetracking Systems*. Bedford, MA USA. *http://www.a-s-l.com*, 1998.
6. Argyle, M. *The Psychology of Interpersonal Behaviour*. London, UK: Penguin Books, 1967.
7. Argyle, M. *Social Interaction*. London, UK: Tavistock Publications, 1969.
8. Argyle, M. and Cook, M. *Gaze and Mutual Gaze*. London, UK: Cambridge University Press, 1976.
9. Argyle, M. and Dean, J. Eye-contact, distance and affiliation. *Sociometry* 28, 1965, pp. 289-304.
10. Argyle, M. and Graham, J. The Central Europe Experiment - looking at persons and looking at things. *Journal of Environmental Psychology and Nonverbal Behaviour* 1, 1977, pp. 6-16.
11. Argyle, M. and Ingham, R. Gaze, mutual gaze and proximity. *Semiotica* 6, 1972, pp. 32-49.
12. Argyle, M., Ingham, R., Alkema, F., and McCallin, M. The different functions of gaze. *Semiotica* 7, 1973, pp. 19-32.
13. Argyle, M., Lalljee, M., and Cook, M. The effects of visibility on interaction in a dyad. *Human Relations* 21, 1968, pp. 3-17.
14. Baecker, R.M., Nastos, D., Posner, I.R., and Mawby, K.L. The user-centred iterative design of collaborative writing software. *Proceedings of ACM INTERCHI'93 Conference on Human Factors in Computing Systems*. Amsterdam, The Netherlands: ACM, 1993, pp. 399-405.
15. Barelds, D.P.H. and Luteijn, F. Het meten van persoonlijkheid: de NPV versus de FFPI. *Nederlands Tijdschrift voor de Psychologie*, 1998 (submitted).
16. Benford, S., Greenhalgh, C., Bowers, J., Snowdon, S., and Fahlén, L. User embodiment in collaborative virtual environments. In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems.* Denver, CO USA: ACM, 1995.
17. Bennet, G.K., Seashore, H.G., and Wesman, A.G. *Manual for the Differential Aptitude Tests*. New York, NY USA: The Psych. Corporation, 1959.
18. Biederman, I., Glass, A.L., and Stacy, E.W. Searching for objects in real world scenes. *Journal of Experimental Psychology* 97, 1973, pp. 22-27.
19. Brady, P.T. A statistical analysis of on-off patterns in 16 conversations. *The Bell System Technical Journal* (Jan.), 1968, pp. 73-91.

20. Bridgeman, B., V.d. Heijden, A.H.C., and Velichkovsky, B.M. A theory of visual stability across saccadic eye movements. *Behavioral and Brain Sciences* 17(2), 1994, pp. 247-258.

21. Broadbent, D.E. *Perception and Communication*. London, UK: Pergamon Press, 1958.

22. Buxton, W.A.S. *A Directory of Sources for Input Technologies*. University of Toronto, Canada. *http://www.dgp.toronto.edu/people/BillBuxton/InputSources.html*, 1998.

23. Buxton, W.A.S. There's more to interaction than meets the eye: Some issues in manual input. In Norman, D.A. and Draper, S.W. (Ed.), *User Centered System Design: New Perspectives on HCI.* Hillsdale, NJ USA: Lawrence Erlbaum Associates, 1986, pp. 319-337.

24. Buxton, W.A.S. and Moran, T. EuroPARC's integrated interactive intermedia facility (iiif): Early experience. In Gibbs, S. and Verrijn-Stuart, A.A. (Ed.), *Multi-user interfaces and applications, Proceedings of the IFIP WG 8.4 Conference on Multi-user Interfaces and Applications,* Heraklion, Crete. Amsterdam, The Netherlands: Elsevier Science Publishers B.V. (North-Holland), 1990, pp. 11-34.

25. Buxton, W.A.S. Telepresence: Integrating shared task and person spaces. In *Proceedings of Graphics Interface '92*, 1992, pp. 123-129.

26. Buxton, W.A.S., Sellen, A.J., and Sheasby, M.C. Interfaces for multiparty videoconferences. In Finn, K.E., Sellen, A.J., and Wilbur, S.B. (Ed.), *Video-Mediated Communication.* Mahwah, NJ USA: Laurence Erlbaum Associates, 1997, pp. 385-400.

27. Chapanis, A. Interactive human communication. *Scientific American* 232(3), 1975, pp. 36-42.

28. Cherry, C. Some experiments on the reception of speech with one and with two ears. *Journal of the Acoustic Society of America* 25, 1953, pp. 975-979.

29. Claffy, K.C., Braun, H.-W., and Polyzos, G.C. Tracking long-term growth of the NSFNET. *Communications of ACM* 37(8), 1994, pp. 34-45.

30. Clark, H.H. and Brennan, S.E. Grounding in communication. In Resnick, L.B., Levine, J., and Behreno, S.D. (Ed.), *Socially Shared Cognition.* Washington, DC USA: American Psychology Association, 1991.

31. Cline, M.G. The perception of where a person is looking. *American Journal of Psychology* 80, 1967, pp. 41-50.

32. Cohen, K.M. Speaker interaction: Video teleconferences versus face-to-face meetings. In *Proceedings of Teleconferencing and Electronic Communications.* Madison, WI USA: University of Wisconsin Press, 1982, pp. 189-199.

33. Colston, H. and Schiano, D. Looking and lingering as conversational cues in VMC. In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems.* Denver, CO USA: ACM, 1995.

34. Costa, P.T. and McCrae, R.R. *The NEO Personality Inventory Manual*. Odessa, FL USA: Psychological Assessment Resources, Inc., 1985.

35. De Furia, S. and Scacciaferro, J. *The MIDI Resource Book*. Pompton Lakes, NJ USA: Third Earth Publishing, 1987. ISBN 0-88188587-8.

36. Dennett, D.C. *Conciousness Explained*. London, UK: Penguin, 1991.

37. Descartes, R. *Discours de la Méthode*. Leiden, The Netherlands: Jean Maire, 1637.

38. Deutsch, J.A. and Deutsch, D. Attention: Some theoretical considerations. *Psychological Review* 70, 1963, pp. 80-90.

39. Diebold, A.R. Anthropology of the comparative psychology of communicative behavior. In Sebeok, T.A. (Ed.), *Animal Communication — Techniques of Study and Results of Research.* Bloomington, IN USA: Indiana University Press, 1968, pp. 525-571.

40. Dix, A., Finlay, J., Abowd, G., and Beale, R. *Human-Computer Interaction*. Hertfordshire, UK: Prentice Hall International (UK), 1993.

41. Dourish, P. and Bellotti, V. Awareness and coordination in shared workspaces. In *Proceedings of ACM CSCW'92 Conference on Computer-Supported Cooperative Work.* Toronto, Canada: ACM, 1992, pp. 25-38.

42. Dourish, P. and Bly, S. Portholes: Supporting awareness in a distributed work group. In *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems.* Monterey, CA USA: ACM, 1992, pp. 541-547.

43. Duchowski, A.T. and McCormick, B.H. Gaze-contingent video resolution degradation. In *Human Vision and Electronic Imaging III.* San Jose, CA USA: SPIE, 1998.

44. Duncan, S. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* 23, 1972, pp. 283-293.

45. Ellis, C.A., Gibbs, S.J., and Rein, G.L. Groupware: Some issues and experiences. *Communications of ACM* 34(1), 1991, pp. 38-58.

46. Empedocles. περι φυσεως. Sicily, Italy: 450-441 B.C. Republished in V.d. Ben, N. (Ed.), *The Poem of Empedocles' Peri Physios : Towards a New Edition of All the Fragments: Thirty-one Fragments.* Amsterdam, The Netherlands: B.R. Grüner, 1975, pp. 230.

47. Enheduanna. *Nin-me-sár-ra*. Sumerian cuneiform tablets, Mesopotamia: late 3rd Millennium B.C. Republished in Hallo, W.W. and Van Dijk, J.J.A. (Ed.), *The Exaltation of Inanna.* New Haven, CT USA: Yale University Press, 1968, pp. 101.

48. Ensor, J.R. Virtual meeting rooms. In Finn, K.E., Sellen, A.J., and Wilbur, S.B. (Ed.), *Video-Mediated Communication.* Mahwah, NJ USA: Laurence Erlbaum Associates, 1997, pp. 415-434.

49. Ericksson, H. MBONE: The multicast backbone. *Communications of ACM* 37(8), 1994, pp. 54-60.

50. Exline, R.V. Explorations in the process of person perception: Visual interaction in relation to competition, sex and need for affiliation. *Journal of Personality* 31, 1963, pp. 1-20.

51. Exline, R.V. Visual interaction: the glances of power and preference. In *Nebraska Symposium on Motivation*, 1971, pp. 163-206.

52. Exline, R.V. and Long, B. Visual behavior in relation to power of position in legitimate and illegitimate power hierarchies. Unpublished manuscript, Department of Psychology, University of Delaware, OH USA, 1971.

53. Exline, R.V. and Messick, D. The effects of dependency and social reinforcement upon visual behaviour during an interview. *British Journal of Social and Clinical Psychology* 6, 1967, pp. 256-266.

54. Eysenck, H.J. *The Dynamics of Anxiety and Hysteria*. London, UK: Routledge and Kegan Paul, 1957.

55. Fokkema, S.D. and Dirkzwager, A. Ruimtelijk Inzicht; Taalgebruik II, Zinnen. *Differentiële Aanlegtests.* Amsterdam, The Netherlands: Swets & Zeitlinger, 1960.

56. Galen, C. De placitis Hippocratis et Platonis. 130-200 A.D. Republished in Lacy, P.D. (Ed.), *On the doctrines of Hippocrates and Plato*. Berlin, Germany: Akademie-Verlag, 1978.

57. Galen, C. θεραπευτικη μεθοδος. 130-200 A.D. Republished in Linacrus, T. (Ed.), *Galeni Methodus medendi vel de morbis curandi.* Lutetiae (Paris, France): Godefr. Hittopii et Desiderium Maheu, 1519.

58. Gaver, W. Sound Support for Collaboration. In *Proceedings of ECSCW'91 European Conference on Computer Supported Cooperative Work.* Amsterdam, The Netherlands: Kluwer, 1991.

59. Gaver, W., Moran, T., *et al.* Realizing a video environment: EuroPARC's RAVE system. *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems*. Monterey, CA USA: ACM, 1992, pp. 27-35.

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

61. Gifford Jr., E.S. *The Evil Eye*. New York, NY USA: The MacMillan Company, 1958.

62. Givón, T. The grammar of referential coherence as mental processing instructions. *Linguistics* 30, 1992, pp. 5-55.

63. Gould, J.D. Looking at pictures. In Monty, R.A. and Senders, J.W. (Ed.), *Eye Movements and Psychological Processes.* Hilldale, NJ USA: Lawrence Erlbaum Associates, 1976.

64. Greenberg, S. A fisheye text editor for relaxed WYSIWIS groupware. In *Conference Companion of ACM CHI'96 Conference on Human Factors in Computing Systems.* Vancouver, Canada: ACM, 1996, pp. 212-213.

65. Greenberg, S. Peepholes: Low cost awareness of one's community. In *Conference Companion of ACM CHI'96 Conference on Human Factors in Computing Systems.* Vancouver, Canada: ACM, 1996.

66. Greenhalgh, C. and Benford, S. MASSIVE: A collaborative virtual environment for teleconferencing. *ACM Transactions on Computer-Human Interaction* 2(3), 1995, pp. 239-261.

67. Grosseteste, R. De iride. Oxford, England: 1230-1235. Republished in Baur, L. (Ed.), *Die philosophischen Werke des Robert Grosseteste, Bischofs von Lincoln.* Münster, Germany: Aschendorffsche Verlagsbuchhandlung, 1912, pp. 778.

68. Gutwin, C. and Greenberg, S. Workspace awareness for groupware. In *Conference Companion of ACM CHI'96 Conference on Human Factors in Computing Systems.* Vancouver, Canada: ACM, 1996.

69. Gutwin, C., Greenberg, S., and Roseman, M. Workspace awareness support with radar views. In *Conference Companion of ACM CHI'96 Conference on Human Factors in Computing Systems.* Vancouver, Canada: ACM, 1996, pp. 210-211.

70. Gutwin, C., Roseman, M., and Greenberg, S. A usability study of awareness widgets in a shared workspace groupware system. In *Proceedings of ACM CSCW'96 Conference on Computer-Supported Cooperative Work.* Boston, MA USA: ACM, 1996, pp. 258-268.

71. Hall, E.T. *The Hidden Dimension*. Garden City, NY USA: Doubleday, 1966.

72. Harrison, S. and Dourish, P. Re-Place-ing Space: The roles of place and space in collaborative systems. In *Proceedings of ACM CSCW'96 Conference on Computer-Supported Cooperative Work.* Cambridge, MA USA: ACM, 1996, pp. 67-76.

73. Heath, C. and Luff, P. Disembodied conduct: Communication through video in a multi-media office environment. In *Proceedings of ACM CHI'91 Conference on Human Factors in Computing Systems.* New Orleans, LA USA: ACM, 1991.

74. Hendriks, A.A.J. *The Construction of the Five-Factor Personality Inventory (FFPI)*. PhD Thesis. Groningen University, The Netherlands, 1997.

75. Hess, E.H. Pupillometrics. In Greenfield, N. and Sternbach, R. (Ed.), *Handbook of Psychophysiology.* New York, NY USA: Holt, Rinehart and Winston, 1972.

76. Hudson, S.E. and Smith, I. Techniques for addressing fundamental privacy and disruption tradeoffs in awareness support systems. In *Proceedings of ACM CSCW'96 Conference on*

Computer-Supported Cooperative Work. Cambridge, MA USA: ACM, 1996, pp. 248-257.

77. Infusion Systems. *The I-Cube User Manual*. Vancouver, Canada: Infusion Systems Ltd. *http://www.infusionsystems.com*, 1996.

78. Isaacs, E. and Tang, J. What video can and can't do for collaboration: a case study. In *Proceedings of ACM Multimedia'93*. Anaheim, CA USA: ACM, 1993.

79. Ishii, H. and Kobayashi, M. ClearBoard: A seamless medium for shared drawing and conversation with eye contact. In *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems.* Monterey, CA USA: ACM, 1992.

80. Ishii, H. and Ullmer, B. Tangible Bits: Towards seamless interfaces between people, bits, and atoms. In *Proceedings of ACM CHI'97 Conference on Human Factors in Computing Systems.* Atlanta, GA USA: ACM, 1997, pp. 234-241.

81. Jaffe, J. and Feldstein, S. *Rhythms of Dialogue*. New York, NY USA: Academic Press, 1970.

82. Jaspars, J.M.F., *et al*. Het observeren van oogcontact. *Nederlands Tijdschrift voor de Psychologie* 28, 1973, pp. 67-81.

83. Joch, A. What pupils teach computers. *BYTE* 21(7), 1996, pp. 99-100.

84. Kahneman, D. *Attention and Effort*. Englewood Cliffs, NJ USA: Prentice-Hall Inc., 1973.

85. Kendon, A. Some function of gaze direction in social interaction. *Acta Psychologica* 32, 1967, pp. 1-25.

86. Kendon, A. and Cook, M. The consistency of gaze patterns in social interaction. *British Journal of Psychology* 69, 1969, pp. 481-494.

87. Kleck, R.E. and Nuessle, W. Congruence between the indicative and communicative functions of eye-contact in interpersonal relations. *British Journal of Social and Clinical Psychology* 7, 1968, pp. 241-246.

88. Kowler, E., Anderson, E., Dosher, B., and Blaser, E. The role of attention in programming saccades. *Vision Research* 35(13), 1995, pp. 1897-1916.

89. Kruger, K. and Hückstedt, B. Die Beurteilung der Blickrichtungen. *Zeitschrift für Experimentelle und Angewandte Psychologie* 16, 1969, pp. 452-472.

90. Laurel, B. Interface agents: metaphors with character. In Laurel, B. (Ed.), *The Art of Human-Computer Interface Design.* Reading, MA USA: Addison-Wesley, 1990, pp. 355-365.

91. LC Technologies, I. *The Eyegaze Communication System*. Fairfax, VA USA. *http://www.lctinc.com*, 1997.

92. Libby, W.L. Eye contact and direction of looking as stable individual differences. *Journal of Experimental Research on Personality* 4, 1970, pp. 303-312.

93. Lucretius. De rerum natura. Rome, Italy: 56 B.C. Republished in Avancius, H. (Ed.), *T. Lvcretii Cari Libri sex nvper emendati.* Venetiis (Venice, Italy): Aldum, 1500. *http://classics.mit.edu/Carus/nature_ things.html*.

94. Luteijn, F., Starren, J., and Van Dijk, H. *Nederlandse Persoonlijkheidsvragenlijst - Revisie 1985 (NPV)*. Lisse, The Netherlands: Swets & Zeitlinger, 1985.

95. Mackworth, N.H. and Morandi, A.J. The gaze selects informative details within pictures. *Perception and Psychophysics* 2, 1967, pp. 547-552.

96. Mark of the Unicorn. *Performer MIDI Sequencing Software*. *http://www.motu.com*, 1998.

97. McDaniel, S.E. Providing awareness information to support transitions in remote computer-mediated collaboration. In *Conference Companion of ACM CHI'96 Conference on Human Factors in Computing Systems.* Vancouver, Canada: ACM, 1996, pp. 57-58.

98. Mobbs, N.A. Eye-contact in relation to social introversion-extraversion. *British Journal of Social and Clinical Psychology* 7, 1968, pp. 305-306.

99. Monk, A., McCarthy, J., Watts, L., and Daly-Jones, O. Measures of process. In Thomas, P. (Ed.), *CSCW Requirements and Evaluation.* Berlin, Germany: Springer Verlag, 1996, pp. 125-139.

100. Mooij, H.A. *Eyecatcher System*. Mooij Holding, Oegstgeest, The Netherlands. *http://utopia.knoware.nl/users/mooij/*, 1997.

101. Moray, N. *Listening and Attention*. Penguin Science of Behaviour, Foss, B.M. (Ed.). Harmondsworth, UK: Penguin Books, 1969.

102. Nakanishi, H., Yoshida, C., Nishimura, T., and Ishida, T. Freewalk: Supporting casual meetings in a network. In *Proceedings of ACM CSCW'96 Conference on Computer-Supported Cooperative Work.* Cambridge, MA USA: ACM, 1996, pp. 308-314.

103. Negroponte, N. *Being Digital*. New York: Vintage Books, 1995.

104. Negroponte, N. Talking heads-display techniques for persona. Unpublished paper. *MIT Architecture Machine Group.* Cambridge, MA USA.

105. Neisser. *Cognition and Reality: Principles and Implications of Cognitive Psychology*. San Francisco, CA USA: W.H. Freeman, 1976.

106. Nielsen, G. *Studies in Self Confrontation*. Copenhagen, Denmark: Monksgaard, 1962.

107. Nielsen, J. Noncommand user interfaces. *Communications of ACM* 36(4), 1993, pp. 83-99.

108. Norman, D.A. *The Psychology of Everyday Things*. New York, NY USA: Basic Books, 1988.

109. O'Connaill, B., Whittaker, S., and Wilbur, S. Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human Computer Interaction* 8, 1993, pp. 389-428.

110. Okada, K.-I., Maeda, F., Ichikawaa, Y., and Matsushita, Y. Multiparty videoconferencing at virtual social distance: MAJIC design. In *Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work.* Chapel Hill, NC USA: ACM, 1994, pp. 385-393.

111. Opcode Systems. *Studio64 MIDI/SMPTE interface*. *http://www.opcode.com*, 1998.

112. Ovidius. Metamorphoses. Rome, Italy: 8 A.D. Republished in Puteolanus, F. (Ed.), *Opera.* Bononiae (Bologna, Italy): Balth. Azoguidus, 1471.

113. Posner, M.I. Orienting of attention. *Quarterly Journal of Experimental Psychology* 32, 1980, pp. 3-25.

114. Puckette, M. and Zicarelli, D. *MAX - An Interactive Graphic Programming Environment.* Opcode Systems, Menlo Park, CA USA, 1990.

115. Rabb, F., Blood, E., Steiner, R., and Jones, H. Magnetic position and orientation tracking system. *IEEE Transactions on Aerospace and Electronic Systems* 15(5), 1979, pp. 709-718.

116. Raeithel, A. and Velichkovsky, B.M. Joint attention and co-construction: New ways to foster user-designer collaboration. In Nardi, B. (Ed.), *Context and Consciousness: Activity Theory and Human-Computer Interaction.* Cambridge, MA USA: MIT Press, 1996.

117. Rothkopf, E.Z. Machine adaption to psychological differences among users in instructive information exchanges with computers. In Klix, F. and Wandke, H. (Ed.), *Man-Computer Interaction Research MACINTER I.* Amsterdam, The Netherlands: North Holland, 1986.

118. Rutter, D.R. *Communicating by Telephone*. London, UK: Pergamon Press, 1987.

119. Rutter, D.R., Morley, I.E., and Graham, J.C. Visual interaction in a group of introverts and extraverts. *European Journal of Social Psychology* 2(4), 1972, pp. 371-384.

120. Rutter, D.R. and Stephenson, G.M. The role of visual communication in synchronising conversation. *European Journal of Social Psychology* 7, 1977, pp. 29-37.

121. Rutter, D.R. and Stephenson, G.M. Visual interaction in a group of schizophrenic and depressive patients. *British Journal of Social and Clinical Psychology* 11, 1972, pp. 57-65.

122. Rutter, D.R., Stephenson, G.M., and Dewey, M.E. Visual communication and the content and style of conversation. *British Journal of Social Psychology* 20, 1981, pp. 41-52.

123. Sacks, H. March 2 Turn-taking; Collaborative utterances via appendor questions; Instructions; Directed utterances. In Jefferson, G. (Ed.), *Lectures on Conversation.* Oxford UK, Cambridge MA USA: Blackwell Publishers, 1995, pp. 523-534.

124. Sacks, H., Schegloff, E.A., and Jefferson, G. A simplest systematics for the organization of turntaking. *Language* 50, 1974, pp. 696-735.

125. Sellen, A.J., Buxton, W.A.S., and Arnott, J. Using spatial cues to improve desktop videoconferencing. In *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems.* Monterey, CA USA: ACM, 1992, pp. 651-652.

126. Sellen, A.J. Remote conversations: the effects of mediating talk with technology. *Human Computer Interaction* 10(4), 1995.

127. Sellen, A.J. Speech patterns in video-mediated conversations. In *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems.* Monterey, CA USA: ACM, 1992, pp. 49-59.

128. Shneiderman, B. *Designing the User-Interface: Strategies for Effective Human-Computer Interaction*. Reading, MA USA: Addison Wesley, 1987.

129. Short, J., Williams, E., and Christie, B. *The Social Psychology of Telecommunications*. London, UK: Wiley, 1976.

130. Silicon Graphics. The Virtual Reality Modeling Language 2.0. *ISO/IEC DIS 14772-1 submission. http://vrml.sgi.com/moving-worlds/*, 1997.

131. Smith, D., Irby, C., Kimball, R., Verplank, B., and Harslem, E. Designing the Star user interface. *BYTE* 7(4), 1982.

132. Sohlenkamp, M. and Chwelos, G. Integrating communication, cooperation and awareness: The DIVA virtual office environment. In *Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work.* Chapel Hill, NC USA: ACM, 1994, pp. 331-343.

133. Sony. *Sony Community Place. http://vs.spiw.com/vs/*, 1997.

134. Stapley, B. *Visual Enhancement of Telephone Conversations*. PhD Thesis. Imperial College, London, UK, 1972.

135. Stass, J.W. and Willis, F.N. Eye contact, pupil dilation, and personal preference. *Psychonomic Science* 7, 1967, pp. 375-376.

136. Stefik, M., Bobrow, D.G., Foster, G., Lanning, S., and Tatar, D. WYSIWIS revised: Early experiences with multiuser interfaces. *ACM Transactions on Office Information Systems* 5(2), 1987, pp. 147-167.

137. Stefik, M., Foster, G., *et al.* Beyond the Chalkboard: Computer support for collaboration and problem solving in meetings. *Communications of ACM* 30(1), 1987, pp. 33-47.

138. Stephenson, G.M. and Rutter, D.R. Eye-contact, distance and affiliation: a re-evaluation. *British Journal of Psychology* 61, 1970, pp. 385-393.

139. Stolk, H. *Eye Movements and Study Material*. PhD Thesis. Open Universiteit, Heerlen, The Netherlands, 1995.

140. Theeuwes, J. Visual selective attention: A theoretical analysis. *Acta Psychologica* 83, 1993, pp. 93-154.

141. Trager, G.L. and Smith Jr., H.L. *An Outline of English Structure*. Washington, DC USA: American Council of Learned Societies, 1957.

142. Treismann, A. Features and objects in visual processing. *Scientific American* 255, 1986.

143. Treismann, A. and Gelade, G. A feature-integration theory of attention. *Cognitive Psychology* 12, 1980, pp. 97-136.

144. Tupes, E.C. and Christal, R.E. Recurrent personality factors based on trait ratings. *Journal of Personality* 60, 1992, pp. 225-251 (original report published in 1960).

145. Underkoffler, J. and Ishii, H. Illuminating light: An optical design tool with a luminous-tangible interface. In *Proceedings of ACM CHI'98 Conference on Human Factors in Computing Systems.* Los Angeles, CA USA: ACM, 1998, pp. 542-550.

146. Van der Veer, G.C. Individual differences and the user interface. *Ergonomics* 32, 1989, pp. 1431-1449.

147. Van der Veer, G.C. and Van Wijk, R. Teaching a spreadsheet application — Visual-spatial metaphors in relation to spatial ability, and the effect on mental models. In Tauber, M.J. and Gorny, P. (Ed.), *Visualisation in Human-Computer Interaction.* Berlin, Germany: Springer Verlag, 1978.

148. Van der Veer, G.C. *Human-Computer Interaction: Learning, Individual Differences, and Design Recommendations*. PhD Thesis. Vrije Universiteit, Amsterdam, The Netherlands, 1990.

149. Velichkovsky, B.M. and Hansen, J.P. New technological windows to mind: There is more in eyes and brains for human computer interaction. In *Proceedings of ACM CHI'96 Conference on Human Factors in Computing Systems.* Vancouver, Canada: ACM, 1996, pp. 496-503.

150. Velichkovsky, B.M., Sprenger, A., and Unema, P. Towards gaze-mediated interaction: Collecting solutions of the "Midas touch problem". In *Proceedings of INTERACT'97.* Sydney, Australia, 1997.

151. Velichkovsky, B.M. Communicating attention: Gaze position transfer in cooperative problem solving. *Pragmatics and Cognition* 3(2), 1995, pp. 199-222.

152. Vertegaal, R. Conversational awareness in multiparty VMC. In *Extended Abstracts of ACM CHI'97 Conference on Human Factors in Computing Systems.* Atlanta, GA USA: ACM, 1997, pp. 6-7.

153. Vertegaal, R. GAZE: Visual-spatial attention in communication. Video Contribution. In *Proceedings of ACM CSCW'98 Conference on Computer-Supported Cooperative Work.* Seattle, WA USA: ACM, 1998 (in press)

154. Vertegaal, R. and Guest, S. Network issues in the growth and adoption of networked CSCW services. *SIGCHI Bulletin* 27(4), 1995, pp. 63-68.

155. Vetter, R.J. ATM concepts, architectures, and protocols. *Communications of ACM* 38(2), 1995, pp. 31-38.

156. Vine, I. Judgement of direction of gaze — an interpretation of discrepant results. *British Journal of Social and Clinical Psychology* 10, 1971, pp. 320-331.

157. Von Cranach, M. and Ellgring, J.H. The perception of looking behaviour. In Von Cranach, M. and Vine, I. (Ed.), *Social Communication and Movement.* London, UK: Academic Press, 1973.

158. Wachtenberg, E. *Eye Pointer Device.* Personal Communication. Initiative Systems, Kfar Saba, Israel, 1997.

159. Webster. *Merriam-Webster Online*. *http://www.m-w.com/home.htm*, 1998.

160. Weisbrod, R.M. Looking behavior in a discussion group. Unpublished paper, Department of Psychology, Cornell University, Ithaca, NY USA, 1965.

161. Wertheimer, M. Untersuchungen zur Lehre von der Gestalt, II. *Psychologische Forschung* 4, 1923, pp. 301-350.

162. Williams, E. Visual interaction and speech patterns: An extension of previous results. *British Journal of Social and Clinical Psychology* 17, 1978, pp. 101-102.

163. Witkin, H.A., Moor, C.A., Goodenough, D.R., and Cox, P.W. Field-dependent and field-independent cognitive styles and their educational implementations. *Review of Educational Research* 77, 1977, pp. 7-64.

164. Wittenborn, J. Factorial equations for tests of attention. *Psychometrika* 8, 1943, pp. 119-35.

165. Yarbus, A.L. Eye movements during perception of complex objects. In Riggs, L.A. (Ed.), *Eye Movements and Vision, Ch. VII.* New York, NY USA: Plenum Press, 1967, pp. 171-196.

166. Zhang, J. and Norman, D. Representations in distributed cognitive tasks. *Cognitive Science* 18, 1994, pp. 87-122.

167. Zimmerman, T.G., Lanier, J., Blanchard, C., Bryson, S., & Harvill, Y. A hand gesture interface device. *Proceedings of ACM CHI+GI'87 Conference on Human Factors in Computing Systems and Graphics Interface*. Toronto, Canada: ACM, 1987, pp. 189-192.

Samenvatting

Dit proefschrift onderzoekt de functies, effecten en ontwerpimplicaties van het overdragen van de visuele aandacht van anderen — in de vorm van hun blikrichting — in synchrone, interactieve gemedieerde groepscommunicatie (en samenwerken op afstand). In het gebruik van huidige gemedieerde systemen (zoals telefoons en video-vergadersystemen) voor groepscommunicatie kunnen, volgens kwalitatieve voorstudies, problemen ontstaan in o.a. het reguleren van het beurtwisselingsproces en het gebruik van deiktische verwijzingen (zoals "Wat denk *jij*?"). Onze vooronderstelling was dat deze problemen veroorzaakt worden door een gebrek aan aandachtsgerelateerde informatie in dit soort systemen. Hierdoor zou het moeilijk zijn voor gespreksgenoten om de dialogische aandacht van anderen vast te stellen: tegen wie anderen spreken, of naar wie anderen luisteren. Informatie over de blikrichting van anderen zou wellicht functioneren als de voornaamste manier om de dialogische aandacht van anderen over te dragen, op een manier die de verbale communicatie van inhoudelijke informatie niet hindert. Wellicht met uitzondering van handgebaren, zouden andere non-verbale visuele middelen (zoals gelaatsuitdrukkingen etc.) geen significante bijdrage leveren tot dit doel. De empirische studies in dit proefschrift bevestigden de bovenstaande vooronderstellingen grotendeels.

In de eerste studie, gepresenteerd in hoofdstuk 3, is onderzocht in hoeverre de focus van visuele aandacht van anderen effectief zou zijn als indicatie van de focus van hun dialogische aandacht in groepsgesprekken tussen vier personen. Met behulp van een oogmeter is het percentage van de tijd vastgesteld dat proefpersonen keken naar het gelaat van de gesprekspartner(s) tegen wie zij spraken, of naar wie zij luisterden. Dit is vergeleken met het percentage van de tijd dat proefpersonen keken naar het gelaat van anderen dan de persoon tegen wie zij spraken, of naar wie zij luisterden. De conclusie is dat het aankijken van gelaten een goede indicatie geeft van de dialogische aandacht richting individuen in groepsgesprekken. Als iemand luistert naar een individu, is de kans dat het aangekeken individu de persoon is naar wie men luistert ongeveer 88 procent. Als iemand spreekt tegen een individu, is deze kans ongeveer 77 procent. In deze situatie is ongeveer 1.6 maal meer aangekeken tijdens luisteren (62%) dan tijdens spreken (40%). Als een spreker meer dan één persoon adresseert, lijkt de indicatiefunctie van aankijken behouden te blijven. Wanneer drie personen aangesproken worden, stijgt het totaal percentage van aankijken significant tot ongeveer 59% van de tijd. In zulke situaties is de hoeveelheid visuele aandacht ontvangen door individuele luisteraars (20%) daardoor nog altijd significant groter dan de hoeveelheid aandacht die men zou hebben ontvangen als men niet zou zijn geadresseerd (12%). Onze inschatting van de effectiviteit van visuele aandacht als indicator van dialogische aandacht is

echter onderhevig aan individuele verschillen in persoonlijkheid. Zo hebben introverte individuen de neiging om relatief meer te kijken naar *andere* personen dan de spreker naar wie zij luisteren, dan extraverte individuen. Autonome individuen hebben de neiging relatief *minder* te kijken naar een persoon voor wie zij dialogische aandacht hebben, dan niet-autonome individuen. Tevens kunnen onze schattingen alleen gegeneraliseerd worden naar situaties waarin kijken naar taakobjecten niet vereist is.

In de tweede studie, gepresenteerd in hoofdstuk 4, is onderzocht of de aanwezigheid van een representatie van visuele aandacht — in de vorm van blikrichting — een geïsoleerd effect heeft op gemedieerde groepscommunicatie. De tweede factor was de aanwezigheid van andere non-verbale visuele uitingen van het menselijk bovenlichaam, zoals overgedragen door huidige video-gemedieerde systemen. Zowel subjectieve als objectieve parameters van het communicatieproces werden gemeten tijdens interactie in groepen van drie participanten (één proefpersoon en twee acteurs). De taak was het oplossen van taalpuzzels in drie gemedieerde condities (die alle geluid weergaven):

1) Simuleerde het gebruik van een normaal video-gemedieerd systeem met bewegend beeld en enkele camera's, waarin alle non-verbale visuele uitingen van het bovenlichaam van de acteurs correct werden overgedragen behalve hoofdoriëntatie en aankijken in het gezicht van proefpersonen.
2) Simuleerde het gebruik van een video-gemedieerd systeem met bewegend beeld en meerdere camera's, waarin alle non-verbale visuele uitingen van het bovenlichaam van de acteurs correct werden overgedragen (inclusief hoofdoriëntatie), behalve het aankijken in het gezicht van proefpersonen.
3) Simuleerde het gebruik van een systeem met stilstaande plaatjes, met de hand geselecteerd door de acteurs, waarin hoofdoriëntatie, aankijken in het gezicht van proefpersonen en fysieke verschijning werden weergegeven, maar geen overige non-verbale visuele uitingen.

Uit de resultaten blijkt dat de aanwezigheid van informatie over blikrichting in de vorm van hoofdoriëntatie een significante stijging met een factor twee veroorzaakte in het aantal deiktische verbale verwijzingen naar personen (bv. "Wat denk *jij*?"). Dit was vermoedelijk het gevolg van verschillen tussen condities in de inschatting van proefpersonen van de effectiviteit van het wijzen met het hoofd tijdens deiktische verwijzingen. De aanwezigheid van non-verbale visuele uitingen van het bovenlichaam anders dan blikrichting had geen significant effect op onze afhankelijke variabelen. Echter, een significant positief lineair verband is aangetoond tussen het percentage van de tijd dat de proefpersonen werden aangekeken in het gelaat en het aantal sprekerswisselingen (r=.37) en beurten door proefpersonen (r=.34). De aanwezigheid van een representatie van de visuele aandacht van anderen verhoogt dus de beurtfrequentie, maar alleen als ruimte- en tijdsaspecten van aankijken in het gelaat correct worden overgedragen. Uit de resultaten in conditie 3 blijkt dat een toename van het aantal beurtwisselingen met 25 procent

mogelijk is als aankijken in het gelaat wordt overgedragen. Het vermoeden bestaat dat de gevonden stijging in beurtfrequentie een indicatie is van een efficiënter of natuurlijker verloop van het beurtwisselingsproces.

Het anderen aankijken in het gelaat heeft vermoedelijk drie belangrijke, onlosmakelijk verbonden functies. De eerste twee functies betreffen een communicatie van interesse, de derde functie het waarnemen van deze en andere visuele informatie:

1) *Het weergeven van dialogische aandacht.* Aankijken is een goede non-verbale manier om dialogische aandacht te communiceren. Hoewel het bewijs indirect is, lijkt het waarschijnlijk dat deze functie bijgedragen heeft aan de vondst van een verhoogde beurtfrequentie in hoofdstuk 4. Een duidelijke aanwijzing hiervoor werd gevonden in de vragenlijsten van die studie. Proefpersonen vonden het gemakkelijker te observeren wie tegen wie sprak in de conditie waarin het percentage aankijken in het gezicht het hoogst was (conditie 3).
2) *Het reguleren van de sociale intimiteit.* Het in de ogen aangekeken worden lijkt een direct effect te hebben op het arousal-niveau. Als zodanig is aankijken één van de meest intieme handelingen die uitgevoerd kunnen worden op afstand. Door het aankijken te zoeken of vermijden, lijken gesprekspartners de arousal van zichzelf en van anderen op een gemeenschappelijk passend niveau te houden (een zogenaamd *Evenwicht van Intimiteit*). Hoewel hiervoor geen bewijs is gevonden, kan deze functie hebben bijgedragen aan de vondst van een verhoogde beurtfrequentie in hoofdstuk 4.
3) *Selectieve (visuele) aandacht.* De selectie van relevante visuele informatie mag zeer waarschijnlijk worden beschouwd als de belangrijkste reden om het gelaat van anderen aan te kijken. Het gezicht is één van de rijkste en meest relevante bronnen van visuele informatie — bijvoorbeeld over voornoemde functies — van het menselijk lichaam gedurende communicatie. Hierdoor is het niet vreemd dat dit deel van het lichaam gemiddeld het meest wordt aangekeken. Deze functie kan echter geen verklaring vormen voor de resultaten uit hoofdstuk 4.

We concluderen dat het overdragen van het aankijken van het gelaat van anderen een vereiste lijkt voor het correct ondersteunen van groepscommunicatie in gemedieerde systemen. De volgende informatie zou in deze systemen moeten worden weergegeven:

1) *Relatieve positie van participanten;*
2) *Hoofdoriëntatie van participanten;*
3) *Het aankijken van gezichten van participanten.*

Als bewegend beeld wordt overgedragen, is het in principe mogelijk aan bovenstaande vereisten te voldoen door gebruikmaking van videotunnels.

Hiermee kan een videocamera op dezelfde locatie worden geplaatst als de representatie van de ogen van een participant. Dit leidt echter wel tot een bijna kwadratische stijging met het aantal gebruikers van het aantal unieke video-verbindingen, waardoor gebruik van Multicast technieken voor de compressie van informatie op het onderliggende netwerk onmogelijk wordt. Als zodanig is het overdragen van blikrichting dmv. ruimtelijk gepositioneerde videocamera's niet goed schaalbaar met het aantal deelnemers, een belangrijke vereiste voor het gebruik van gemedieerde systemen voor *groeps*communicatie.

In hoofdstuk 5 van dit proefschrift wordt een oplossing voor dit probleem gepresenteerd gebaseerd op het gescheiden overdragen van beeld en blikrichting, naast geluid. Het GAZE Groupware System maakt gebruik van geavanceerde, op het bureau geplaatste oogmeters voor het meten van kijkgedrag van individuele participanten tijdens een gemedieerd gesprek. Dit kijkgedrag wordt vervolgens metaforisch gerepresenteerd in een virtuele vergaderkamer op het World-Wide Web. Dit gebeurt door het ruimtelijk oriënteren van een foto van iedere participant aan de hand van de gemeten kijkrichting van die participant (zie de foto op de kaft van dit proefschrift). Tevens kan het systeem exact weergeven waar participanten kijken binnen gezamenlijke documenten.

De belangrijkste voordelen van het GAZE Groupware System zijn:

1) Een geïntegreerde weergave van (visuele) aandacht van participanten voor andere participanten en gezamenlijke taakobjecten;
2) Het impliciet vergaren van informatie over de (visuele) aandacht van participanten;
3) Het overdragen van kijkgedrag vergt nauwelijks extra bandbreedte;
4) Een lineaire schaalbaarheid van bandbreedtegebruik met het aantal participanten, ook indien bewegend beeld zou worden overgedragen;
5) Het gebruik van natuurlijke metaforen voor het overdragen van informatie over de aandacht van anderen.

Appendix A

Glossary of Terms

Articulatory Attention	Speaking to one or more persons. A mode of *dialogic attention*.
Attentive State	A description of someone's focus of attention during an activity. At a syntactical level this involves describing the spatial and temporal properties of someone's (visual) attention, at a semantical level which actions, objects or people someone is attending to.
Audio	A medium for sound.
Auditory Attention	Listening to a person. A mode of *dialogic attention*.
Awareness	Knowledge about the attention of others in distributed communication and collaboration.
Communilaboration	Contraction of *communication and collaboration*.
Conversational Awareness	Knowledge about the (dialogic) attention of others towards persons during communication. See also *Selfcentric — .*
Cue	An act that conveys information from one human to another.
Deixis	The pointing or specifying function of some words (as definite articles and demonstrative pronouns) whose denotation changes from one discourse to another [159]. For example: "Look at *this*."
Dyadic	Involving 2 persons.
Dialogic Attention	Listening to a person (auditory mode of dialogic attention) or speaking to one or more persons (articulatory mode of dialogic attention). The *focus* of dialogic attention identifies these persons, the *extent* of dialogic attention describes the number of persons within this focus.
Eye-contact	Occurs when two persons gaze at each other's eye region. Virtually synonymous with *mutual gaze*.
Eye not found	A message generated by an eyetracking device indicating that it could not locate the pupil or corneal reflection.
Eyetracker	A device for measuring the orientation of the pupil(s), often relative to a visual scene.
Fixation	Prolonged foveation by suspension of eye movement. This allows detailed higher-level processing of visual information.
Fovea	The small area of the retina equipped for acute vision.
Foveation	The process of centering a retinal image for acute observation by the fovea.

Gaze	The act of looking at other humans [6]. Since the facial region is typically the focal point in this act, the term is virtually synonymous with *gaze at the facial region*.
Gaze at the facial region	The act of looking at the face of other humans. Often abbreviated as *gaze*.
Gaze direction	The absolute direction of looking, as constituted by the summation of body, head and eye orientation.
Groupware System	A computer based system that supports groups of people engaged in a common task, providing an interface to a shared environment [45].
Head orientation	The angular direction of the head, as a component of gaze direction.
Macro-level Awareness	Knowledge about the attention of others prior to, after or outside the scope of a synchronous distributed communication and collaboration session.
Micro-level Awareness	Knowledge about the attention of others during a synchronous distributed communication and collaboration session. Includes *Conversational* and *Workspace Awareness*.
Mutual gaze	Occurs when two persons gaze at each other. Virtually synonymous with *eye-contact*.
Multiparty	Involving more than two persons.
Nonverbal cues	All cues expressed without words, including paralinguistic expression.
Paralinguistic expression	Nonverbal use of vocal cues (such as pitch, timbre and loudness of voice through time) that accompany or modify the phonemes of words and that may communicate meaning [159].
Phoneme	Any of the abstract units of the phonetic system of a language that correspond to a set of similar speech sounds (as the velar /k/ of *cool* and the palatal /k/ of *keel*) which are perceived to be a single distinctive sound in the language [159].
Phonemic Clause	A basic syntactic unit of speech, as a string of 2-10 words in which there is one primary stress and which is terminated by a juncture, a slight slowing of speech, with slight intonation changes at the very end [81].
Point of gaze	A representation of the orientation of the pupil(s) in terms of coordinates within a visual scene.
Relaxed-WYSIWIS	The relaxation of WYSIWIS constraints on the following dimensions: display space, display time, subgroup population, and congruence of view [136]. See also *WYSIWIS.*
Saccade	The rapid, ballistic, reorientation of the pupil that occurs in between fixations in order to foveate a new area of the visual field.
Selfcentric Conversational Awareness	Knowledge about the (dialogic) attention of others towards *oneself* during communication. See also *Conversational Awareness*.

Simultaneous Speech	Overlapping talkspurts, produced by more than one speaker.
Speaker Switch	The act of exchanging the role of speaker and listener. Occurs when a new speaker has a talkspurt of one or more phonemic clauses, with others (including the previous speaker) being silent for at least one phonemic clause.
Talkspurt	A series of phonemic clauses by the same speaker.
Turn	A series of talkspurts, bounded by a speaker switch, *including* the silence that may occur before the speaker switch. See also *Utterance*.
Turntaking	The continuous process of exchanging the role of speaker and listener. This allows, as a rule, only a single person to speak at any moment in time.
Utterance	A series of talkspurts, bounded by a speaker switch, *not including* the silence that may occur before the speaker switch. See also *Turn*.
Verbal cues	All cues expressed with words, with the exception of paralinguistic expression.
Video	A medium for motion images.
Video-Mediated System	A system in which motion images, typically of humans, are conveyed, typically with sound, in order to communicate or collaborate across large distances.
Video-Mediated Communication (VMC)	Communication using a *Video-Mediated System*.
Workspace Awareness	Knowledge about the attention of others towards tasks during collaboration.
WYSIWIS	"What You See Is What I See". The provision of a consistent and coordinated display of shared information across participants during distributed work [137]. See also *Relaxed-WYSIWIS*.

Appendix B

Sample Materials

This appendix provides examples of materials and procedures used to conduct the experiment presented in Chapter 4 of this thesis.

SCORING DEICTIC 2nd-PERSON PRONOUNS

This section provides examples of our procedure for scoring deictic 2nd-person pronouns. Examples are presented in Dutch only. For an English-language overview of the rules used for scoring, see Chapter 4, *Analysis* section, page 98. Use of the word "je" (English: "you") was only scored if a 2nd-person singular deictic reference ("jij") was intended. Examples where internal context was *required* to ascertain this are printed in italics. Where possible, context is given between parentheses.

Utterance	Score	Interpretation
als jij denkt, jouw zin was	TRUE 2x	als jij denkt, jouw zin was
je... je zei net	TRUE 1x	jij... jij zei net
wat was jouw zin?	TRUE	wat was jouw zin?
weet je dat zeker?	TRUE	weet jij dat zeker?
je moet het gewoon intypen	TRUE	jij moet het gewoon intypen
dus je hebt die andere niet ingevoerd?	TRUE	dus jij hebt die andere niet ingevoerd?
hoe bedoel je?	TRUE	hoe bedoel jij?
wat was jouw zin?	TRUE	wat was jouw zin?
heb je er nu 3?	TRUE	heb jij er nu 3?
kun je hem nog een keer doen?	TRUE	kun jij hem nog een keer doen?
als je dat wilt testen	TRUE	als jij dat wilt testen
je kunt snel typen	TRUE	jij kunt snel typen
hoeveel heb je er nu?	*TRUE*	*hoeveel heb jij er nu?*
je kan hem nog een keer doen?	*TRUE*	*jij kan hem nog een keer doen? (intypen)*
je mag het proberen	*TRUE*	*jij mag het (van mij) proberen*
die kun je typen	*TRUE*	*die kun jij (in)typen*
die jij had,... Mirjam	*TRUE*	*die jij had, (pauze), Mirjam*
die kun je proberen	*FALSE*	*die kunnen we proberen*
dus die kun je voorop doen	*FALSE*	*dus die (zin) kunnen we voorop doen*
je kunt het proberen	*FALSE*	*we kunnen het proberen*
anders moet je maar proberen	*FALSE*	*anders moeten we het maar proberen*
wat had jij, Bart?	FALSE	verwijzing + naam direct achter elkaar
die andere twee kun je wisselen	FALSE	die andere twee kunnen we wisselen
dan kun je daar nog mee schuiven...	FALSE	dan kunnen we daar nog mee schuiven...
als je begint met...	FALSE	als we beginnen met...
dus dan mag je niet zeggen	FALSE	dus dan mogen we niet zeggen
kan je doen	FALSE	kunnen we doen
en die kun je ook omdraaien	FALSE	en die kunnen we ook omdraaien
je moet alleen die stukjes verschuiven	FALSE	we moeten alleen die stukjes verschuiven
als je zegt "het KNMI"	FALSE	als we zeggen "het KNMI"
maar als je zegt	FALSE	maar als men zegt
je kan zeggen "het KNMI verwacht"	FALSE	we kunnen zeggen "het KNMI verwacht"
haar/zijn	FALSE	alleen 2e persoon is correct
jullie	FALSE	alleen enkelvoud is correct

LANGUAGE PUZZLES

This section provides the sentence fragments used as language puzzles in the experiment presented in Chapter 4. Sentence fragments are presented in Dutch only. For an English-language example, see Box 4-1 on page 91.

Below, puzzles are presented in the same order as during the experiment. The first presented permutation of segments is always correct. Other sentences deemed correct are provided as permutations of fragment numbers in the fourth column. The assignment of fragments to conversational partners was random, but the same for each session.

	Fragment 1	**Fragment 2**	**Fragment 3**	**Other Correct Permutations**
1	de schoonmaakster	sloeg zijn kamer	altijd over	
2	als je het niks vindt	gooi ik het	op Internet	132, 213, 231, 312, 321
3	de laatste loodjes	wegen	het zwaarst	213, 321
4	kunnen wij	geen dag zonder	het koffiezetapparaat	231
5	ondanks	het mooie weer	was het koud	312
6	zonder een paraplu	gaan we	niet weg	213, 213
7	het KNMI verwacht	hier en daar	een bui	132
8	omdat het regent	gaan wij	maar niet	312, 321, 231, 213
9	niet alleen	met de auto	kunnen wij reizen	312
10	zijn verjaardagskaart	komt	veel te laat	231, 213, 321
11	nooit	ging de kat	op de kattenbak	213, 321
12	hij zat	niet onderuit gezakt	op zijn stoel	132
13	de melkboer	had	geen melkflessen	213, 321
14	bij de buurman	is het gras	groener	321, 312, 231, 213
15	hoge bomen	vangen	veel wind	321, 213

QUESTIONNAIRE

This section provides an English translation of the questionnaire used in the experiment presented in Chapter 4. Each subject was asked to answer this questionnaire immediately after the experimental session.

1. Did you find it hard to create those sentences with the three of you?
   very hard *hard* *undecided* *easy* *very easy*

2. Did you work together previously with the other two persons?
   never *sometimes* *regularly* *often* *very often*

3. Did you find the collaboration with the two partners pleasant?
   very unpleasant *unpleasant* *undecided* *pleasant* *very pleasant*

4a. Did you find this way of communicating pleasant?
   very pleasant *pleasant* *undecided* *unpleasant* *very unpleasant*

4b. Can you explain why?

5a. Did you find this communication system easy to work with?
   very hard *hard* *undecided* *easy* *very easy*

5b. Can you explain why?

6. Was it always clear whom your partners were talking to?
   very clear *clear* *undecided* *unclear* *very unclear*

7. Could you easily see what your partners were looking at?
   very hard *hard* *undecided* *easy* *very easy*

8. Would you like to collaborate this way again? Please motivate your answer.

9. Do you have any further comments?