Effects of Gaze on Multiparty Mediated Communication

Roel Vertegaal
Computing and Information Science
Queen's University
Canada
E-mail: roel@acm.org

Gerrit van der Veer
Computer Science Department
Vrije Universiteit Amsterdam
The Netherlands
E-mail: gerrit@cs.vu.nl

Harro Vons
Usability Consultancy
Baan Apps
The Netherlands
E-mail: hvons@baan.nl

ABSTRACT

We evaluated effects of gaze directional and other non-verbal visual cues on multiparty mediated communication. Groups of three participants (two actors, one subject) solved language puzzles in three audiovisual communication conditions. Each condition presented a different selection of images of the actors to subjects: (1) frontal motion video with 14% gaze; (2) motion video with head orientation and 7% gaze; (3) still images with head orientation and 32% gaze. Presence of head orientation cues caused subjects to use twice as many deictic references to persons. We found a linear relationship between the amount of actor gaze perceived by subjects and the number of speaking turns taken by subjects. Lack of gaze can decrease turn-taking efficiency of multiparty mediated systems by 25%. This is because gaze conveys whether one is being addressed or expected to speak, and is used to regulate social intimacy. Support for gaze directional cues in multiparty mediated systems is recommended.

KEYWORDS CSCW, videoconferencing, gaze direction.

INTRODUCTION

Humans exhibit great sensitivity to the look (or *gaze*) of others [2]. Gaze at their eyes reveals that a person is looking at them. From a distance of about 1 m, people can discriminate gaze at their eyes by someone facing them with an accuracy of approximately .6 degrees [6]. Head orientation reveals a person is looking at others. From 1.5 m distance and at right angles to two interactors, humans can discriminate one person looking at the eyes of the other in 60% of cases, simply by judging the angle of head orientation [28]. However, most video-mediated communication systems are not very good at preserving gaze directional cues [21]. This is because each person has only one camera (allowing a single frontal picture), and because that camera is typically placed well above the eyes of the other person on the screen. Due to this parallax, eye gaze appears lowered. Isaacs & Tang [9] and O’Connaill et al. [14] observed that single-camera video mediated systems may cause problems in mediating multiparty communication. They noticed difficulties in floor control, and in referring to other participants. Our assumption was that these problems were directly caused by the lack of information about the gaze direction of the participants. Gaze directional cues code who is talking or listening to whom with great accuracy [25], and we expected the lack of such information to have a great effect on the management of group conversations. However, the isolated effect of gaze directional cues on multiparty conversation was never demonstrated empirically. We therefore conducted an experiment in which we gauged the effect of such cues on a variety of dependent variables in a triadic video-mediated collaborative setting. To estimate their relative importance, we compared effects to those of other visual cues typically conveyed in video mediated systems. We will first discuss our independent variables, and how they were used to constitute experimental conditions. For each dependent variable, we will then discuss why it was measured, how this was done, and predictions toward treatment effects.

In Proceedings of Graphics Interface 2000, Montreal, Canada.

INDEPENDENT VARIABLES

We tried to isolate the effect on multiparty communication of three independent variables: (a) the presence of head orientation information; (b) the amount of gaze at the eyes conveyed; (c) the presence of other non-verbal visual cues such as facial expressions and lip movements, as conveyed by motion video. We used levels of variable (a) and (c) to constitute the following three conditions:

1) A condition in which moving upper-torso visual cues were presented, but no head orientation (hereafter referred to as *motion video-only*).
2) A condition in which moving upper-torso visual cues were presented, including head orientation (hereafter referred to as *motion video with gaze direction*).
3) A condition in which *no* moving upper-torso visual cues were presented, except for head orientation (hereafter referred to as *still images with gaze direction*).

As Sellen [18] showed, the use of different mediated systems to create these conditions is not possible without introducing other, potentially confounding, differences. Instead, we controlled our factors towards subjects using the same system in all conditions, by using actors as their conversational partners. These actors would alter their behavior towards subjects according to experimental conditions. Using triads of one replaceable subject and two reusable actors, we thus constituted the simplest form of multiparty communication, keeping the number of subjects and actors required to an absolute minimum. However, control over variable (b), the amount of gaze at the eyes of subjects, proved more difficult. Our experiment was aimed at evaluating the effect of human cues, rather than the technology used to convey them. As said, however, video mediation does not allow gaze at the eyes to be conveyed due to the parallax between camera and screen. Rosenthal tried to solve this problem [16]. By placing a half-silvered mirror at a 45° angle between camera and screen the camera could be virtually positioned behind the eyes of the person on the screen [1]. The great drawback of this *video tunnel* technology is that subjects would have to sit perfectly still

– their heads in a tunnel construction – to keep their eyes exactly aligned with the lens of an actor's camera [21]. This, in turn, would impair individual gaze at their eyes by the other actor, block head orientation cues, and restrict the natural behavior of subjects. To ensure subjects were able to perceive gaze at their eyes we therefore had to take a different approach, borrowed from TV presenters. We instructed the actors to look into the camera as much as possible when looking at their video monitors, thus simulating gaze at the eyes of subjects. This did mean the amount of gaze was allowed to potentially vary between conditions. We controlled for this confounding influence retroactively by measuring the amount of gaze at the eyes received by subjects, using this as a covariate in our statistical tests. Predictions with regard to most dependent variables were therefore difficult to make, requiring post-hoc testing in most cases.

DEPENDENT VARIABLES AND PREDICTIONS

We measured treatment effects on three dependent variables: task performance; the number of deictic references to persons; and turn frequency.

Task Performance

As Monk et al. [13] demonstrate, results obtained in comparing different mediated settings may depend very much on the experimental task used. Tasks that are highly personal and/or involve conflict are much more sensitive to differences in mediation than, e.g., problem-solving tasks. Thus, they are more likely to affect dependent variables other than task performance itself. We therefore devised a collaborative problem-solving task based on language puzzles. For each problem, each participant would obtain one of three pieces of information required to solve that problem. Participants would need to put these pieces in the correct order to score a point. By verbal communication of pieces and permutations of pieces, participants would collaborate to perform the task. Performance measure was the number of correct permutations given per session.

Deictic Verbal References

In their usability studies on video-mediated vs. face-to-face communication, Isaac and Tang observed many instances in face-to-face interaction when people used their eye gaze to indicate whom they were addressing [9]. However, when using a video-mediated system, participants would often use each other's names to indicate whom they were addressing. In general, the use deictic references to persons may be problematic when visuo-spatial cues are not conveyed. For example, if "You can try" is a direct response to something the addressed person just said, the meaning of the word "*you*" is easily disambiguated by knowledge about the identity of the previous speaker. If "You can try" is used imperatively, extra information is needed to ascertain whom is being addressed. This can be provided by head pointing. We believed it likely the availability of head orientation cues would thus affect the use of deictic referencing [10]. We measured the ability to use deixis towards persons by counting singular deictic use of second-person pronouns (i.e., the *you* in "Do *you* think so?"). As we did not expect a confounding influence of our covariate, we planned the evaluation of the following hypothesis:

Predictions Regarding Deictic Verbal References "The presence of head orientation cues causes the number of personal deictic verbal references used to rise significantly."

Speaker Switching and Turn Frequency

Isaacs and Tang [9] also observed that during video conferencing, people would control the turn-taking process explicitly by requesting others to take the next turn. In face-to-face interaction, however, they saw many instances where people used their eye gaze to indicate whom they were addressing and to suggest a next speaker. Kendon [12] suggested gaze directional cues play an important role in keeping the floor, taking and avoiding the floor, and suggesting who should speak next. As such, Short et al. [20] attributed problems in turn-taking behavior with mediated systems to a lack of gaze directional cues. We therefore decided to measure the number of turns taken by participants. Like Sellen [19], we did this by automated analysis of participants' speech patterns. There is little comparable evidence on which to base predictions regarding the effect of gaze directional cues on multiparty speaker switching. Firstly, there is only one study, by Sellen [18], in which gaze directional cues were part of experimental treatment. Sellen failed to find significant differences in the number of turns between several multiparty conversational contexts: face-to-face, video-mediated with gaze direction, video-mediated without gaze direction, and audio-only communication. Secondly, most studies, particularly the early ones, were based on dyadic (two-person) communication. Finally, most studies, including Sellen's, compared communication settings that differed on too many variables at once. The most confirmed result from dyadic studies is a significant increase in the number of turns in face-to-face conditions, as compared with audio-only conditions [4, 17]. These results may well be explained by a lack of gaze directional cues yielding a worse synchronization of turn-taking in audio-only conditions [12]. Most studies suggest that with regard to turn-taking, adding motion video to speech communication has little effect (see Sellen [18] for an overview).

METHOD

We used an independent samples design for our experiment, comparing performance between three matched groups of subjects, each group treated with one of the three conditions. We treated this design as single-factor, using post-hoc testing for most dependent variables.

Conditions

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

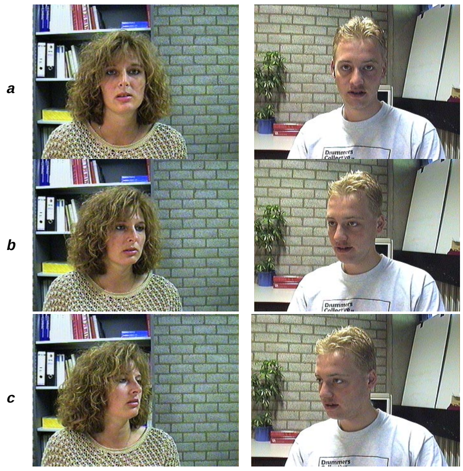

Figure 1. *Three different directions of actor gaze as experienced by the subjects: a) facing the subject; b) looking at computer screen; and c) looking at other actor.*

1) *Motion video-only*. In this condition, the subjects saw a full-motion video image of the actors, with the actors always facing the subject (Figure 1a).
2) *Motion video with gaze direction.* In this condition, the subjects saw a full-motion video image of the actors. Actors were allowed to turn their heads in different directions, indicating whom or what they looked at: the subject (Figure 1a), their computer screen (Figure 1b), or the other actor (Figure 1c). As actors were in the same room, it would have been possible to achieve eye contact between them in this condition. To avoid this potentially confounding effect, when looking at each other, they looked at a common reference point instead.
3) *Still images with gaze direction.* At any moment in time, actors would manually select one of three still images for display to the subject: actor looking at subject (Figure 1a), actor looking at computer screen (Figure 1b), or actor looking at other actor (Figure 1c). Actors were instructed to base their selection on whom or what they would actually be looking at. This looking behavior essentially replicated that of condition 2. Note that the frontal picture was taken with the actors looking straight into the camera lens.

Experimental Subjects and Actors

Our experimental subjects were paid volunteers, mostly university students from a variety of technical and social disciplines. Prior to the experiment, we tested all subjects on eyesight and a number of relevant matching variables: Dutch language competence (using a pen-and-paper aptitude test [8]); age; sex; and field of study. We allocated each subject to a treatment group in a way that matched groups on these variables. The 56 subjects used for further analysis were assigned to treatment groups as follows:

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

Subjects believed the actors were subjects also. None of the subjects in this subset knew or had any suspicion regarding the actors. None had any previous experience with video-mediated communication. Subjects believed we were interested in how people cooperate via the Internet, and were only informed of the true purpose of the experiment after treatment. We used one female and one male actor, seated in a separate room from the experimental subject. The difference in sex between the actors may have aided identification of voices in the still images with gaze direction condition. Both actors were about the same age as the subjects.

Task

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

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

For the subject, each sentence fragment appeared on a computer screen. The actors pretended this was the case for them also, having their fragments listed on paper instead. To prevent a practice effect, this paper listed all correct answers to each puzzle. It prescribed which correct solutions they were allowed to give away, and when to give incorrect solutions. This was done to minimize the influence of actors on task performance while keeping their act credible towards the subject. In order to ensure an exchange of information between the subject and each actor:

1) Nobody could see the sentence fragment of the other participants.
2) Each fragment remained on the subject's screen for only 10 seconds.
3) Each participant had a specific role. The subject's role was to submit each solution they collectively agreed on to be correct. Actor 1 would pretend to enter this solution for verification by computer, while Actor 2 would report its correctness, pretending this was indicated on her computer screen.

When all correct permutations were given, a computer would provide a new sentence fragment on the subject's computer screen, generating an audio signal to inform the actors. The number of correct permutations generated per 15 minute session was used as a measure of task performance. Correct permutations that were given more than once counted only once, and uncompleted language puzzles were discarded.

Instructions and Session Procedure

Prior to the experiment, actors were instructed with regard to their behavior in the different conditions, which they practiced in several training sessions. Actors memorized all answers to all problems solved in the experimental task prior to the experiment. They were not informed until after the experiment of the purpose of the experiment or reasons behind the experimental treatments. Actors were instructed to behave as if they were subjects, with a similar system setup. However, actors were told to allow the actual subject to take the initiative. This resulted in a situation in which much of the interaction was between the subject and one of the actors, rather than between actors only. For each subject, the session was structured in the following way. After introducing the subject to the system, the session would start with the participants seeing and hearing each other. After introducing themselves, the experimenter explained the role of each participant using a simple practice game. After exactly one minute, the experimenter interrupted the game to explain the rules of the actual task. The session proceeded with the first puzzle, ending after 15 minutes. After each session, subjects filled in a questionnaire and were debriefed by the host.

MATERIALS

All equipment was set up in a way that minimized differences between conditions to treatment variables only. All video and audio equipment was analog, with no discernable lag. All video and audio signals were recorded in sync on video tape using a video splitter device and two audio tracks.

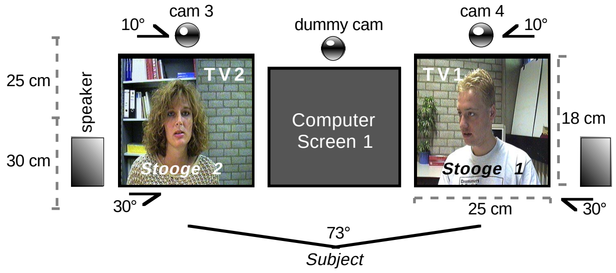

Figure 2. *The video-mediated system used by the subjects.*

Subject Configuration

Figure 2 shows the setup as experienced by subjects. The subject was seated in front of two video monitors, with the right monitor (TV1) displaying the image of Actor 1 and the left monitor (TV2) the image of Actor 2. Between these video monitors, a computer monitor (Screen 1) was placed, used to display the subject's sentence fragment. The average distance from the head of the subject to each monitor was approximately 60 cm. From the subject's point of view, the angle between the center of the left monitor image and the center of the right monitor image was approximately 73°.

Actor Configuration

The equipment for each actor was about half the subject configuration. Actors each had only one video monitor (TV3 and TV4) on which they always saw live video images of the subject. An Apple videoconferencing camera was placed on top of each monitor, with its lens 17 cm above the center of the monitor. The cameras pointed almost horizontally at the eyes of the actors, seated about 80 cm away. In all conditions, Actor 1 got the live image of camera 4, and Actor 2 got the live image of camera 3. The actors each had a unidirectional microphone placed in front of them. The signal from each microphone was amplified and fed to the respective speaker in the subject room. Actors used a numeric keyboard for selecting images in Powerpoint in the still-image condition. These images looked identical to the live camera feeds. Actor 1 had a disconnected computer keyboard with which he pretended to feed answers into a computer for verification.

ANALYSIS

We will now discuss how we analyzed video tape recordings to obtain measurements for each dependent variable.

Analysis of Deictic Verbal References

The experimenter and an independent observer scored the number of deictic verbal references used by subjects and actors during each full 15 minute session. Both observers were blind to experimental conditions. The independent observer was also blind to any experimental predictions or details. Before scoring, rules were agreed on what constituted a correct reference. Only deictic second-person pronouns were scored (i.e., the words *you* and *your* in "Are *you* sure that was *your* sentence?"), using these criteria:

- the reference was to one person only.
- the reference was not preceded or followed by a name.
- the reference was not used in a generic way.
- repetitions were scored only once (e.g., "You, you said").
- references in puzzle sentences were not scored.

Before scoring, both observers practiced the use of the above criteria on a subset of sessions not used for further analysis. After the training, the inter-observer reliability was determined on a new set of unused data. We obtained a significant correlation of r=.86 between observers (p<.001, 2-tailed). Subsequent analysis of 56 sessions, averaged between observers, yielded a mean number of deictic verbal references per session for the subject and each actor.

Turn-taking Analysis

We analyzed the first five minutes of each session for turn-taking behavior of subjects and actors. We used an automated procedure to analyze the speech patterns of individual speakers. As automated analysis could only be carried out on the isolated speech data of individual speakers, the two-track recordings were separated by hand into three separate digital audio tracks (22 KHz, 8- bit) for each session (inter-observer reliability r=.97, p<.001, 2-tailed). Like Sellen [18], we then used a fuzzy algorithm that counted the number of turns by each speaker [25]. First, this algorithm filled in 240 ms pauses to account for stop consonants, effectively removing pauses within words [5]. Then, talkspurt analysis removed pauses between consecutively spoken words. This way, talkspurts with a length of at least one phonemic clause were identified. The phonemic clause is regarded as the basic syntactic unit of speech. On average, it consists of 2-10 words with a duration of approximately 1.5 s, providing an estimate for finding the shortest uninterrupted vocalizations (see [11, 19] for a discussion). To identify talkspurts, a 13-sample (1.56 s) window moved over the speech data, filling samples within a 70% confidence interval around its mean position with speech energy if more than half of the samples in the window indicated speech activity, and if this speech activity was balanced within the window. Finally, if one of the speakers had a talkspurt of longer than a phonemic clause (i.e., 1.56 s) with everybody else being silent for the same length of time, a turn was assigned to her. The total number of turns by all participants, minus one, constituted the number of speaker switches. We checked the validity of the above turn-taking analysis algorithm by calculating the correlation over time between a turn classification produced by the algorithm and that produced by a trained linguist (see [25] for details). With a correlation of r=.64 (p<.001, 2-tailed) between classification methods the algorithm, which identified phonemic clauses simply by checking the duration of consecutive speech, did well against the human expert who used intonation and semantics of speech to identify phonemic clauses.

Analysis of Gaze at the Eyes of Subjects

Since we allowed the amount of gaze at the eyes received by subjects to vary between and within conditions, we needed to measure and control for this factor after the fact. We used an independent observer, with no knowledge about the experiment, to score the number of video frames in which actors appeared to gaze at the eyes of the subject. Before scoring, both the observer and the experimenter practiced observation on a subset of experimental data not used for further analysis. Inter-observer reliabilities averaged across conditions (Fisher Z transformed) were r=.94 for Actor 1 gaze and r=.87 for Actor 2 gaze (both p<.001). Subsequent scoring by the independent observer of the first 5 minutes of each of the 56 used sessions yielded a mean percentage of gaze at the subject's eyes per actor per session.

RESULTS

Results for each variable were calculated over the same 56 sessions. Where appropriate, analyses of variance (one-way ANOVAs) were carried out, evaluated at α=.05 level. Post-hoc comparisons were carried out using Student-Newman-Keuls (SNK) evaluated at α=.05 level. One planned comparison was carried out using a one-tailed t-test evaluated at α=.05 level. All data was normally distributed (Kolmogorov-Smirnov test, p>.05) with equal variances between conditions (Levene test, p>.05) unless indicated.

Task Performance

Analysis of variance showed no significant differences between conditions in the number of problems solved (F(2, 53)=1.39, p=.26).

	Deictic 2nd-pers. Pronouns *mean (s.e.)*		
Variable	Motion	Motion+GD	Still+GD
Subject number of references	1.3 (.3)	2.6 (.7)	3.6 (.9)
Mean actor # of references	1.8 (.3)	2.4 (.3)	2.2 (.3)

Table 1. *Means and standard errors for the number of deictic verbal references per first 5 session minutes.*

Deictic Verbal References

Table 1 presents the data summary for the number of deictic verbal references using second-person pronouns during the first five session minutes. A planned comparison showed that subjects used twice as many deictic verbal references in the condition conveying motion video with gaze direction than in the condition conveying motion video only (t(26.93)=1.82, p<.04, uneq. var., 1-tailed), thus confirming our hypothesis. Analysis of variance showed no significant differences between conditions in the mean number of deictic verbal references made by the actors (F(2, 53)=.88, p=.42).

	Speaker Turns *mean (s.e.)*		
Variable	Motion	Motion+GD	Still+GD
Number of speaker switches	14.7 (1.0)	15.1 (1.1)	18.9 (1.5)
Subject number of turns	5.9 (.4)	6.3 (.5)	7.7 (.7)
Actor 1 number of turns	6.1 (.5)	5.3 (.4)	7.8 (.7)
Actor 2 number of turns	3.8 (.4)	4.6 (.5)	4.5 (.5)

Table 2. *Means and standard errors for the number of speaker turns per first 5 session minutes.*

Turn-taking Behavior

Table 2 shows the data summary for the number of speaker switches and individual turns during the first five session minutes. Analysis of variance showed the number of speaker switches differed significantly across conditions (F(2, 53)=3.75, p<.03). Post-hoc comparisons showed this difference lay in the condition conveying still images with gaze direction (SNK, p<.05). There were over 25% more speaker switches in this condition. Differences across conditions in the number of individual turns by subjects showed a similar trend (F(2, 53)=3.17, p=.05). Here, post-hoc comparisons showed that the still image with gaze direction condition was different from the motion video-only condition (SNK, p<.05). Differences across conditions in the number of turns by Actor 1 were significant (F(2, 53)=5.39, p<.01). Post-hoc comparisons showed the still image with gaze direction condition was different from the other conditions (SNK, p<.05). There was no significant difference across conditions in the number of turns by Actor 2 (F(2, 53)=.94, p=.40). Actor 2 did show a practice effect over sessions (correlation between session order and number of turns per session r=.46, p<.001).

	Amount of Actor Gaze *mean (s.e.)*		
Variable	Motion	Motion+GD	Still+GD
Amount of gaze (% time)	13.8 (1.2)	6.6 (.8)	31.6 (1.5)

Table 3. *Means and standard errors for the percentage of actor gaze at the eyes of subjects per first 5 session minutes.*

	Estimated Means Adjusted for Gaze		
Variable	Motion	Motion+GD	Still+GD
Number of speaker switches	15.4	17.1	16.3
Subject number of turns	6.2	7.3	6.3

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

Removing Effects of Gaze at the Eyes

Table 3 shows the data summary of the percentage of actor gaze at the eyes of subjects during the first five session minutes. Analysis of variance showed differences in the mean percentage of actor gaze were significant across conditions (F(2, 53)=112.05, p<.0001). Post-hoc comparisons showed differences were significant between all conditions (SNK, p<.05). Subjects experienced about four times more actor gaze in the still image with gaze direction condition than in the motion video with gaze direction condition. There was a modest, but significant, linear relationship across conditions between the percentage of actor gaze at the eyes of subjects and the number of speaker switches (r=.37, p<.01 2-tailed) and between the percentage of actor gaze at the eyes of subjects and the number of subject turns (r=.34, p<.02 2-tailed). To adjust for this confounding effect, we performed a covariance analysis (with Roy Bargman Stepdown test). All assumptions for this analysis were met. Table 4 shows the resulting adjusted mean number of speaker switches and subject turns. With the effect of gaze at the eyes of subjects removed, differences between conditions in the number of speaker switches (F(2, 52)=.56, p=.58) or subject turns (F(2, 52)=.92, p=.41) were no longer significant.

Questionnaire

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

DISCUSSION

We will first consider potential confounding effects of actor behavior. For each of our dependent variables, we will then discuss possible explanations for our findings.

Confounding Effects of Actor Behavior

For the main dependent variables, we will now discuss to what extent results could have been due to differences in actor behavior other than treatment.

Confounding Effects on Deixis

On average, we found no significant differences between conditions in the number of deictic verbal references made by actors, making it unlikely they induced subject behavior by verbal means. We therefore believe it likely effects were in fact due to treatment variables.

Confounding Effects on Turn-taking

Most of the speaker switches occurred between subjects and Actor 1. Although Actor 2 demonstrated no treatment effect, like the subjects, Actor 1 did have more turns in the still image condition (see Table 2). One might therefore suspect that treatment effects on subject turn-taking were due to the turn-taking behavior of Actor 1. The positive linear relationship between the amount of actor gaze at the eyes of subjects and the number of subject turns, across conditions, makes this unlikely. Firstly, Actor 1 did not see this information. Secondly, if the act of looking at the camera lens confounded his turn-taking behavior in both motion video conditions, we would have found a *negative* linear relationship in those conditions. We therefore believe it likely effects were in fact due to treatment variables.

Explaining Findings on Task Performance

We found no significant differences in task performance between conditions. Monk et al. [13] already suggested that measures of task performance are typically sensitive only to gross manipulations of experimental treatment. Our task may simply not have been very sensitive to experimental treatment. This does, however, suggests that effects with regard to other dependent variables were in fact due to differences between conditions in the communication process itself, rather than to differences in the nature of the experimental task. This allows us to generalize findings to other task situations in which efficiency of the turn-taking process is the parameter of interest.

Explaining Effects on Deixis

Results regarding the number of deictic verbal references to persons were in line with expectations. Subjects used twice as many references when head orientation was conveyed. Our hypothesis was confirmed, stating that the presence of head orientation cues causes the number of personal deictic verbal references used to rise significantly. The actual ability of subjects to use deixis towards the actors did not differ between conditions. We believe subjects judged the usefulness of deixis by assessing visuo-spatial properties of the system on the basis of actor head orientation behavior.

Explaining Effects on Turn-taking

Results with regard to the turn-taking variables ran contrary to expectations. We might have expected the motion video-only condition to show fewer speaker switches and consequently fewer turns than conditions in which gaze directional cues of the head were conveyed. Instead, the still image condition scored over 25% more speaker switches than *both* motion video conditions. Differences across conditions in the number of individual turns by subjects showed a similar trend. It is evident that the explanation for these results cannot lie in the absence of non-verbal visual cues in the still image condition. Both literature and earlier presented arguments suggest that any potential effects of this treatment variable would have gone in the opposite direction, with fewer turns when there are fewer nonverbal visual cues [4]. The sense of anonymity in the still image condition may have had a positive effect on turn-taking, but only one subject stated this in the questionnaire. As the analysis of covariance demonstrates, a much more satisfactory explanation for our findings is the confounding influence of actor gaze at the eyes of subjects. The linear relationship between the amount of gaze at the eyes perceived by subjects and their turn-taking behavior was sufficiently strong to explain our findings. In the still image condition, whenever the frontal image was selected, the actors would *always* appear to gaze at the eyes of the subject. This was not the case in the other conditions. When differences in the amount of gaze at the eyes of subjects were removed, differences in turn-taking behavior disappeared also. As discussed, people are very good at judging the angle of frontal eye gaze at their faces. One the one hand, we can therefore regard our covariate as a measure for the ability of subjects to discriminate whether they *themselves* were being looked at. On the other, our covariate was a measure for the *amount* of visual attention subjects received. This yields two, possibly complementary, explanations as to why there were more speaker switches in the still image condition:

1) *Knowing Your Turn.* According to a study by Vertegaal [25], in multiparty conversation, gaze at the eyes codes whom is being addressed or listened to with great certainty. This information is not coded by other non-verbal means. Although head orientation might be used to see whom others address, this cue was not vital to the turn-taking process. Instead, subjects used gaze at their eyes to see when they *themselves* were addressed or expected to speak. Difficulties in conveying gaze at the eyes caused deficiencies in subjects obtaining and yielding the floor in both motion video conditions. This explanation is consistent with Kendon's findings regarding functions of gaze in dyadic turn synchronization [12].
2) *Keeping Your Distance.* According to Argyle and Dean's Equilibrium of Intimacy theory [3], like proximity, people use gaze at the eyes to regulate social distance to each other. When the level of intimacy between people is disturbed (either too high or too low), they feel uncomfortable [2]. The mean percentage of gaze at the eyes in the still image condition was almost exactly that found by Exline for triadic face-to-face conversations [7]. In the other conditions percentages were much lower, yielding a much lower level of intimacy. As this lower level of intimacy could not easily be compensated by other means, subjects were less inclined to take the floor in those conditions.

With regard to our explanations for the mechanism behind the effect of gaze at the eyes on turn-taking, we found clear support for our first explanation in the questionnaire. Subjects found it significantly easier to observe who was talking to whom in the still image condition. Other qualitative observations, including comments by subjects, seemed mostly in line with our empirical findings.

CONCLUSIONS AND RECOMMENDATIONS

We first present our empirical conclusions, after which we outline our recommendations for the design of multiparty mediated systems.

Empirical Conclusions

In this paper, we presented an empirical evaluation of the effects of gaze directional and other non-verbal visual cues on multiparty mediated communication. Groups of three participants (two actors and one subject) solved language puzzles in three mediated communication conditions. In addition to speech, each condition presented a different selection of images of the actors' upper torsos to subjects: (1) frontal motion video showing actor gaze 14% of time; (2) motion video with head orientation and 7% actor gaze; (3) still images with head orientation and 32% actor gaze. Effects of the amount of actor gaze perceived by subjects were isolated retroactively. Results show the presence of head orientation cues caused subjects to use twice as many deictic verbal references to persons. We believe this was due to differences between conditions in the subjects' estimate of the effectiveness of head pointing in disambiguating deixis. Across conditions, we also found a significant positive linear relationship between the amount of actor gaze at the eyes of subjects and the number of subject turns (r=.34) and speaker switches (r=.37). We did not find a similar effect on turn frequency of the presence of other non-verbal upper-torso visual cues, including head orientation. As evidenced by subject performance in our still image condition, the potential increase in turn frequency may be in the order of 25% when gaze at the eyes is conveyed in a manner that preserves face-to-face characteristics. We believe there are two reasons why the presence of gaze at the eyes has a positive effect on turn frequency in multiparty mediated communication. Firstly, gaze is used to determine when a person is speaking or listening to you. In group communication, it not obvious who will be the next speaker when the current speaker falls silent. Seeing when they were being addressed or expected to speak made it easier for subjects to obtain or yield the floor. We found clear support for this in our questionnaires. Subjects found it easier to observe who was talking to whom in the condition with normal percentages of gaze. Secondly, gaze seems to be used to regulate social distance. Subjects may have felt the level of intimacy with their conversational partners was disturbed when there was not enough gaze at their eyes, making them less inclined to take the floor.

Design Recommendations

We believe that a higher turn frequency is an indication of a more natural, and perhaps more efficient, turn-taking process. As discussed, most empirical studies seem to confirm this rationale. Although effects of a higher turn-taking efficiency may be dependent on the task situation, we believe one *can* generalize that synchronous interactive group communication systems should preserve gaze directional cues, especially gaze at the eyes. With respect to the design of such systems, we therefore formulated the following incremental requirements:

1) Preservation of relative position. Relative viewpoints of participants should be based on a common reference point (e.g., around a shared workspace), providing basic support for the use of a common external context in deictic referencing.
2) Preservation of head orientation. Its representation eases the use of deictic references to persons.
3) Preservation of gaze at the eyes. Allowing participants to perceive gaze at each other's eyes eases management of turn-taking and may aid tele-presence.

Our findings do *not* suggest that motion video should not be conveyed. Rather, they suggest that when developing software for group communication, one should consider conveying gaze directional cues first. Whether it is for highly personal or business communication, participants need to be able to seek or avoid gaze at each other's eyes according to their own personal or cultural preferences.

We will now briefly discuss how the above requirements could be implemented in group communication systems. If motion video *is* conveyed, we suggest the use of a multiple camera setup, in which each participant has a camera for each other participant. By placing each camera behind a semi-transparent screen displaying the image of that participant, basic support for the above requirements can be provided [1]. The larger the distance of head to screen, or the smaller the projected images, the more head movement of users is tolerable without impairing conveyance of gaze at the eyes. We believe office-size systems such as MAJIC [15] therefore provide the best implementation of our requirements currently possible with motion video. Note that the need for multiple video streams does mean bandwidth use will not scale well with the number of users in such systems [23]. When still images are used, our requirements can be implemented using very little bandwidth indeed. The GAZE Groupware System [23] implements all requirements in a transparent and noncommand fashion. It measures whom participants look at using a desktop eyetracking system. It then orients their picture so that it faces the person they look at [22-27].

ACKNOWLEDGEMENTS

We gratefully acknowledge the contributions of Mirjam Netten, Bart Schermer, TU Delft Wit-lab, Ronald Leenes, Armanda Zandberg, Luuk Lagerwerf, Andrew Monk and Bill Buxton.

REFERENCES

1. Acker, S. and Levitt, S. Designing videoconference facilities for improved eye contact. Journal of Broadcasting & Electronic Media 31(2), 1987, pp. 181-191.
2. Argyle, M. and Cook, M. Gaze and Mutual Gaze. London: Cambridge University Press, 1976.
3. Argyle, M. and Dean, J. Eye-contact, distance and affiliation. Sociometry 28, 1965, pp. 289-304.
4. Argyle, M., Lalljee, M., and Cook, M. The effects of visibility on interaction in a dyad. Human Relations 21, 1968, pp. 3-17.
5. Brady, P.T. A statistical analysis of on-off patterns in 16 conversations. The Bell System Technical Journal (Jan.), 1968, pp. 73-91.
6. Cline, M.G. The perception of where a person is looking. American Journal of Psychology 80, 1967, pp. 41-50.
7. Exline, R.V. Explorations in the process of person perception: Visual interaction in relation to competition, sex and need for affiliation. Journal of Personality 31, 1963, pp. 1-20.
8. Fokkema, S.D. and Dirkzwager, A. Ruimtelijk Inzicht; Taalgebruik II, Zinnen. Differentiële aanlegtests. Amsterdam: Swets & Zeitlinger, 1960.
9. Isaacs, E. and Tang, J. What video can and can't do for collaboration: a case study. In Proceedings of ACM Multimedia '93. Anaheim, CA: ACM, 1993.
10. Ishii, H. and Kobayashi, M. ClearBoard: A Seamless Medium for Shared Drawing and Conversation with Eye Contact. In Proceedings of CHI'92. Monterey, CA: ACM, 1992.
11. Jaffe, J. and Feldstein, S. Rhythms of Dialogue. New York, NY USA: Academic Press, 1970.
12. Kendon, A. Some Function of Gaze Direction in Social Interaction. Acta Psychologica 32, 1967, pp. 1-25.
13. Monk, A., McCarthy, J., Watts, L., and Daly-Jones, O. Measures of Process. In Thomas, P. (Ed.), CSCW Requirements and Evaluation. Berlin: Springer Verlag, 1996, pp. 125-139.
14. O'Connaill, B., Whittaker, S., and Wilbur, S. Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication. Human Computer Interaction 8, 1993, pp. 389-428.
15. Okada, K.-i., Maeda, F., Ichikawaa, Y., and Matsushita, Y. Multiparty Videoconferencing at Virtual Social Distance: MAJIC Design. In Proceedings of CSCW '94. Chapel Hill, NC: ACM, 1994, pp. 385-393.
16. Rosenthal, A.H. Two-way television communication unit. United States Patent 2 420 198, 1947.
17. Rutter, D.R. and Stephenson, G.M. The Role of Visual Communication in Synchronising Conversation. European Journal of Social Psychology 7, 1977, pp. 29-37.
18. Sellen, A.J. Remote conversations: the effects of mediating talk with technology. Human Computer Interaction 10(4), 1995.
19. Sellen, A.J. Speech Patterns in Video-Mediated Conversations. In Proceedings of CHI'92. Monterey, CA: ACM, 1992, pp. 49-59.
20. Short, J., Williams, E., and Christie, B. The Social Psychology of Telecommunications. London: Wiley, 1976.
21. Stapley, B. Visual enhancement of telephone conversations. PhD Thesis. Empirial College, 1972.
22. Vertegaal, R. Conversational Awareness in Multiparty VMC. In Extended Abstracts of CHI'97. Atlanta, GA: ACM, 1997, pp. 6-7.
23. Vertegaal, R. The GAZE Groupware System: Mediating Joint Attention in Multiparty Communication and Collaboration. In Proceedings of CHI'99. Pittsburg, PA: ACM, 1999.
24. Vertegaal, R. GAZE: Visual-spatial Attention in Communication. Video Paper. In Proceedings of CSCW '98. Seattle, WA USA: ACM, 1998.
25. Vertegaal, R. Look Who's Talking to Whom. PhD Thesis. Enschede, The Netherlands: Cognitive Ergonomics Department, Twente University, 1998.
26. Vertegaal, R., Velichkovsky, B., and Van der Veer, G. Catching the Eye: Management of Joint Attention in Cooperative Work. SIGCHI Bulletin 29(4), 1997.
27. Vertegaal, R., Vons, H., and Slagter, R. Look Who's Talking: The GAZE Groupware System. In Summary of CHI'98. Los Angeles, CA: ACM, 1998.
28. Von Cranach, M. and Ellgring, J.H. The perception of looking behaviour. In Von Cranach, M. and Vine, I. (Ed.), Social Communication and Movement. London: Academic Press, 1973.