Paper submitted to Discourse Processes
Please contact authors before citing

# A look is worth a thousand words: full gaze awareness in video-mediated conversation

Andrew F. Monk and Caroline Gale[1]
University of York

[1] Now at BTexaCT, Adastral Park, Martlesham Heath, Ipswich, UK

*Contact:*
Andrew Monk
Department of Psychology,
University of York, York, YO1 5DD, U.K.
Tel. +44 1904 433148, Fax: +44 1904 433181
Email: A.Monk@psych.york.ac.uk

*Running head:*
A look is worth a thousand words

**A look is worth a thousand words: full gaze awareness in video-mediated conversation**


## Abstract

Full gaze awareness, defined here as knowing what someone is looking at, might be expected to be a powerful communicative resource when the conversation concerns some object of common interest in the environment. This paper sets out to demonstrate this possibility in the context of video-mediated communication. An experiment is reported where pairs complete a communication task using a novel apparatus that supports full gaze awareness and mutual gaze (eye contact). This "GA Display" is contrasted with two control conditions, mutual gaze without full gaze awareness and audio only. The GA Display reduced the number of turns and number of words required to complete the task by about half in comparison with the two control conditions. The results of a subsequent Conversational Games Analysis suggest that at least part of this saving comes about because full gaze awareness provides an alternative non-linguistic channel for checking one's own and the other person's understanding of what is said.


# Introduction

*Full gaze awareness*

Under the right circumstances, people can distinguish with some accuracy what someone else is currently looking at in their immediate environment. Elsewhere we report an experiment where an estimator had to guess what position a gazer was looking at on a flat stimulus between them (Gale & Monk, 2000). The mean root mean square error of estimation equated to a change in the position of the gazer's head-and-eye position of 3.8˚ of pan and 2.6˚ of tilt. This low degree of error was essentially the same in a video-mediated condition and does not depend on allowing the estimator to see the head-and-eye movement to the target position. We describe this ability to gauge the current object of someone else's visual attention as full gaze awareness. We contrast it with partial gaze awareness, and mutual gaze. Partial gaze awareness is knowing the general direction someone is looking, for instance, up or down, left or right. Mutual gaze is knowing whether someone is looking at you. This is more commonly known as eye contact and has some well documented functions in regulating conversation. It is used in the process of transferring the role of speaker smoothly from one participant to another (Duncan & Niederehe, 1974; Goodwin, 1981;

Kendon, 1967; Levine & Sutton-Smith, 1973); it can also act as a social cue (Argyle, Lefebvre & Cook, 1974).

Full gaze awareness has received much less attention than mutual gaze in the research literature on language use. Studies of gaze following in monkeys (Anderson, 1996) and very young children (Butterworth & Jarrett, 1991) suggest that full gaze awareness may be a pervasive and primitive cognitive mechanism that precedes verbal language. Clark (1996) uses full gaze awareness as an example of how adult language uses many signals in addition to words and sentences. He describes how head-and-eye movements can be used to indicate references in conversation, for instance, "I want you [gazes at A] and you [gazes at B] to come with me".

The aim of the experiment described here is in the spirit of Clark (1996) and others who see language as a collaborative activity (see for example: Grice, 1957; Schegloff, 1991). According to this account, conversation is only possible because participants have implicit obligations to one another. As a speaker, one has an obligation to design utterances for the hearer and to monitor the hearer's subsequent utterances and behaviour for problems. As a hearer, one has an obligation to signal when one cannot understand the speaker sufficiently for current purposes. The possibility we wish to raise in this paper is that in certain contexts full gaze awareness may be a further resource in this process. Imagine, for example, an engineer, Ann, who is an expert on some piece of equipment, explaining its function to a novice, Ben, who knows little about it. By monitoring the appropriateness of where Ben is looking Ann can judge if he is understanding what is being said. Similarly, Ben can monitor where Ann is looking to get extra information regarding what she is talking about. Such a function for gaze has not been previously demonstrated in a quantitative experiment.

*Video-mediated conversation and full gaze awareness*

The applied context of this work is the design of multi-media communication equipment. It is commonly the case that, in the kind of work activities that video and audio links are designed to support, people use drawings, documents or a whiteboard to communicate and coordinate the work. Tang (1991) points out how a shared visual artefact can be used as a conversational resource in design meetings to store information, express ideas and mediate communication. Whittaker, Geelhoed and Robinson (1993) discuss how shared electronic workspaces can serve similar functions. Indeed, studies that have compared the value of shared data (teledata) with the value of an image of the person one is talking to (telepresence) have shown large effects in favour of shared data (Anderson, Mullin, Katsavras, McEwan, Grattan, Brundell, et al., 1999a; Gaver, Sellen, Heath & Luff, 1993; Kraut, Miller & Siegel, 1996). This need for shared visual artefacts suggests there may be a place for supporting full gaze awareness in mediated communication.

In video-mediated communication full gaze awareness is often not possible because the view provided excludes the immediate environment. A convention has grown up of setting the focal length of a camera so that the image includes only the other person's head and shoulders. This can support only partial gaze

awareness as one cannot see the objects they may be looking at. Were a wider angle view to be provided, including at least the top half of the body of the other person and thus some of their immediate environment, then full gaze awareness would become possible. Of course, this means that some of the limited visual resolution that had been devoted to depicting the face is now given over to depicting the environment. The arguments to be made for having a high definition full movement image of facial detail are that this is necessary to discriminate subtle facial expressions and to support lip reading. However, the importance of subtle nuances in facial expression may be over stated in most task contexts (Daly-Jones, Monk & Watts, 1998; Whittaker, 1995). Also, if a multi-media communication link makes lip reading necessary it suggests that there is something seriously wrong with the sound quality. It may then be better to devote available effort and bandwidth to improving the poor sound quality rather that worrying about video at all.

*Video-mediated communication and mutual gaze*

Video conferencing equipment does not typically allow mutual gaze because of a discrepancy between the camera position and the image of the other person's eyes. Commonly, a camera will be mounted on top of the monitor displaying an image of the person one is talking to. Mutual gaze is difficult to achieve with this configuration. If one looks at the other person's eyes in the monitor, the vertical discrepancy between camera and the facial image results in the impression that you are looking at their chest. Looking at the camera gives a strong illusion that one is looking them in the eyes, but then one can no longer judge whether they are looking at you. Thus, the only way that mutual gaze could be achieved would be if the participants were to learn that, under these circumstances, when the other person appears to be looking at a particular position on their chest they are really making eye contact. This is not at all natural and users commonly comment on the problem of "making proper eye contact". The video tunnel, described below, circumvents this problem of providing true mutual gaze in a two party conversation (Buxton & Moran, 1990).

In summary, there is plenty of evidence that mutual gaze is one of the conversational resources we rely on to successfully collaborate in conversations under normal circumstances. Full gaze awareness has been studied far less extensively but again there is good reason to think that knowing the focus of someone else's visual attention would have similar advantages. However, to supporting mutual gaze or full gaze awareness with electronically mediated communication it is not straightforward. The next section describes some previous attempts to do this. It also describes the novel apparatus that supports full gaze awareness and mutual gaze to be used in the experiment described in the rest of the paper.

*Mediating full gaze awareness and mutual gaze*

Various ingenious inventions have been suggested to mediate full gaze awareness using multi-media communication equipment. For example, Velichkovsky (1995) used an eye tracker to detect and signal the visual attention of another person artificially. A circle indicating where the other person is looking is projected onto the shared electronic workspace. Velichkovsky's

invention can thus be thought of as artificial full gaze awareness. Vertegaal (1999) also used an eye movement monitor to signal visual attention in this way but only while the remote participant was looking at the shared electronic workspace. In addition, participants were represented by recorded pictures or personas that were rotated to give the illusion they were looking either towards each other, or towards the viewer, depending on where the eye movement monitor computed they were looking. Both of these systems can provide artificial full gaze awareness over a narrow bandwidth link as they do not require full video signals to be transmitted. Vertegaal's GAZE system is particularly impressive as it solves the problem of gaze awareness in multi-party conversations where gaze awareness may be particularly important in regulating turn taking (see also: Sellen, 1995; Okada, Maeda, Ichikawaa & Matsushita 1994).

The video tunnel is an invention to circumvent the problem of providing true mutual gaze in a two party conversation (Buxton & Moran, 1990). This uses a half-silvered mirror to put the camera in the same virtual position as the monitor. Each participant sees the image of the other through this half silvered mirror. The camera sees their face reflected in it. By adjusting the position of the camera and face it is thus possible to look at the eyes of the other participant and at the camera at the same time. For this to be effective the position of the participants has to be restricted. In the original Xerox design this was achieved by putting a long cowl on the front, hence the description video tunnel.

Ishii, Kobayashi & Grudin's (1993) Clearboard 2 design provides mutual gaze in a similar manner. Here, the image of one's conversational partner is projected onto the underside of a transparent drawing surface. The drawing surface is a half silvered mirror, the "board" of the title. The camera is placed so as to look down on the board thus capturing a reflection of the participant in the board. The board also contains a polarising sheet that prevents the multiple images that would arise if the camera picked up the image projected from the other camera.

Clearboard 2 also has the aim of providing full gaze awareness. This aim is best explained by describing their first prototype, Clearboard 0. Clearboard 0 was not video-mediated. Two people sit in the same room facing each other across a table. In between them is a transparent sheet of glass with the object they are conversing about written on it. If they both look at the drawing they are looking at the same object, though of course one of them will be looking at the wrong side and would not be able to read letters. If they look through the glass they can see each other's facial expressions and are also able to achieve mutual gaze as they normally would. If one person is looking at the drawing on the glass then the other can look through the glass and judge what part of the drawing they are looking at, that is, they can also achieve full gaze awareness. In order to provide the features of Clearboard 0 in a mediated context where the conversants are in different locations, Clearboard 2 simply displayed the drawing on the board at the same time as the image of the other person.

Some informal experiments by the current authors question how accurate this full gaze awareness could have been. The problem is that, unlike Clearboard 0, Clearboard 2 provides no spatial separation between the drawing and the image of the other participant. To estimate where someone is looking one needs to make two judgements: (a) the direction faced by the gazer's head and eyes, that

is, the angle of head-and-eye rotation of the gazer with respect to some arbitrary direction such as straight ahead and (b) the position of the object being looked at with respect to the head and eyes of the gazer. The latter information is absent if the objects being looked at are presented in the same plane as the image of the participant.

In our pilot experiments to measure gaze accuracy a target stimulus was mixed onto the image of the other person in a video tunnel. As with Ishii et al's (1993) Clearboard 2 the target stimulus is thus in the same plane as the image of the face, that is, actually on the same screen. Although the direction in which the gazer was looking could be estimated, the relationship between this and the objects being looked at was ambiguous or nonsensical. In our pilot experiments, gaze estimation accuracy was very poor, essentially at chance, and logically it is difficult to see how such an arrangement could provide anything better than partial gaze awareness, that is, whether the other person is looking up, down, left or right.

*The GA Display*

The above analysis of why full gaze awareness might be difficult to achieve with Clearboard 2 suggested the design used in the experiment to be described in this paper. This physically separates the objects being looked at from the image of the other participant, as in Clearboard 0, but in a mediated communication context. Figure 1 shows how this "GA Display" works. Each participant views the image of the other participant through a half silvered mirror in a conventional video tunnel arrangement. The camera is mounted to the side of the participant it views such that the mirror serves to right-left reverse the image. The focal length of the camera is adjusted so that the image of the participant's face is actual size. A translucent display is positioned half way between the participant and the image of the other participant. While both participants are effectively looking at the same side of the objects in this display, that is, they could both read writing on it, there is still full gaze awareness. This is because the mirror left-right transforms the image of the other person apparently behind the translucent display. When your partner looks at an object on your right hand side, that is, their left, they are actually looking to the right of the display that they can see, but the mirror transforms the image appropriately for full gaze awareness.

- - - - - -

Figure 1 about here

- - - - - -

Data to be presented in the results section show that this arrangement allows good gaze estimation accuracy so that the design provides both mutual gaze and full gaze awareness for any stimulus that can be viewed with an image of the other participant behind it. While this might be a problem with images that depend on small contrast changes, such as X-rays, high contrast images such as technical drawings and circuit diagrams can easily be viewed in this way. The depth of field of the eye is such that only the face or the object can be fully in focus at any one time.

*The experimental task and control conditions*

The theoretical objective of the experiment is to demonstrate that full gaze awareness can, at least under some circumstances, be another conversational resource in the cooperative activity of conversation. In particular, it is predicted to reduce the need to use verbal language to check one's own understanding or the understanding of your partner. The practical objective is to demonstrate the GA Display has certain advantages over control conditions in terms of linguistic efficiency.

The experimental task involved pairs of participants. One member of the pair was designated as the expert. Their task was to describe the position of a point on a slide that they could both see (see Figure 2 for an example). The other role, the receiver, had to say what point the expert was describing, but only once they were reasonably confident that they were correct. This task was chosen as one that would maximise the advantages of full gaze awareness. Using the GA Display the expert can monitor the receiver's current locus of visual attention and so can judge whether the instructions given are being understood. Similarly, the receiver can use where the expert is looking as a parallel source of information with which to check his or her understanding regarding the target point.

---

Figure 2 about here

---

The slides were all pictures selected to contain relatively abstract material that was difficult to describe, for instance, a circuit diagram or a schematic section of a rat's brain. In this way it was similar to the tangram task used by Schober and Clark (1989) and other authors. In all these tasks, communication eventually becomes very efficient as partners develop codes and systems for describing the stimuli. The prediction is that full gaze awareness will be most useful when verbal communication is least efficient, that is, when the pairs are still developing systematic ways of describing the stimuli. For this reason, rather than using the same stimulus throughout the experiment, there were 10 different slides, each used only once within the experiment.

The context for this research is the practical problem of designing multi-media communication facilities. For this reason the GA Display was compared with two video-mediated control conditions: (i) Video Tunnel Only, and (ii) Audio Only. The former control condition was achieved by reducing the size of the picture by half and presenting it at a different position for each member of the pair. The expert's drawing was printed onto the bottom right-hand quarter of the display while the receiver's drawing was printed onto the top left-hand quarter. Gaze accuracy data presented in the results shows that this very considerably reduces full gaze awareness yet all the information given by the video tunnel (mutual gaze and facial expressions) is still available. While the objects to be described are smaller they are still readily visible and it is difficult to see how this confound could account for the results obtained. The latter control condition was achieved by switching off the image of the other person's face. In this control condition there is no view of the other participant, that is, no mutual gaze, no full gaze awareness and no facial expressions.

While performance data (accuracy and time to completion) are reported here the most important data concern process (number of turns and number of words spoken). Performance data are rarely sensitive to manipulations of video mediation (Doherty-Sneddon, Anderson, O'Malley, Langton, Garrod & Bruce, 1997; Whittaker, 1995). On the whole, process data are also more revealing about the effects of the manipulation (Monk, McCarthy, Watts & Daly-Jones, 1996) and no attempt was made to make performance more sensitive as a dependent variable. The experimental instructions emphasised accuracy. The receiver was simply told to guess as soon as they were reasonably confident that they knew which pair of points the expert was describing. The expectation is that the receiver will guess correctly on most trials. As minimal time pressure was applied there was also no strong reason to expect effects on time to completion.

The process data speak to our theoretical and practical hypotheses. The GA Display is predicted to result in more efficient conversations, that is, less turns and less words to complete the task than in either of the control conditions where full gaze awareness is not possible. The comparison with the Video Tunnel Only condition is most relevant to the theoretical objectives of the experiment, because here the other video resources (facial expressions and mutual gaze) are still available. The Audio Only control condition is most relevant to the practical objectives. It provides a reference point, a control condition that is commonly used in experiments on video-mediated communication.

The theoretical prediction is that there will be a reduced need to use verbal language to check one's own understanding or the understanding of your partner when these communicative functions can be carried out using full gaze awareness. This prediction was tested using Conversational Games Analysis (Kowtko, Isard & Doherty-Sneddon, 1991). Conversational Games Analysis has been used successfully to evaluate mediated communication (Anderson, O'Malley, Doherty-Sneddon, Langton, Newlands, Mullin, et al., 1997; Doherty-Sneddon, et al., 1997). It proceeds by marking up each conversation to indicate the function of each utterance. Utterances are viewed as moves in different sorts of games that both speakers are participants in. The way this works, and the full set of games identified are set out in the Results section.

Align Games mark out a set of utterances where one person confirms the other person's understanding of an utterance. Check Games mark out a set of utterances where someone checks their own understanding by requesting confirmation from the other. The prediction then is that both Align and Check Games will be less frequent in the speech of pairs using the GA display as their function can be carried by full gaze awareness.

Finally, it should be mentioned that there was a second independent variable, audio quality. 81.5dBA noise was mixed into the sound channels for half the trials. Veinott, Olson, Olson & Fu (1999) found that participants who did not have English as a first language benefited more from a video link than participants who did. The result was interpreted as suggesting that video-mediation may be more important when verbal communication is difficult. It was thought that poor audio quality might have the same effect in our experiment. In the event the manipulation of audio quality was ineffective,

participants simply shouted to overcome the noise. This independent variable is therefore mentioned mainly for its impact on the experimental design.

## Method

### *Design*

The sampling unit in all the analyses presented here is the participant pair. Audio-visual configuration was a between-pairs variable thus, eight pairs completed the task using the GA Display, eight pairs the Video Tunnel Only control and eight the audio only control. Audio quality was a within-pairs independent variable. Pairs performed two trials with each of ten pictures. Half the pairs in each condition conducted the trials with the poor audio quality (noise) for the first five pictures and the good quality audio (no noise) for the second five pictures. The other half of the pairs experienced the audio quality conditions in the opposite order (first five pictures - no noise; second five pictures - noise). The pictures were presented in a different random order for each pair.

### *Participants*

48 participants were used in 24 pairs and assigned at random to one of the audio-visual configurations. They were recruited as pairs from the University of York and all participants knew their partner prior to the experiment. One member of each pair was randomly designated as expert and the other as receiver.

### *Materials*

The stimuli used in the communication task were ten real pictures that had elements that were difficult to describe. The ten pictures were: two circuit diagrams; two architectural blueprints of houses; two diagrammatic representations of the rat's brain; two extracts from sheet music printed backwards, and two electron microscope images featuring benzene rings, of which one is depicted in Figure 2.

Around 15 arbitrary points (15 plus or minus 2) were marked and labelled with letters on each of the receiver’s pictures. A previous experiment had used a single circuit board stimulus where both partners could see all the labelled points. It was found that pairs quickly learned to describe the points, rather than the stimulus. Accordingly, in this experiment the expert's transparent foil had only two points on it, the two points that had to be described to the receiver in that trial. This prevented the points being described only in terms of their position relative to other points, for instance, "the point nearest the bottom edge", instead they had to use features in the picture, as intended.

### *Apparatus*

Two video tunnels were constructed as described above (see Figure 1). Each participant sat 150 cm from a monitor, the mid-point of which was at eye level and marked with a coloured dot. They viewed the monitor through a half-silvered mirror, placed at an angle of 45˚ such that the middle of the mirror was also at eye level. The stimuli were placed in a vertical wooden frame, halfway

between the participant and monitor. A coloured dot marked the mid-point of each stimulus; this was again at eye level. Participants could thus line up the coloured dots on the stimuli and monitor to keep the position of their head and eyes constant. A tripod-mounted camera, also at eye level, was placed at 45˚ to the mirror and 90˚ to the participant. This captured the image of the participant's face as reflected by the mirror and sent it to the other participant's monitor. To prevent participants from seeing the camera reflected in the mirror, a wooden frame was placed over the whole set-up and black material draped over it.

Each participant's camera sent an input to an audio-video mixer (a Panasonic Digital AV mixer). Each of these sent one output to the other participant's monitor, and one output to a remote location where sound and vision were recorded on a standard VCR and cassette recorder. The pair communicated over an audio link using clip-on microphones connected to enclosed headphones. Participants sat on height-adjustable chairs. To constrain their movement wooden frames were constructed such that the participant could only hold in place if they kept relatively still. Due to constraints on available space, the expert and receiver actually worked in the same room with a screen between them.

For pairs in the GA Display and Audio Only video configurations three copies of the original picture were made onto A4-size clear acetate sheets. One copy was the receiver's and had around 15 labelled points added to it. The other two copies of each picture were the expert's. Each of these had two points marked, these were the two points that the expert had to describe in each of the two trials with that picture. Thus, in the course of the two trials four points in all were described.

For pairs in the Video Tunnel Only configuration, the same 10 pictures described above were used but reduced in size by half, that is, each stimulus was approximately 10.5 cm by 15 cm instead of 21 cm by 30 cm. The stimuli were identical to those described above in every respect apart from their size, that is, the same points were marked in the same relative positions. The crucial difference lay in the positioning of the stimuli on the A4 clear acetates. The expert's drawing was printed onto the bottom right-hand quarter of the acetate while the receiver's drawing was printed onto the top left-hand quarter. Thus, although the stimuli were identical in appearance, they were not placed at exactly the same location, and when either participant gazed at a point, their partner could not tell what they were looking at. This effectively supported partial gaze awareness, that is, information about the general direction of gaze (left versus right, up versus down) without supporting full gaze awareness. This view also supported mutual gaze and made facial expressions visible. Thus the Video Tunnel Only configuration allows a direct comparison between configurations where a view of the face was available with and without full gaze awareness.

*Procedure*

Participants were told that this was an experiment examining how people communicate over a video link, and what effects a high- and low-quality audio link have on this. No mention of gaze or gaze awareness was made. It was

explained that the experiment was in two halves, and that during one half they would have to communicate over a noisy audio channel. To ensure that both participants were comfortable with the noise level, a sample was then played and they were given the option of withdrawing from the experiment. No participants withdrew at this point.

The expert was told that their task was to describe to the receiver the location of a particular pair of points on various different pictures. In addition, the expert was asked to give feedback once the receiver has made a guess. It was emphasised to both participants that they could interrupt each other at any point, and that the receiver should do so as soon as they understood what point was being described. In the GA Display and Audio Only configurations, participants were told that their partner had the same view as they did. In the Video Tunnel Only configuration, both were told that although their pictures were identical, they were placed in different locations. After settling the participants in the apparatus, the chair height being adjusted to give a constant eye height, they performed one practice trial.

Each pair completed two trials, in each of which they described two points, with each picture. In order to save time changing the transparent sheets these trials were consecutive. After the pairs had worked through all 10 pictures (20 trials) they performed a gaze accuracy task. A stimulus similar to that used in the main experiment was created containing 20 randomly distributed points. The same transparent sheet was mounted in the video tunnels of both expert and receiver. The expert gazed fixedly at each of the 20 points in turn in a random order indicated by the experimenter and the receiver simply guessed which point was being looked at. The pairs in the Video Tunnel Only configuration did this twice, first with the same full size stimulus used by the other pairs, and second with smaller offset stimuli prepared as in the main part of the experiment.

## Results

*Manipulation checks: gaze accuracy and performance*

The gaze accuracy task, run after the main communication task, was included as a manipulation check, that is, to demonstrate that the GA Display did indeed provide full gaze awareness and the Video Tunnel Only configuration did not. Table 1 shows that the GA Display was extremely effective. Even the pairs who had not used it in the main part of the experiment got 92% correct. In contrast, with the Video Tunnel Only configuration less than 10% of the points were guessed correctly

| | |
|---|---|
| *Using GA Display* | |
| Pairs in GA Display condition | 18.63 (1.11) |
| Pairs in Video Tunnel Only condition | 18.25 (1.30) |
| Pairs in Audio only condition | 18.38 (0.99) |
| *Using Video Tunnel Only Configuration* | |
| Pairs in Video Tunnel Only condition | 1.88 (1.17) |

Table 1. Results from the gaze estimation trials carried out after the main experiment. Mean correct out of 20 (and standard deviation).

Table 2 gives the mean number of trials correct in the main part of the experiment. Performance is close to the maximum of 10 in all conditions, showing that the receivers were following the instructions and not guessing until they were relatively confident that they knew which point was being described. Completion time was measured using the time stamped video recordings. Table 3 gives the mean time for each audio condition, for the pairs in each video configuration. Here there is an apparent difference in favour of the GA Display and for the high quality audio condition. The completion time data were subject to a three-way split-plot analysis of variance. The between-pairs independent variables were video configuration and order of audio condition (poor quality audio followed by good quality audio or vice versa). The within-pairs independent variable was audio condition. In the event there was no significant main effect of video configuration ($F(2,18) = 1.87$, n.s.) or order ($F(1,18) = 2.59$, n.s.). There was a significant effect of audio condition ($F(1,18) = 4.91$, $p < .05$). There were no significant two- or three-way interactions.

| | High quality audio | Poor quality audio |
|---|---|---|
| GA Display condition | 9.5 (0.76) | 9.75 (0.46) |
| Video Tunnel Only condition | 9.5 (0.53) | 9.38 (0.74) |
| Audio Only condition | 9.88 (0.35) | 9.5 (0.76) |

Table 2. Mean number of trials correct out of 10 (and standard deviation) for pairs in each video configuration for each audio condition.

| | High quality audio | Poor quality audio |
|---|---|---|
| GA Display condition | 10.71 (2.87) | 11.45 (3.74) |
| Video Tunnel Only condition | 12.94 (2.38) | 13.73 (2.26) |
| Audio Only condition | 12.70 (2.68) | 12.94 (2.45) |

Table 3. Mean time to complete the ten trials in each audio condition in minutes (and standard deviation) for pairs in each video configuration for each audio condition.

As indicated above, it was not expected that measures of performance would be sufficiently sensitive to result in significant effects for the between-pairs variable video configuration. However, there was a main effect of the within-pairs variable audio condition showing that that this manipulation did, in some sense, make communication more difficult. It has already been noted that participants coped with the noise by shouting into the microphones. This may have made speech slower without having the desired effect on intelligibility. Anticipating the analyses to be presented in subsequent sections, there were no significant audio condition main effects or audio condition by video configuration interactions in the measures computed. There was no sign of the predicted interaction between audio condition and video configuration. The effect of video configuration was the same whether or not the noise was present. One simple interpretation of these results is that the audio conditions used were not effective in manipulating intelligibility.

*Measures of communication process*

An independent transcriber used audio recordings of the dialogues and a set of instructions from the experimenter to compile transcripts for all 24 pairs. The transcriber was required to write down everything that was said by each person, including non-words, back-channel responses and partial words, and to use no punctuation. Where overlap occurred, the transcriber marked the overlapping speech in square brackets. Pauses were also marked, defined as any break in speech longer than 0.5 seconds. Filled pauses such as "uuum" were also coded. These transcripts were verified against the audio recordings and corrected where necessary by the experimenter.

Using these transcripts the following counts were made: turns, overlaps, words spoken by each participant (not including non-words or half-words) and pauses. A *turn* was defined as any vocal utterance that was: (i) not a back-channel response (e.g., "mhm"), (ii) not a failed interruption (i.e., an attempt to interrupt that failed to result in the other person stopping their speech or responding to the content of the interruption) and (iii) not some non-verbal vocalisation such as a grunt or a sigh. Each of these measures of process was subject to the same three-way split-plot analysis of variance as time to completion. None of these analyses demonstrated a significant main effect of audio quality nor any significant interaction with audio quality.

Only the means for the three different video configurations thus need to be presented. There are dramatic differences between the GA Display configuration and the two control conditions for turns and words spoken by the expert (see Table 4), there being about twice as much talk in the control conditions. Similar effects are seen in the number of overlaps and pauses (see Table 5). There were no significant interactions.

| | Turns | Words E | Words R |
|---|---|---|---|
| GA Display (1) | 181.88 (69.79) | 1334.00 (488.69) | 362.13 (145.95) |
| Video Tunnel Only (2) | 405.63 (80.12) | 2662.13 (496.34) | 521.50 (121.01) |
| Audio Only (3) | 332.38 (120.21) | 2111.00 (882.90) | 534.38 (170.40) |
| F (2, 18) | 14.92 | 4.53 | 3.95 |
| p | <0.001 | <0.05 | <.05 |
| HSD | 75.38 | 593.54 | 123.25 |
| Sig. difference between | 1&2, 1&3 | 1&2, 1&3, 2&3 | 1&2, 1&3 |

Table 4. Mean number of turns and words spoken by the expert, E, and receiver, R, (and standard deviations) for the three video configurations. F is the F value for the video configuration main effect, HSD is Tukey's HSD.

| | Overlaps | Pauses E | Pauses R |
|---|---|---|---|
| GA Display (1) | 18.12 (12.40) | 61.38 (16.18) | 11.00 (4.11) |
| Video Tunnel Only (2) | 44.38 (8.57) | 95.13 (44.11) | 30.88 (12.40) |
| Audio Only (3) | 28.25 (8.72) | 110.88 (36.22) | 28.63 (11.11) |
| F (2,18) | 13.34 | 4.53 | 13.73 |
| p | < 0.001 | < 0.05 | < 0.001 |
| HSD | 9.25 | 30.34 | 7.50 |
| Sig. difference between | 1&2, 1&3, 2&3 | 1&2, 1&3 | 1&2, 1&3 |

Table 5. Mean number of overlaps and pauses made by the expert, E, and receiver, R, (and standard deviations) for the three video configurations. F is the F value for the video configuration main effect, HSD is Tukey's HSD.

The considerable reduction in the number of turns required to complete this task when using the GA Display is the most important result in this paper: there is simply less talk when full gaze awareness is available to the participants. The other measures of process follow this primary result. The large effect on the number of words spoken by the expert probably arises because the task requires much longer turns from the expert than the receiver. Interestingly there is a much smaller difference between the two control conditions, Video Tunnel Only and Audio only. It would appear, that for this task, being able to see your partner's facial expression and being able to achieve mutual gaze is of much less importance than being able to see what they are looking at in the task domain. The effect on the number of pauses in the speech also follows the turns data, as one might expect. Table 6 gives overlaps and pauses per 100 turns. When reported in this way there are no significant effects of video configuration.

| | Overlaps / 100turns | Pauses E / 100turns | Pauses R / 100turns |
|---|---|---|---|
| GA Display (1) | 9.47 (4.92) | 40.23 (18.85) | 7.00 (3.65) |
| Video Tunnel Only (2) | 13.25 (5.14) | 28.73 (17.13) | 8.42 (3.27) |
| Audio Only (3) | 8.94 (2.84) | 34.41 (10.56) | 8.89 (2.15) |
| F (2,18) | 2.09 | <1 | <1 |
| p | n.s | n.s | n.s |

Table 6. Mean number of overlaps and pauses made by the expert, E, and receiver, R, per 100 turns (and standard deviations) for the three video configurations. F is the F value for the video configuration main effect.

*Conversational Games Analysis*

The view of conversation taken here is as a collaboration dependent on both parties continuously cooperating to monitor their own and the other person's conversation for trouble that may need to be repaired. Our hypothesis is that full gaze awareness may play a role in this process. By monitoring the receivers' current locus of visual attention the experts can check if their instructions are being understood. Similarly, receivers can use where the experts are looking as an additional source of information to verify their understanding regarding the point under discussion.

To test this hypothesis it is necessary to classify the function of different parts of the conversations that occurred. Conversational Games Analysis is a technique for dialogue analysis developed concurrently at the Universities of Edinburgh, Glasgow and Nottingham that categorises speech fragments according to their conversational function. It was designed on the basis of dialogues arising from the Map Task (Kowtko, et al., 1991) and has been applied to a variety of task-based dialogues (see for example Anderson, et al., 1997).

Conversational Games Analysis has three levels: games, moves and utterances. A game is composed of at least two, but potentially many more, moves, and a single utterance constitutes at least one move. Table 7 lists the different games that may be coded.

**INSTRUCT:** Communicates a request for action.

**CHECK:** The speaker checks his or her self-understanding by requesting confirmation that an interpretation is correct.

**ALIGN:** The speaker confirms the other's: understanding of an utterance, attention, agreement or readiness.

**QUERY - YES/NO:** A question which requires a yes / no response concerning new or unmentioned information (not checking the interpretation of a previous message).

**QUERY - W.** An open-ended question such as what, why or where regarding new information (not checking the interpretation of a previous message).

**EXPLAIN:** Freely offered information regarding the task not elicited by coparticipant.

Table 7. Six types of game found necessary and sufficient to code transcripts in this experiment (adapted from Anderson, et al., 1997).

In the experimental task used in this experiment, each trial consists of an Instruct Game as the expert is communicating a direct request to the receiver to name two points on the picture. The successful completion of this Instruct Game is evidenced by a Query Y / N Game initiated by the receiver embedded within this Instruct Game. An example from the transcripts is given below. Each game is bracketed by its start and end and each move is labelled after the utterance containing it. Thus, the reply-y move from the expert ends both the embedded Query Y / N Game and the Instruct Game. "*" indicates a pause of more than 0.5 seconds.

1 START GAME INSTRUCT

E: er first point if you take the chip in the very top right it would be at the bottom right corner of that chip in the space between the outer wiring um the second point would be if you go to the er from the centre * one chip over and then down to the cluster of transistors it's in the centre of the cluster

Move: {ready} instruct

R: okay

Move: acknowledge

2 START GAME QUERYY/N (EMBEDDED)

R: e and k

Move: query-y/n

E: yes you're right

Move: reply-y

END GAME 2

END GAME 1

This is the simplest form the dialogue could take to complete the trial. Most trials contain other embedded games, most notably games in which one partner checks their own or the other person's understanding. Checking one's own understanding is coded as a Check Game. The demands of this task meant that Check Games were always initiated by the receiver in this experiment. An example from the transcripts is given below. The receiver interrupts the Instruct move that was to communicate the position of the second point to check that she has understood what is meant by "at the top" in the Instruct move for the first point.

1 START GAME INSTRUCT

E: the first point's * in the middle and at the top * almost exactly in the middle

Move: {ready} instruct

R: mhm

Move: acknowledge

E: and * the * second point is

Move: instruct

2 START GAME CHECK (EMBEDDED)

R: you mean at the top * uh top of the screen yeah

Move: check

E: yeah

Move: reply-y

R: yeah

Move: acknowledge

END GAME 2

E: and * the second point is * about in the centre of the bottom left corner

Move: instruct

R: right

Move: acknowledge

Checking the other person's understanding is coded as an Align Game. The demands of the task meant that Align Games were always initiated by the expert in this experiment. An example from the transcripts is given below.

11 START GAME ALIGN (EMBEDDED)

E: you've got it

Move: align

R: I think so yeah

Move: reply-y

E: yeah

Move: align

END GAME 11

The final two types of game needed to needed to code all the utterances in the transcripts were Query-W and Explain Games. As an example of a Query-W Game consider the example below. The Query-W Game is embedded in an Align Game which is itself embedded in the Instruct Game that forms the trial. Square brackets indicate overlapping speech.

E: erm if er * it's a it's in er the m- centre of a group of three which are running in a diagonal towards the erm north east

Move: instruct

50 START GAME ALIGN (EMBEDDED)

E: * can you see the er there's a group of three circles [which are]

Move: align

51 START GAME QUERY W (EMBEDDED)

R: [how high up is it]

Move: query-w

E: it's um a bit below the middle but not particularly [if you]

Move: reply-w

END GAME 51

R: [right ok] got it

Move: reply-y

END GAME 50

In an Explain Game a participant gives information that is not elicited by the other participant as part of another game, and that is not a request for action (part of the Instruct Game that forms the trial). For example, the exchange given below starts with a Check Game initiated by the receiver utilising a code they have developed to describe the stimulus which is a flower. Following this exchange the receiver decides it would be useful to elaborate the code and does so in an Explain Game.

27 GAME CHECK (EMBEDDED)

R: on to the next um the signature petal kind of thing
Move: check

E: yeah
Move: reply-y

R: up to the next signature petal
Move: check

E: yeah
Move: reply-y

R: up to the small one
Move: check

E: and then up to the bigger one above that
Move: clarify

END GAME 27

28 GAME EXPLAIN (EMBEDDED

R: there's not a bigger one it's just getting smaller now * like coz I'm getting into the core [of the flower]
Move: explain

E: [oh right right]
Move: acknowledge

END GAME 28

*Results from the Conversational Games Analysis*

To re-iterate, the Conversational Games Analysis was carried out to determine whether there was a reduced need for Check and Align Games with the GA Display. Counts were made of the number of each of the types of game previously specified in Table 7. Each trial was defined as an Instruct Game with an embedded Query yes/no Game. The instruct game was the expert instructing the receiver to guess the two points being described, the Query yes/no Game being the guess. There were thus necessarily 20 Instruct Games and 20 Query yes/no Games additional to those listed in Tables 8 and 9.

The number of games per session for Check and Align Games, where a session is taken as the 10 trials in each audio condition, was subject to the same three-way split-plot analysis of variance as time to completion in and the process measures. Neither Check or Align Games showed a significant main effect of audio quality. Audio quality did interact with order of presentation of audio condition for both Check and Align Games but these interactions can both be interpreted as simple practice effects, such that participants use less of these games in the second half of the experiment. In neither case was there a three-way interaction. Only the means for the three different video configurations thus need to be presented in Table 8.

The advantage for the GA Display configuration is even more dramatic than that observed with the process measures. There are many fewer Check and Align Games with the GA Display configuration than with the two control conditions. Tukey's HSD test shows that the significant differences are between the GA Display configuration and the control configurations. For the Check Games this comparison is significant for both control groups. For the Align Games the comparison between GA Display and Video Tunnel only does not quite reach significance.

| | Check | Align |
|---|---|---|
| GA Display (1) | 6.12 (4.17) | 6.12 (6.19) |
| Video Tunnel Only (2) | 27.25 (11.4) | 19 (8.87) |
| Audio Only (3) | 17.88 (9.35) | 29.62 (16.3) |
| F (2, 21) | 9.10 | 8.12 |
| p | < 0.01 | < 0.01 |
| HSD | 11.71 | 14.07 |
| Sig. difference between | 1&2, 1&3 | 1&3 |

Table 8. Mean number of Check and Align Games per session (and standard deviation). F is the F value for the video configuration main effect, HSD is Tukey's HSD.

There were very few Query-W or Explain Games (see Table 9). Similarly, the number of Query Y/N Games additional to the twenty necessary to complete the task was very small. For this reason alone, the differences observed in Table 9 cannot explain the reduction in turns observed in Table 4.

The low frequency of these games makes these data very skewed, there being many zeros in the raw data. Accordingly the three-way analysis of variance used for Align and Check Games was not deemed appropriate. New counts were computed ignoring audio condition, that is, treating the session as all 20 trials and these scores submitted to a Kruskal-Wallace non-parametric test with the single independent variable video configuration with three levels. The only significant difference is in the number of Query-W Games. There are more of these in the Video Tunnel Only configuration that makes any attempt at full gaze awareness misleading.

| | Query-W | Explain | Query Y / N |
|---|---|---|---|
| GA Display (1) | 1.63 (0.74) | 1.00 (1.20) | 0.63 (0.92) |
| Video Tunnel Only (2) | 4.63 (2.88) | 0.88 (0.99) | 1.50 (1.69) |
| Audio Only (3) | 1.50 (1.69) | 0.88 (0.99) | 1.50 (2.07) |
| Chi-Square (d.f. = 2) | 9.05 | 0.04 | 1.25 |
| p | < 0.05 | n.s. | n.s. |

Table 9. Mean number of Query-W, Explain and Query Y / N Games extra to the 20 needed to complete the trials.

The dramatic effect observed for both Check and Align Games supports the hypothesis that full gaze awareness can perform the function of Check and Align games non-verbally. To further underline this point, the Conversational Games Analysis protocols were re-examined, counting moves rather than games. Moves equate closely to turns though some turns may contain more than one move. Table 10 gives the number of moves per 100 moves for Check Moves made by the receiver and Align Moves made by the expert. While, as one might expect, the differences between video conditions are less dramatic they are still significant. In both cases Tukey's HSD is only able to demonstrate a significant difference between the two most different means. As in Table 8, the Video Tunnel Only control condition results in the most Check Moves and the Audio Only condition the most Align Moves, though there is no significant difference between the two control groups in either moves or moves per 100 moves.

| | Check Moves / 100 (receiver) | Align Moves / 100 (expert) |
|---|---|---|
| GA Display (1) | 2.67 (1.31) | 2.56 (2.30) |
| Video Tunnel Only (2) | 6.20 (2.76) | 4.62 (3.45) |
| Audio Only (3) | 4.55 (1.98) | 6.93 (3.05) |
| F (2, 21) | 5.65 | 4.33 |
| p | < 0.05 | < 0.05 |
| HSD | 2.66 | 3.76 |
| Sig. difference between | 1 & 2 | 1 & 3 |

Table 10. Mean number of Check moves made by receiver and Align Moves made by the expert per 100 moves (and standard deviation). F is the F value for the video configuration main effect, HSD is Tukey's HSD.

## Discussion

The apparatus devised for this experiment provided all the information available to copresent conversational partners but in the context of multimedia communication. Partners had good access to the conversational resource of facial

expression through a life-size, television-quality image of the other partner at a distance of 1.5 m. They could establish true mutual gaze through the video tunnel and full gaze awareness through the spatially separated translucent displays (transparent sheets).

The apparatus also made it possible to selectively remove these resources. The Video Tunnel Only configuration lacks only full gaze awareness, whereas the Audio Only configuration lacks all visual information about the behaviour of one's partner. Comparison of the three configurations thus allows us to assess the potential value of the conversational resources available, in this experimental context. Two conclusions are drawn.

1. Making full gaze awareness possible considerably reduces the amount of talk and, even more notably, the degree to which participants need to verbally check their own and the other person's understanding of what has been said.

2. Any advantage provided by a view of the face (facial expression and mutual gaze) is smaller than the advantage provided by full gaze awareness and was not detectable in this experiment.

The remainder of this section discusses the theoretical and practical implications of these two results.

*The value of a view of the face: Video Tunnel compared with Audio Only*

While it is difficult to interpret null results, the lack of a difference between the Video Tunnel Only and Audio Tunnel Only configurations is worthy of comment. For most of the process measures and Conversational Games Analysis counts there was no significant difference. The lack of a need to read facial expression would be predicted for an information transfer task with little social input (Short, Williams & Christie, 1976). Mutual gaze might also be expected to be of lesser importance as turn taking should be easy with a well structured task with only two people conversing (Whittaker, 1995).

Other investigators have, however, obtained statistically significant differences between similar audio-visual configurations. Using their map task Doherty-Sneddon, et al. (1997) were able to demonstrate a significant difference in Align Games initiated by instruction givers, but not Check Games, when comparing video-mediated with Audio Only communication. There are a number of differences between Doherty-Sneddon, et al.'s (1997) experiment and that described here, particularly in the tasks used. The map task depends on the two participants not being able to see each others' maps as they differ in detail. Our task depends on being able to see identical stimuli. Also, in Doherty-Sneddon, et al.'s (1997) experiment video configuration was a within pairs variable. This leads to more sensitive statistical comparisons but could have resulted in subtle effects due to participants being able to experience and compare all three configurations.

Doherty-Sneddon, et al. (1997) had two video-mediated conditions, one supported mutual gaze via a video tunnel, the other did not. Interestingly, their video configuration allowing mutual gaze resulted in more turns and more overlaps than the one that did not. A subsequent analysis of gaze behaviour showed that this configuration also resulted in 56% more gazes per dialogue than the other video-mediated condition or a comparable face-to-face interaction

from a previous experiment. They describe this as "over gazing" and attribute it to the novelty of establishing mutual gaze with a TV image. The experiment described here does not have these control configurations so it is not possible to say whether over gazing occurred here. It is a possible explanation of the non-significant trend towards a larger number of turns and the significantly larger number of words spoken by the expert in the Video Tunnel Only configuration. The increased number of overlaps with the Video Tunnel Only configuration parallels the results of Doherty-Sneddon, et al. (1997) and may also be explained by the novelty of video-mediated mutual gaze.

In the experiment presented here, the other exception to the general rule that there were no significant differences between the two control configurations was in the number of Query-W Games. There were significantly more overlaps and Query-W Games in the Video Tunnel Only configuration than the audio only condition. The increase in Query-W Games is possibly due to the way the stimuli were presented in this configuration. To prevent full gaze awareness the stimuli for the expert and receiver were presented in different quadrants of the field of view. One's partner thus always appeared to be looking at the wrong portion of the screen and may have resulted in a lack of confidence in positional information in general. It should however be noted that the use of Query-W Games in the two control conditions was very infrequent, compared to the use of Check and Align Games.

In conclusion, it is difficult to judge the generality of the lack of any apparent advantage for the Video Tunnel Only configuration over Audio Only in our experiment. It is however consistent with other studies that have questioned the value of a view of the face when the work to be carried out involves mainly information transfer concerning a shared visual object and in a context where speech communication is effective (Anderson, Smallwood, MacDonald, Mullin & Fleming, 1999b; Chapanis, 1975; Gaver, et al., 1993; Veinott, et al., 1999).

*The value of full gaze awareness: the GA Display compared with the two controls*

The main focus of this study was the GA Display. The significant difference observed between the GA Display and the Video Tunnel Only video configurations is of theoretical interest as a demonstration that full gaze awareness can facilitate communication in comparison with a control condition that provides mutual gaze and facial expression. While an analysis of the information provided by full gaze awareness suggests that it should be an important conversational resource, this is to our knowledge, the first demonstration that it actually is.

A distinction can be drawn between the explicit and implicit use of gaze . The example attributed to Clark (1996) "I want <u>you</u> [gazes at A] and <u>you</u> [gazes at B]" illustrates the explicit use of gaze. This is effectively pointing with the head and eyes and could equally well have been done with a hand. The implicit use of gaze involves less conscious gaze activity on the part of the gazer. For example in our experiment the expert could monitor the visual attention of the receiver to see if it was appropriate given the instructions given.

There was some evidence in the transcripts that gaze was used explicitly for deixis.

*Pair 12*

E: ok this first one is is sort of in the top right

R: [yeah

E: [corner] and its sort of um by a little knobby [bit

R: [knobby] bit

[yeah

E: [yeah that] looks right there

R: there

E: yeah

In the above dialogue the expert (E) examines the receiver's (R) focus of attention and decides that it is correct ("yeah that looks right"). The receiver indicates explicitly that she is looking at the point in question, almost as if she is pointing with a finger, ("there"). Finally the expert concurs ("yeah"). Also one pair (pair 6) decided early on to use gaze explicitly as a strategy for solving the task. More generally however, the transcripts do not suggest that gaze is used explicitly in this way, leaving open the possibility that gaze is implicit behaviour on the part of the encoder that is monitored by the decoder, as implied by the term gaze awareness.

As well as the theoretically important demonstration that full gaze awareness can facilitate communication, the results provide support for the practical value of the GA Display. The most likely alternative in a real work context where communication has to be electronically mediated is represented by the Audio Only control condition. The GA Display resulted in participants using 949 fewer words and 55% fewer turns than in the Audio Only condition.

Alternative control conditions were considered when designing this experiment. It would have been interesting to know how the GA Display compared with a copresent control condition. This possibility was rejected because of the difficulty in choosing an appropriate copresent configuration. Two people communicating about a stimulus hung on a sheet between them would be visually equivalent but is a most unlikely work context; if they are copresent, why should they transfer the stimulus to a special display? A more realistic work context would be a printed version of the stimulus placed horizontally between two participants but this is visually very different from the GA Display.

Reducing the amount of talk required to complete the task is of practical importance in contexts where communication is difficult or may interfere with some other element of the task: delicate manual controls for example. One may also speculate that in the longer term, longer that is than a 30 minute experimental session, this saving could be reflected in practically important performance effects. Monk, McCarthy, Watts & Daly-Jones (1996) argue that participants in an experiment protect their performance of the main task given them by the experimenter. This will be reflected in differences in process measures and may also be experienced as an increase in work load. In continuous realistic work it may be much more difficult or stressful to maintain performance in this way and it is possible that a performance advantage for the

GA Display could be demonstrated in long term use. We have demonstrated that full gaze awareness can be mediated and is indeed "worth 1000 words" (949 in our experiment). Further research is needed to determine the generality of this finding with regard to longer term performance effects and in comparison with other control conditions.

## References


Anderson, A. H., Mullin, J., Katsavras, R., McEwan, R., Grattan, E., Brundell, P. & O'Malley, C. (1999a). Multimediating multiparty interactions. In A. M. Sasse & C. Johnson (Eds.), *Interact'99*, Edinburgh. (pp. 313-320). Amsterdam: IOS Press.

Anderson, A. H., O'Malley, C., Doherty-Sneddon, G., Langton, S., Newlands, A., Mullin, J., Fleming, A. M. & van der Velden, J. (1997). The impact of VMC on collaborative problem solving: an analysis of task performance, communicative process, and user satisfaction. In K. E. Finn, A. J. Sellen & S. B. Wilbur (Eds.), *Video-mediated communication.* (pp. 133-156). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Anderson, A. H., Smallwood, L., MacDonald, R., Mullin, J. & Fleming, A. (1999b). Video data and videolinks in mediated communication: what do users value? *International Journal of Human-Computer Studies, 51*, 165-187.

Anderson, J. R. (1996). Chimpanzees and capuchin monkeys: Comparative Cognition. In A. E. Russon, K. A. Bard & S. T. Parker (Eds.), *Reaching into thought: The minds of the great apes.* (pp. 23-56). Cambridge: Cambridge University Press.

Argyle, M., Lefebvre, L. M. & Cook, M. (1974). The meaning of five patterns of gaze. *European Journal of Social Psychology, 4*, 125-136.

Butterworth, G. & Jarrett, N. (1991). What minds have in common in space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology, 9*, 55-72.

Buxton, W. A. S. & Moran, T. (1990). EuroPARC's integrated interactive intermedia facility (iiif): early experience. In S. Gibbs & A. A. Verrijn-Stuart (Eds.), *Multi-user interfaces and applications.* (pp. 11-34). Amsterdam: Elsevier Science Publishers.

Chapanis, A. (1975). Interactive Human Communication. *Scientific American, 232*, 36 - 42.

Clark, H. H. (1996). *Using Language.* Cambridge: CUP.

Daly-Jones, O., Monk, A. F. & Watts, L. A. (1998). Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus. *International Journal of Human-Computer Studies, 49*, 21 - 59.

Doherty-Sneddon, G., Anderson, A., O'Malley, C., Langton, S., Garrod, S. & Bruce, V. (1997). Face-to-face and video mediated communication: a comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied, 3*, 105-125.

Duncan, S. & Niederehe, G. (1974). On signalling that it's your turn to speak. *Journal of Experimental Social Psychology, 10*, 234 - 247.

Gale, C. & Monk, A. F. (2000). Where am I looking? The accuracy of video-mediated gaze awareness. *Perception and Psychophysics, 62*, 586-595.

Gaver, W., Sellen, A., Heath, C. & Luff, P. (1993). One is not enough: multiple views in a media space. In *Proceedings of ACM INTERCHI'93 Conference on Human Factors in Computing Systems.* (pp. 335-341).

Goodwin, C. (1981). *Conversational organisation: interaction between speakers and hearers.* New York: Academic Press.

Grice, H. P. (1957). Meaning. *Philosophical Review, 66*, 377-388.

Ishii, H., Kobayashi, M. & Grudin, J. (1993). Integration of interpersonal space and shared workspace: clearboard design and experiments. *ACM Transactions on Information Systems, 11*, 349-375.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta psychologia, 26*, 22 - 63.

Kowtko, J. C., Isard, S. & Doherty-Sneddon, G. (1991). Conversational games analaysis in dialogue. In A. Lascarides (Eds.), *Tech. Rep. No. HCRC/RP-26 Publications.* University of Edinburgh.

Kraut, R. E., Miller, M. D. & Siegel, J. (1996). Collaboration in performance of physical tasks: effects on outcomes and communication. In M. S. Ackerman (Eds.), *CSCW96*, Boston, MA. (pp. 57-66). New York: ACM.

Levine, M. H. & Sutton-Smith, B. (1973). Effects of age, sex and task on visual behaviour during dyadic interaction. *Developmental Psychology, 9*, 400-405.

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

Okada, K., Maeda, F., Ichikawaa, Y. & Matsushita, Y. (1994). Multiparty video conferencing at virtual social distance: MAJIC design. In *CSCW'94*, Chapel Hill, NC. (pp. 385-393). New York: ACM Press.

Schegloff, E. A. (1991). Conversation analysis and socially shared cognition. In L. B. Resnick, J. M. Levine & S. D. Teasley (Eds.), *Perspectives on socially shared cognition.* (pp. 150 - 171). Washington DC: American Psychological Association.

Schober, M. F. & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*, 211 - 232.

Sellen, A. J. (1995). Remote conversations: The effects of mediating talk with technology. *Human-Computer Interaction, 10*, 401 - 444.

Short, J., Williams, E. & Christie, B. (1976). *The social psychology of telecommunications.* London: John Wiley and Sons.

Tang, J. C. (1991). Findings From Observational Studies of Collaborative Work. *International Journal of Man-Machine Studies, 34*, 143-160.

Veinott, E. S., Olson, J., Olson, G. M. & Fu, X. (1999). Video helps remote work: speakers who need to negotiate common ground benefit from seeing each other. In *CHI99*, Pittsburgh. (pp. 302-9). New York: ACM Press.

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

Vertegaal, R. (1999). The GAZE groupware system: mediating joint attention in multiparty communiation and collabroation. In M. G. Williams, M. W. Altom, K. Ehrlich & W. Newman (Eds.), *CHI'99*, Pittsburgh, PA. (pp. 294-301). New York: ACM Press.

Whittaker, S. (1995). Rethinking video as a technology for interpersonal communications: theory and design implications. *International Journal of Human-Computer Studies, 42,* 501-529.

Whittaker, S., Geelhoed, E. & Robinson, E. (1993). Shared Workspaces: How Do They Work and When Are They Useful? *International Journal of Man-Machine Studies, 39,* 813-842.

Figure legends

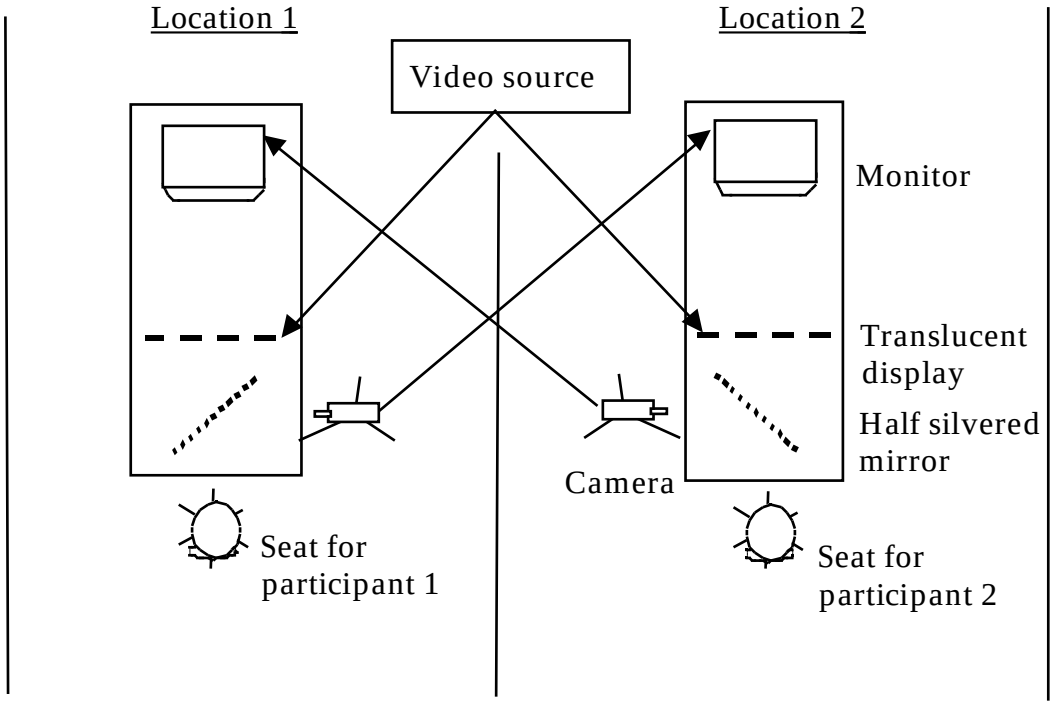


Figure 1. The GA (gaze awareness) display designed for use in the experiment. This supports mutual gaze and full gaze awareness, see text for explanation. In the experiment the translucent video displays were replaced by transparent acetate sheets.

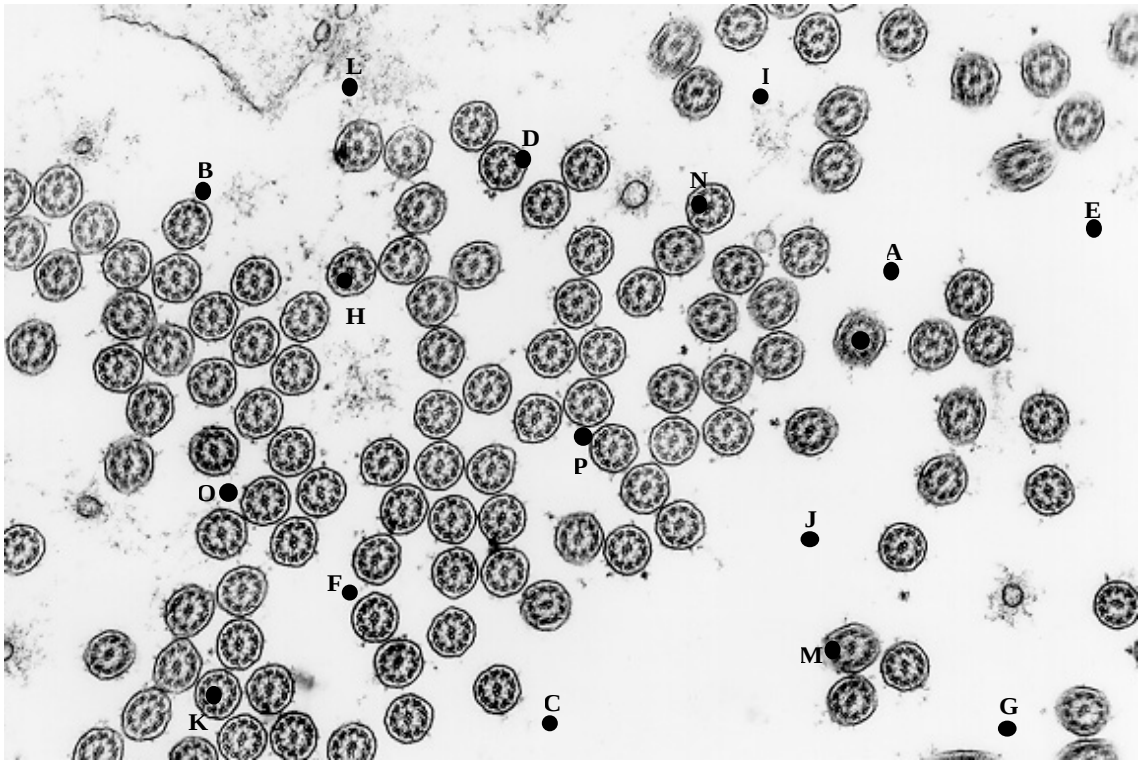

Figure 2. An example of one of the stimuli, an electron microscope slide of benzene molecules. This is the version seen by the receiver, the expert's had only two lettered points marked on it (see Method for explanation).