DISCOURSE PROCESSES, *33*(3), 257–278


# A Look Is Worth a Thousand Words: Full Gaze Awareness in Video-Mediated Conversation

Andrew F. Monk and Caroline Gale
*Department of Psychology*
*University of York*
*York, England*

*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 article sets out to demonstrate this possibility in the context of video-mediated communication. An experiment is reported in which pairs complete a communication task using a novel apparatus that supports full gaze awareness (GA) and mutual gaze (eye contact). This "GA display" was contrasted with 2 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 ½ in comparison with the 2 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 nonlinguistic channel for checking one's own and the other person's understanding of what was said.

Under the right circumstances, people can distinguish with some accuracy what someone else was currently looking at in their immediate environment. Elsewhere we have reported an experiment in which an estimator had to guess what position a gazer was looking at on a flat stimulus between them (Gale & Monk, 2000). The estimator was quite accurate; 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 did 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 atten-

Correspondence and requests for reprints should be sent to Andrew Monk, Department of Psychology, University of York, York, YO1 5DD England. E-mail: A.Monk@psych.york.ac.uk

tion as full gaze awareness (GA). 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

Full gaze awareness has received much less attention than mutual gaze in the research literature on language use. Studies of gaze following in monkeys (J. R. Anderson, 1996) and very young children (Butterworth & Jarrett, 1991) have suggested that full gaze awareness may be a pervasive and primitive cognitive mechanism that precedes verbal language. Clark (1996) used full gaze awareness as an example of how adult language uses many signals in addition to words and sentences. He described how head-and-eye movements can be used to indicate references in conversation, for instance, when a speaker says, "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 (e.g., Grice, 1957; Schegloff, 1991). According to this account, conversation is possible only 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 behavior for problems. As a hearer, one has an obligation to signal when one cannot understand the speaker sufficiently for these purposes. The possibility we wish to raise in this article 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 was 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.

### The Problem: Mediating Full Gaze Awareness and Mutual Gaze

The applied context of this work is the design of multimedia 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) pointed out how a shared visual artifact can be used as a conversational resource in design meetings to store information, express ideas, and mediate communication. Whittaker, Geelhoed, and Robinson (1993) discussed 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 favor of shared data (A. H. Anderson, Mullin, et al., 1999; Gaver, Sellen, Heath, & Luff, 1993; Kraut, Miller, & Siegel, 1996). This need for shared visual artifacts 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 the other person 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 in favor of a high definition full movement image of facial detail are that this is necessary to discriminate subtle facial expressions and that it supports lip reading. However, the importance of subtle nuances in facial expression may be overstated in most task contexts (Daly-Jones, Monk, & Watts, 1998; Whittaker, 1995). Also, if a multimedia communication link makes lip reading necessary, it suggests that there was 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 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 is mounted on top of the monitor displaying an image of the person one was talking to. Mutual gaze is difficult to achieve with this configuration. If you look 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 you are looking them in the eyes, but then you 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."

## Previous Solutions: Inventions That Mediate Full Gaze Awareness

Various ingenious inventions have been suggested to mediate full gaze awareness using multimedia 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 was looking was projected onto the shared electronic workspace. Velichkovsky's invention can thus be thought of as artificial full gaze awareness, although it did not support mutual gaze.

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 Velichkovsky (1995) and Vertegaal's inventions 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 was particularly impressive, as it solved the problem of gaze awareness in multiparty conversations in which gaze awareness may be particularly important in regulating turn taking (see Sellen, 1995; Okada, Maeda, Ichikawaa, & Matsushita 1994). However, the mutual gaze may have been rather artificial as it was mutual gaze with an avatar, not with a video image of the other participant.

## Previous Solutions: Inventions That Mediate Mutual Gaze

The video tunnel was invented to circumvent the problem of providing true mutual gaze in a two party conversation (Buxton & Moran, 1990). This used half-silvered mirrors to put the cameras in the same virtual positions as the monitors. Each participant saw the image of the other through a half silvered mirror. Each camera saw a face reflected in it. By adjusting the position of the camera and face, it was 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 had to be restricted. In the original Xerox design, this was achieved by putting a long cowl on the front, hence the description video tunnel. Because of this limited view of the other participant and their immediate environment, the video tunnel did not support full gaze awareness.

Ishii, Kobayashi, and Grudin's (1993) Clearboard 2 design provided mutual gaze in a similar manner to the video tunnel. Here, the image of one's conversational partner was projected onto the underside of a transparent drawing surface. The drawing surface was a half-silvered mirror, the "board" of the title. The camera was placed so as to look down on the board, thus capturing a reflection of the participant in the board. The board also contained a polarizing sheet that prevented

the multiple images that would arise if the camera picked up the image projected from the other camera.

Clearboard 2 also had 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 sat in the same room facing each other across a table. In between them was a transparent sheet of glass with the object they were conversing about written on it. If they both looked at the drawing, they were looking at the same object, although of course one of them was looking at the wrong side and would not have been able to read letters. If they looked through the glass they could see each other's facial expressions and were also able to achieve mutual gaze as they normally would. If one person looked at the drawing on the glass then the other could look through the glass and judge what part of the drawing the first person was looking at; that is, they could also achieve full gaze awareness. To provide the features of Clearboard 0 in a mediated context in which the conversants were in different locations, Clearboard 2 simply displayed the drawing on the board at the same time as the image of the other person.

Some of our informal experiments question how accurate this full gaze awareness could have been. The problem is that, unlike Clearboard 0, Clearboard 2 provided 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 was, the angle of head-and-eye rotation of the gazer, 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 was 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 relation 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 was difficult to see how such an arrangement could provide anything better than partial gaze awareness, that is, whether the other person was looking up, down, left, or right.

## A New Solution: The GA Display

None of the inventions reviewed previously were able to mediate full gaze awareness and mutual gaze. Vertegaal's (1999) GAZE system comes closest, but in a very artificial way, using avatars and arbitrary signals. Our display was based on Clearboard 2 but physically separated the objects being looked at from the image of the other participant. Figure 1 shows how this "GA display" worked. Each participant viewed the image of the other participant through a half-silvered mirror in

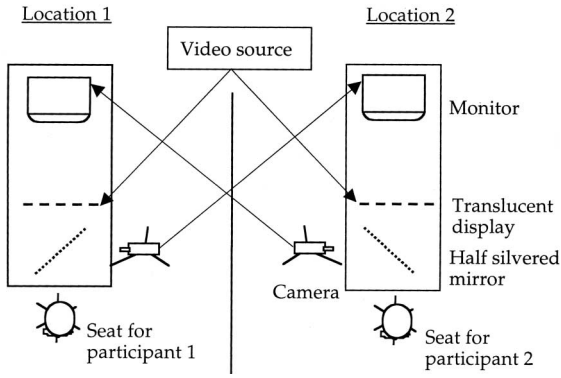


FIGURE 1 The gaze awareness display designed for use in the experiment. This supports mutual gaze and full gaze awareness. In the experiment the translucent video displays were replaced by transparent acetate sheets.

a conventional video tunnel arrangement. The camera was mounted to the side of the participant it viewed such that the mirror served to right–left reverse the image. The focal length of the camera was adjusted so that the image of the participant's face was actual size. A translucent display was positioned half way between the participant and the image of the other participant. Although both participants were effectively looking at the same side of the objects in this display, that is, they could both read writing on it, there was still full gaze awareness. This was because the mirror left–right transformed the image of the other person apparently behind the translucent display. When one partner looked at an object on their right hand side, that is, the other person's left, they were actually looking to the right of their display, but the mirror transformed the image appropriately to give the other partner full gaze awareness.

Our results show that this arrangement allowed good gaze estimation accuracy. The design provided both mutual gaze and full gaze awareness for any stimulus that can be viewed with an image of the other participant behind it. Although 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 problem of distinguishing the image of the face from the image of the object was reduced because the depth of field of the eye was such that only the face or the object could be fully in focus at any one time.

## This Experiment

The first objective of the experiment was to demonstrate that full gaze awareness can, at least under some circumstances, be another conversational resource in the cooperative activity of conversation. In particular, full gaze awareness was pre-

dicted to reduce the need of a speaker to use verbal language to check her own understanding or the understanding of her conversational partner. The second objective was 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. The expert's 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 member, the receiver, had to identify 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 maximize the advantages of full gaze awareness. Using the GA display the expert could monitor the receiver's current locus of visual attention and so could judge whether the instructions given are being understood. Similarly, the receiver could use where the expert was looking as a parallel source of information with which to check his or her understanding regarding the target point.

The GA display was compared with two video-mediated control conditions: (a) video tunnel only and (c) audio only. The first control condition was achieved by reducing the size of the picture 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 whereas the receiver's drawing was printed onto the top left-hand quarter. This very considerably reduced full gaze awareness yet all the information given by the video tunnel (mutual gaze and facial expressions) was still available. The second control condition was achieved by switching off the image of the other person's face. In this control condition there was no view of the other participant, that is, no mutual gaze, no full gaze awareness, and no facial expressions.

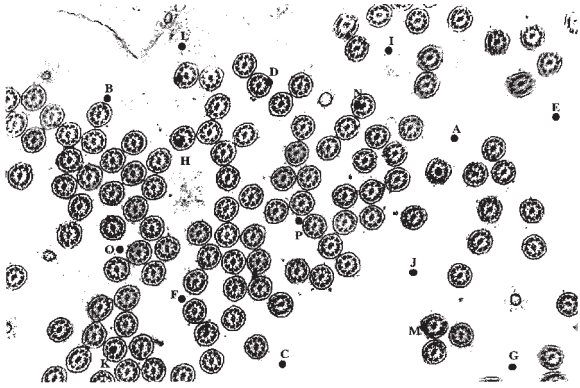


FIGURE 2 An example of one of the stimuli, an electron microscope slide of benzene molecules. This is the version seen by the receiver; the expert's had only two lettered points marked on it.

The GA display was predicted to result in more efficient conversations, that is, fewer turns and fewer words to complete the task than in either of the control conditions where full gaze awareness was not possible. More specifically, people were expected to do less verbal checking of their own or their partner's understanding in the GA display condition. This latter prediction was tested using conversational games analysis (Kowtko, Isard, & Doherty-Sneddon, 1991). Comparison of the GA display condition with the video tunnel only condition was most relevant to the first objectives of demonstrating a role for mutual gaze, because here the other video resources (facial expressions and mutual gaze) were still available. The audio only control condition provided a reference point or control condition that is commonly used in experiments on video-mediated communication.

Finally, there was a second independent variable, audio quality: 81.5 dBA noise was mixed into the sound channels for half the trials. It was predicted that the visual information provided by the GA display might be most advantageous when the verbal channel was impeded in this way. However, 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

### Participants and Design

Forty-eight participants were used in 24 pairs and assigned at random to one of the three 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. Pairs performed two trials with each of 10 pictures. Half the pairs in each condition conducted the trials with the poor audio quality (noise) for the first 5 pictures and the good quality audio (no noise) for the second 5 pictures. The other half of the pairs experienced the audio quality conditions in the opposite order. The pictures were presented in a different random order for each pair.

For all the statistical analyses presented in the next section, the sampling unit was the participant pair. Audio-visual configuration was a between-pairs variable, and audio quality was a within-pairs variable. In addition, there was a between-pairs variable, order (of audio quality condition).

### Materials

The stimuli used in the communication task were pictures that had elements that were difficult to describe. In this way, it was similar to the tangram task used by

Schober and Clark (1989) and other authors. In such tasks, communication eventually becomes very efficient as partners develop codes and systems for describing the stimuli. The prediction was that full gaze awareness would be most useful when verbal communication was least efficient, that is, when the pairs were 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 10 pictures were 2 circuit diagrams, 2 architectural blueprints of houses, 2 diagrammatic representations of the rat's brain, 2 extracts from sheet music printed backwards, and 2 electron microscope images featuring benzene rings, of which one is depicted in Figure 2.

Around 15 arbitrary points (plus or minus 2) were marked and labeled with letters on each of the receiver's pictures. A previous experiment (Gale, 1998) had used a single circuit board stimulus in which both partners could see all the labeled 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 2 points on it, the 2 points that had to be described to the receiver in that trial. This prevented the points from 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.

## Apparatus

Two video tunnels were constructed (see Figure 1). Each participant sat 150 cm from a monitor, the midpoint of which was at eye level and marked with a colored 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 the monitor. A colored dot marked the midpoint of each stimulus; this was again at eye level. Participants could thus line up the colored 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. 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 labeled points added to it. The other two copies of each picture were the expert's. Each of these had 2 points marked; these were the 2 points that the expert had to describe in each of the two trials with that picture. Thus, in the course of the two trials 4 points in all were described.

For pairs in the video tunnel only configuration, the same 10 pictures described previously were used but reduced in dimensions by half, that is, each stimulus was approximately 10.5 cm × 15 cm instead of 21 cm × 30 cm. The stimuli were identical to those described previously 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 whereas the receiver's drawing was printed onto the top left-hand quarter. Although the objects to be described were smaller, they were still readily visible. Note also that the video tunnel only condition supported partial gaze awareness, that is, information about the general direction of gaze (left vs. right, up vs. down).

## 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 the task was to describe to the receiver the location of a particular pair of points on various pictures. In addition, the expert was asked to give feedback once the receiver had made a guess. It was emphasized to both participants that they could interrupt each other at any point and that the receiver should do so as soon as he or she 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. The chair height was adjusted to give a constant eye height, and they performed one practice trial before starting the experiment proper. Each pair completed two trials with each picture. In each trial they described two points. 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 as a manipulation check. 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 Check: Gaze Accuracy

The manipulation check showed that the GA display did indeed provide full gaze awareness and the video tunnel only configuration did not. The GA display resulted in greater than 90% correct responses (*M* correct out of 20 = 18.6, 18.3, and 18.4, for the pairs originally in the GA display, video tunnel only, and audio only conditions, respectively). In contrast, with the video tunnel only configuration, less than 10% of the points were guessed correctly (*M* correct out of 20 = 1.9).

### Task Performance

Table 1 gives the mean number of trials correct in the main part of the experiment. Performance was 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 1 also gives the mean time for each audio quality condition for the pairs in each video configuration. An analysis of variance showed no significant main effect of video configura-

TABLE 1
Means and Standard Deviations for Number of Trials Correct and Time to Complete the 10 Trials

| | *GA Display* | | *Video Tunnel Only* | | *Audio Only* | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Trials correct out of 10 | | | | | | |
| High quality audio | 9.50 | 0.76 | 9.50 | 0.53 | 9.88 | 0.35 |
| Poor quality audio | 9.75 | 0.46 | 9.38 | 0.74 | 9.50 | 0.76 |
| Time to complete in minutes | | | | | | |
| High quality audio | 10.71 | 2.87 | 12.94 | 2.38 | 12.70 | 2.68 |
| Poor quality audio | 11.45 | 3.74 | 13.73 | 2.26 | 12.94 | 2.45 |

*Note.* GA = gaze awareness.

tion or order, $F(2, 18) = 1.87$, *ns*; $F(1, 18) = 2.59$, *ns*. There was a significant effect of audio quality condition, $F(1, 18) = 4.91$, $p < .05$. There were no significant two- or three-way interactions.

It was not expected that measures of performance would be sufficiently sensitive to result in significant effects for the between-pairs variable video configuration, so our hypotheses are framed in terms of process variables extracted from transcripts. Previous work has shown that performance data are rarely sensitive to manipulations of video mediation (Doherty-Sneddon et al., 1997; Monk, McCarthy, Watts & Daly-Jones, 1996; Whittaker, 1995). Our experimental instructions emphasized accuracy, and the expectation was that the receiver would guess correctly on most trials. Minimal time pressure was applied, and therefore, there was no strong reason to expect effects on time to completion.

The significantly longer times to completion in the poor audio quality condition suggests 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 other significant main effects of audio quality. More important, none of the dependent variables demonstrated the predicted interaction between audio quality and video configuration: The effect of video configuration was the same whether or not the noise was present. One simple interpretation of these results was that the audio quality 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 nonwords, back-channel responses (e.g., “mhm”), 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 s. 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 nonwords or half-words), and pauses. A turn was defined as any vocal utterance that was (a) not a back-channel response, (b) 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 (c) not some nonverbal vocalization such as a grunt or a sigh.

Each of these measures of process was subject to the same three-way analysis of variance as time to completion. None of these analyses demonstrated a significant main effect of audio quality or any significant interaction with audio quality; therefore, the data were collapsed across audio quality condition and order in Table 2. However, there were dramatic differences between the GA display configuration and the two control conditions for turns and words spoken by the expert, with about twice as much talk in both of the control conditions, $F(2, 18) = 14.92$, $p < .001$; $F(2, 18) = 4.53$, $p < .05$; $F(2, 18) = 3.95$, $p < .05$, for turns, words spoken by the expert, and words spoken by the receiver, respectively. Post hoc tests using Tukey's HSD showed that there were significant differences between all three means for words spoken by the expert (HSD = 594 words) but only between GA display and each of the control conditions for turns (HSD = 75.4 turns) and words spoken by the receiver (HSD = 123 words).

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 article: There was simply less talk when full gaze awareness was available to the participants. Interestingly there was 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.

Similar effects were seen in the number of overlaps and pauses; however this is what one might expect from the fact there was less talk in the GA display condition. Table 2 gives overlaps and pauses per 100 turns; there were no significant effects of video configuration.

## 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

TABLE 2
Means and Standard Deviations for Number of Turns, Words Spoken, Overlaps per 100 Turns, and Pauses per 100 Turns

| | *GA Display* | | *Video Tunnel Only* | | *Audio Only* | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Turns | 181.88 | 69.79 | 405.63 | 80.12 | 332.38 | 120.21 |
| Words (expert) | 1334.00 | 488.69 | 2662.13 | 496.34 | 2111.00 | 882.90 |
| Words (receiver) | 362.13 | 145.95 | 521.50 | 121.01 | 534.38 | 170.40 |
| Overlaps/100 turns | 9.47 | 4.92 | 13.25 | 5.14 | 8.94 | 2.84 |
| Pauses/100 turns (expert) | 40.23 | 18.85 | 28.73 | 17.13 | 34.41 | 10.56 |
| Pauses/100 turns (receiver) | 7.00 | 3.65 | 8.42 | 3.27 | 8.89 | 2.15 |

*Note.* GA = gaze awareness.

for trouble that may need to be repaired. Our hypothesis was 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. This was operationalized here as a reduction in check and align games coded in a conversational games analysis.

Conversational games analysis is a technique for dialogue analysis developed concurrently at the Universities of Edinburgh, Glasgow, and Nottingham that categorizes 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 (e.g., see A. H. 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 3 lists the different games that may be coded.

In the experimental task used in this experiment, each trial consisted of an instruct game as the expert was communicating a direct request to the receiver to name two points on the picture. The successful completion of this instruct game was evidenced by a query yes/no game initiated by the receiver and embedded within this instruct game. An example from the transcripts is given next. Each game is numbered and bracketed by its start and end, and each move is labeled after the utterance containing it. Thus, the reply-yes move from the expert ends both the embedded query yes/no game and the instruct game. An asterisk indicates a pause of more than 0.5 s.

TABLE 3
Six Types of Games Found Necessary and Sufficient to Code Transcripts in This Experiment

**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.

*Note.* From *Video-Mediated Communication* (p. 138), by K. E. Finn, A. J. Sellen, and S. B. Wilbur, 1997, Mahwah, NJ: Lawrence Erlbaum Associates, Inc. Adapted with permission.

**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: instruct**

R: okay

**Move: acknowledge**

**2 Start Game Query yes/no (embedded)**

R: e and k

**Move: query-yes/no**

E: yes you're right

**Move: reply-yes**

**End Game 2**

**End Game 1**

This is the simplest form the dialogue could take to complete the trial. Most trials contained other embedded games, most notably games in which one partner checked their own or the other person's understanding. Checking one's own understanding was 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 next. The receiver interrupted the instruct move that was to communicate the position of the second point to check that she had understood what was 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: 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-yes**

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 was 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 next.

**11 Start Game align (embedded)**

E: you've got it

**Move: align**

R: I think so yeah

**Move: reply-yes**

E: yeah

**Move: align**

**End Game 11**

The number of games per session for check and align games, where a session is taken as the 10 trials in each audio quality condition, was subject to the same analysis of variance as time to completion. Audio quality had no significant main effect on the number of check or align games. 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 fewer of these games in the second half of the experiment. In neither case was there a three-way interaction; thus the means presented in Table 4 collapse across audio quality condition and order.

The advantage for the GA display configuration was even more dramatic than that observed with the process measures. There were many fewer check and align

games with the GA display configuration than with either of the two control conditions, $F(2, 21) = 9.10$, $p < .01$; $F(2, 21) = 8.12$, $p < .01$, respectively. Tukey's HSD test shows that the significant differences were between the GA display configuration and the control configurations. For the Check games this comparison was significant for both control groups (HSD = 11.7 games). For the align games the comparison between GA display and Video tunnel only did not quite reach significance (HSD = 14.1). 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 nonverbally.

## 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 lacked only full gaze awareness, whereas the audio only configuration lacked all visual information about the behavior 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 not detectable in this experiment.

TABLE 4
Means and Standard Deviations for Number of Check and Align Games per Session

| | *GA Display* | | *Video Tunnel Only* | | *Audio Only* | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Check | 6.12 | 4.17 | 27.25 | 11.4 | 17.88 | 9.35 |
| Align | 6.12 | 6.19 | 19 | 8.87 | 29.62 | 16.3 |

*Note.* GA = gaze awareness.

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

Although 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 between these two control conditions. 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 less 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 were a number of differences between Doherty-Sneddon et al.'s experiment and ours, particularly in the tasks used. The map task depended on the two participants not being able to see each others' maps, as they differed in detail. Our task depended on being able to see identical stimuli. Also, in Doherty-Sneddon et al.'s experiment, video configuration was a within pairs variable. This led 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, and 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 behavior 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 "overgazing" (i.e., gazing more than is natural) and attribute it to the novelty of establishing mutual gaze with a TV image. The experiment described here did not have these control configurations, therefore it is not possible to say whether overgazing occurred here. Overgazing is a possible explanation of the nonsignificant 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 paralleled the results of Doherty-Sneddon et al. and may also be explained by the novelty of video-mediated mutual gaze.

In sum, 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 in which speech communication is effective (A. H. Anderson, Smallwood, et al., 1999; Chapanis, 1975; Gaver et al., 1993; Veinott, Olson, Olson, & Fu, 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 condition 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. Although 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 experimental demonstration that it actually is.

A distinction can be drawn between the explicit and implicit use of gaze. Clark's (1996) example, "I want *you* [gazes at A] and *you* [gazes at B]," illustrated 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 uttered.

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 previous dialogue the expert (E) examined the receiver's (R) focus of attention and decided that it was correct ("yeah that looks right"). The receiver indicated explicitly that she was looking at the point in question, almost as if she was pointing with a finger, ("there"). Finally the expert concurred ("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 was used explicitly in this way, leaving open the possibility that gaze is implicit behavior on the part of the encoder that is monitored by the decoder, as implied by the term *gaze awareness*.

As well as demonstrating 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 in which 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 in which 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-min experimental session, this saving could be reflected in practically important performance effects. Monk et al. (1996) argued that participants in an experiment protect performance of the main task given them by the experimenter. This would be reflected in differences in process measures and might also be experienced as an increase in workload. 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 1,000 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.

## ACKNOWLEDGMENTS

Caroline Gale is now at BTexaCT, Adastral Park, Martlesham Heath, Ipswich, England.

During this work, Andrew F. Monk was supported by the England Economic and Social Research Council through their Cognitive Engineering Programme and Caroline Gale by a Research Studentship from the England Engineering and Phys-

ical Science Research Council. We thank Leon Watts for his contribution setting up these experiments, and we also thank Gordon Baxter and anonymous reviewers for their comments on drafts of this article.

## REFERENCES

Anderson, A. H., Mullin, J., Katsavras, R., McEwan, R., Grattan, E., Brundell, P., & O'Malley, C. (1999). Multimediating multiparty interactions. In A. M. Sasse & C. Johnson (Eds.), *Interact'99* (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, NJ: Lawrence Erlbaum Associates, Inc.

Anderson, A. H., Smallwood, L., MacDonald, R., Mullin, J., & Fleming, A. (1999). 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, England: 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.

Chapanis, A. (1975). Interactive human communication. *Scientific American, 232,* 36–42.

Clark, H. H. (1996). *Using language.* Cambridge, England: Cambridge University Press.

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. (1998). *The effects of gaze awareness on communication in video-mediated spatial instruction tasks.* Unpublished master's thesis, The University of York.

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.

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 analysis in dialogue. In A. Lascarides (Ed.), *Tech. Rep. No. HCRC/RP-26 Publications.* University of Edinburgh, Edinburgh, Scotland.

Kraut, R. E., Miller, M. D., & Siegel, J. (1996). Collaboration in performance of physical tasks: Effects on outcomes and communication. In M. S. Ackerman (Ed.), *CSCW96* (pp. 57–66). New York: Association of Computing Machinery.

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 (Ed.), *CSCW Requirements and Evaluation* (pp. 125–139). Berlin, Germany: Springer Verlag.

Okada, K., Maeda, F., Ichikawaa, Y., & Matsushita, Y. (1994). Multiparty video conferencing at virtual social distance: MAJIC design. In *CSCW'94* (pp. 385–393). New York: Association of Computing Machinery.

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: Wiley.

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 M. G. Williams, M. W. Altom, K. Ehrlich, & W. Newman (Eds.), *CHI99* (pp. 302–9). New York: Association of Computing Machinery.

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* (pp. 294–301). New York: Association of Computing Machinery.

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.