Remote Conversations: The Effects of Mediating Talk With Technology

Article *in* Human-Computer Interaction · December 1995

DOI: 10.1207/s15327051hci1004_2

1 author:

Abigail Sellen
Microsoft
233 PUBLICATIONS **14,615** CITATIONS

HUMAN–COMPUTER INTERACTION, 1995, Volume 10, pp. 401–444

Remote Conversations: The Effects of Mediating Talk With Technology

Abigail J. Sellen
MRC Applied Psychology Unit and
Rank Xerox Research Centre, Cambridge Laboratory

ABSTRACT

Three different videoconferencing systems for supporting multiparty, remote conversations are described and evaluated experimentally. The three systems differed by how many participants were visible at once, their spatial arrangement, and control over who was seen. Conversations using these systems were compared to same-room (Experiment 1) and audio-only (Experiment 2) conversations. Specialized speech-tracking equipment recorded the on–off patterns of speech that allowed objective measurement of structural aspects of the conversations, such as turn length, pauses, and interruptions. Questionnaires and interviews also documented participants' opinions and perceptions in the various settings.

Contrary to expectation, systems in which visual cues such as selective gaze were absent produced no differences in turn-taking or in any other aspect of the structure of conversation. In fact, turn-taking was unaffected even when visual information was completely absent. Overall, only the same-room condition showed any significant differences from any other condition; people in the same room produced more interruptions and fewer formal handovers of the floor than in any of the technology-mediated conditions. In this respect, the audio-only and video systems examined in these studies were equivalent. However, analyses of

Abigail J. Sellen is a cognitive psychologist interested in the design and evaluation of video technology in support of distributed work; she is a research scientist jointly appointed to Rank Xerox Research Centre, Cambridge Laboratory (EuroPARC) and the MRC Applied Psychology Unit.

CONTENTS

participants' perceptions showed that participants felt that visual access in mediated conversations was both important and beneficial in conversation. Further, there were indications that the particular design of the different video systems did affect some aspects of conversational behavior, such as the ability to hold side and parallel conversations.

1. INTRODUCTION

For decades, telephones have allowed people to "meet" and talk from remote locations. But, with recent technological advancements, the possibility of also providing a visual channel for communication is now more financially and technologically feasible. A renewed interest in videoconferencing technology as a way of enabling people to meet from remote locations is reflected by the growing number of experimental video systems, or *media spaces,* in research laboratories around the world (e.g., Bly, Harrison, & Irwin, 1993; Fish, Kraut, Root, & Rice, 1992; Gaver

et al., 1992; Mantei et al., 1991). Many of these have become the focus of a flurry of research on technology for collaborative work.

Part of the promise of video technology is the possibility of being able to simulate for remote participants some of what people share when they meet in the same physical space. What researchers have yet to understand, however, is the extent to which the visual channel can provide these benefits. Conversely, researchers have yet to fully understand how video-mediated interaction might be fundamentally different from sharing the same physical space, including the possibility that it might significantly alter conversational behavior, either for the worse or for the better.

In this article, I examine and compare the properties of conversations in five different situations: when participants are in the same room (i.e., face-to-face); when only an audio connection is available; and when held over three kinds of videoconferencing systems. I pay particular attention to how the medium of interaction might affect the temporal surface structure of conversation, measured by automatically tracking the on–off patterns of speech by the participants.

1.1. Role of Visual Information in Conversation

Videoconferencing technology is built on the assumption that the ability of two people to see as well as hear one another in conversation is valuable. There are clearly situations in which sharing visual information is critical to the collaborative task or at least makes interaction easier (i.e., redesigning some physical artifact). Recently, studies of the use of video in collaborative tasks have begun to focus on the importance of such task-centered views (e.g., Gaver, Sellen, Heath, & Luff, 1993; Nardi et al., 1993). However, the explicit goal of videoconferencing is often stated as one of simulating face-to-face meetings; the implicit assumption is that visual access to at least the head and shoulders of the person with whom one is conversing is a critical part of the interactional process. Accordingly, most videoconferencing systems (both commercial and experimental) focus on making this face-to-face view available.

To some extent, this assumption makes intuitive sense. Conversation is much more than the exchange of verbal messages. Talk is accompanied by a rich array of nonverbal behaviors that we use and respond to in conversation, including gaze, posture, gesture, and facial expression. Gaze alone is thought to serve a number of important functions including: helping to regulate and synchronize conversation, providing feedback on how the communication is being perceived by the listener, communicating emotions, and communicating the nature of the interpersonal relationship (Argyle & Cook, 1976; Kendon, 1967).

Despite the assumption that visual access matters, there is very little empirical evidence to suggest that video adds significantly to audio in the support of collaborative tasks. Ochsman and Chapanis (1974) concluded,

as a result of comparing 10 different communication modes, that, although inclusion of a voice channel is critical, the presence of a video channel has no effect on communication behavior. Research has also shown that, in comparisons among face-to-face, video plus audio, and audio-only modes of interaction, access to visual information has no significant effect on the outcome of intellectual, decision-making, and creative tasks. However, tasks that rely on social cues such as situations of conflict, bargaining, and negotiation do tend to be affected by the presence or absence of the visual communication channel (see Argyle & Cook, 1976; McGrath, 1984; Short, Williams, & Christie, 1976 for reviews.) This led researchers such as Short et al. (1976) to conclude that the most important reason for access to visual information is the transmission of "social presence." When social presence is relatively unimportant for successfully accomplishing the task, access to audio alone seems to be sufficient.

Although empirical studies have failed to find overwhelming effects with regard to task outcome, one may also ask how face-to-face visual access affects the interactional process of conversation. The process of conversation is, to a large extent, rule-bound, and the conversants highly skilled at respecting and adhering to the rules of the game. There are many things that are, thus, predictable about conversation; one is that, in general, only one person at a time speaks or "holds the floor." Speaker turns, interruptions, and passing the floor from one speaker to the next are accomplished in a variety of subtle and mutually understood ways (Duncan, 1972; Sacks, Schegloff, & Jefferson, 1974).

Conversation, therefore, can be characterized according to its surface structure. For example, it can be characterized in terms of the frequency and duration of turns, length of pauses, and number of interruptions. Because these are measurable, this offers an alternative to task outcome as a way of assessing the effect of conversational medium. Such measures are of interest because they may reveal the manner in which a conversation is conducted. For example, if there are difficulties in turn management and regulation, this may reveal itself in unusually long turns, frequent interruptions, or very long pauses.

Effects of Reducing Visual Cues on the Conversational Process

The rationale for predicting that visual cues will influence measures of conversational process is partly based on the assumption that the ability to selectively gaze at others is important in the regulation of conversation. It has been suggested that gaze and head-turning play an important role in keeping the floor, taking and avoiding the floor, and suggesting who should speak next (Argyle & Cook, 1976; Kendon, 1967). For example, Kendon (1967) found that the gaze by a speaker at a listener increases just before ending a long utterance, and that when there is no such terminal

gaze, there is more likely to be a pause before switching speakers. He found that a speaker will tend to look away at the beginning of a turn and then terminate the turn with a sustained gaze, usually at the presumptive next speaker. A speaker wishing to hold the floor at a pause point will look away from the listener.

Since the late 1960s, a number of studies have been carried out measuring the surface structure of speech, with the expectation that there will be detrimental effects on turn-taking when visual cues are reduced. Almost all of these studies have been in the context of dyadic conversations and have typically involved comparisons of face-to-face versus audio-only conditions (e.g., via an intercom system). Such studies generally predicted that when visual cues are reduced, poor turn-taking will result in more interruptions (more simultaneous speech) and more pauses in audio-only conversations. If there are problems in exchanging the floor, exchanges will occur less often, resulting in longer and fewer utterances in audio-only conditions.

Results of these studies, however, have often gone counter to these hypotheses, finding fewer interruptions, shorter periods of simultaneous speech, and shorter pauses in audio-only conditions (see Figure 1 for a summary). To add to the confusion, findings on almost all of these measures have been inconsistent across studies (see Figure 1), with the possible exception of amount of simultaneous speech, which, counter to prediction, is consistently less when visual cues are reduced.

Some of the inconsistencies are undoubtedly due to definitional and methodological differences. Nonetheless, it is safe to say that there is no strong evidence that a reduction in visual cues causes a breakdown in conversational synchronization, at least in the context of dyadic conversations contrasting face-to-face and audio-only conversations. Indeed, despite the literature that suggests the importance of gaze in conversational regulation, there are other studies that question its importance. For example, attempts to replicate Kendon's (1967) results have failed (Beattie, 1983; Rutter, Stephenson, Ayling, & White, 1978). Psychologists such as Duncan (1972) have also downplayed the importance of gaze in managing conversation and have instead stressed a variety of nonvisual cues such as paralanguage and syntax.

Effects of Videoconferencing on the Conversational Process

Because of technological advances in long-distance video, rather than asking what the effects of reduced visual cues may be, a question more pertinent to today's concerns is to ask what the effects of different kinds of video-mediation may be on measures of conversational process. Conventional videoconferencing systems typically involve a single camera and monitor for each remote participant, and the fact that cameras have a fixed field of view and usually cannot be controlled by the viewer limits the

Figure 1. **Summary of the literature on the effects of different conditions on measures of conversational process.**

Dependent Variable Measured	Conditions Compared[a]			Findings[b]	References
	Face to Face	Video Mediated	Audio Only		
Number of interruptions	X		X	FF > AO	Cook & Lalljee (1972)
				FF > AO	Rutter & Stephenson (1977)
				FF < AO	Argyle et al. (1968)
	X	X		FF > VM[c]	O'Conaill et al. (1993)
				FF = VM	O'Conaill et al. (1993)
Duration of simultaneous speech	X		X	FF > AO	Jaffe & Feldstein (1970)
				FF > AO	Rutter & Stephenson (1977)
	X	X	X	FF > VM = AO	Williams (1978)
				FF > VM > AO	Rutter et al. (1981)
Pause length	X		X	FF = AO	Cook & Lalljee (1972)
				FF > AO	Jaffe & Feldstein (1970)
				FF < AO	Argyle et al. (1968)
Length of utterances/turn	X		X	FF < AO	Rutter & Stephenson (1977)
				FF > AO	Cook & Lalljee (1972)
	X	X		FF < VM[c]	Cohen (1982)
				FF < VM[c]	O'Conaill et al. (1993)
				FF = VM	O'Conaill et al. (1993)
Number of utterances/turns	X		X	FF > AO	Argyle et al. (1968)
				FF < AO	Cook & Lalljee (1972)
	X	X		FF > VM[c]	Cohen (1982)
				FF > VM[c]	O'Conaill et al. (1993)
				FF = VM	O'Conaill et al. (1993)
	X	X	X	FF = VM = AO	Rutter et al. (1981)

[a]Xs indicate conditions compared. [b]The symbols < and > are taken here to mean a statistically significant difference; = indicates a failure to find a statistically significant difference. FF = face to face, AO = audio only, VM = video mediated. [c]Video system with a considerable audio and video transmission lag present.

visual cues available. But mediating conversation with video technology implies much more than simply reducing visual cues. The effects of videoconferencing technology may be more complex for a variety of reasons.

In videoconferencing, failure to make eye contact tends to be a problem because of the separation of camera and monitor. Short et al. (1976) suggested that this and the lack of head-turning and directional gaze cues may cause problems in turn-taking behavior. In addition, in video-mediated meetings, the principle of reciprocity does not always hold (i.e., "If I can see you, you can see me") and this inherent asymmetry may also contribute to problems of regulation. For example, one participant may believe that he or she is making eye contact, but this is not perceived by the other participant. Similarly, participants from time to time will look at the camera, and this may be interpreted by the receiver of the look as a signal.

Of course, there are implications–beyond those of turn-taking–that stem from the fact that participants over videoconferencing systems do not share a physical space. Remote participants cannot mutually determine the physical distance between themselves as they would if they shared the same space. Speakers have no sense of how their voices are perceived by listeners. Other differences are more subtle, such as the relative impotence of gestures and gaze in securing another's attention through video (Heath & Luff, 1991).

Few published studies have focused on the surface characteristics of video-mediated conversations. Those that have (also summarized in Figure 1) to some extent suggest that video conditions may have similar effects on conversation patterns as audio-only conditions, in comparison with a face-to-face condition.[1] For example, they found that video-mediated conversations contain less simultaneous speech (Rutter, Stephenson, & Dewey, 1981; Williams, 1978) than face-to-face conversations. Further, Rutter et al. (1981) found that a video condition fell somewhere between a face-to-face and an audio-only condition with respect to the duration of simultaneous speech. On the basis of these results, Rutter (1987) argued that video mediation is essentially attenuating the visual cues available in face-to-face meetings. Whether this attenuation view of video mediation is correct, one inference that can be drawn is that simply giving conversants visual access via video technology does not render the situation similar to being physically copresent, if these speech measures are any indication.

To uncover why differences between face-to-face and videoconferencing conversations exist, it is perhaps more fruitful to concentrate on the

1. The picture with regard to length and number of speaker turns is unclear, however. There is some indication that there are fewer and longer speaker turns in video-mediated conditions than under face-to-face conditions for video systems if they have an audio and video transmission lag (Cohen, 1982; O'Conaill, Whittaker, & Wilbur, 1993).

behaviors that any particular videoconferencing system affords rather than to focus on generic, hypothetical entities such as cues. Research has not really addressed the issue of how the specifics of the design of different videoconferencing systems might affect behavior (although for a discussion of some of the issues, see Gaver, 1992). Videoconferencing systems vary a great deal along a number of different dimensions. These include: audio and video quality, audio or video lag, number of participants visible at any one time, number of speakers who can be heard simultaneously, camera angle, number of people per camera, and the presence of a confidence monitor (picture of the outgoing video signal). For the purpose of the design of these systems, it is important that researchers know how such complex differences affect the nature of interaction among conversants.

1.2. Three Multiparty Videoconferencing Systems

In the two experiments that follow, I focus on three systems designed to support *multiparty* conversations, that is, conversations among more than two people. One reason for this is that designing systems for groups larger than two is a greater challenge.[2] It raises questions such as who should occupy the screen, how much of the group should be visually represented, and how many screens should be used. More design variables are therefore at issue. In this section, I describe the three systems and summarize their differences and similarities.

Picture-in-Picture (PIP) System

The most common method of supporting multiparty videoconferences involving several sites is the use of a PIP approach. Using this technique, a single video screen is divided into quadrants with each quadrant containing the output of one video source. Typically, the images are combined at some central location using a PIP device, and then the output is broadcast to each participant. This means that, in a four-party conversation, each participant sees the other three people in the conversation in addition to himself or herself (see Figure 2), and all participants see the same image.

One implication of this design solution is that it fails to support selective gaze and selective listening. Unlike conversants sharing the same physical space, participants no longer occupy distinct physical spaces but, rather, share a single screen. They also are seen through a single camera that is

2. Apart from the technology, conversational structure in multiparty groups is much more complex than that among dyads. As soon as a third party is introduced, the next turn is no longer guaranteed to the noncurrent speaker. As the numbers increase, competition for the floor also increases. Four-party conversations are notably different from three-party conversations in that the presence of four people provides for the possibility of two different ongoing conversations.

Figure 2. **A meeting using a PIP device.**

usually mounted on top of or to the side of the monitor. This means participants are limited in their ability to show that they are directing their gaze at others through head-turning or selective eye contact with them. In turn, participants are also limited in their ability to be aware of who is visually attending to them.

In addition, because the audio in this kind of system typically emanates from a single audio speaker, it is reasonable to assume that it is more difficult to selectively listen to different, parallel conversations. In part, what makes selective attention to parallel conversations possible (i.e., the "Cocktail Party Effect") is that the different speech streams originate from distinct points in space.

Hydra System

As a direct result of these concerns about the PIP approach, we developed Hydra, a system that uses multiple cameras, monitors, and speakers (Sellen, Buxton, & Arnott, 1992). Hydra simulates a four-way, round-table meeting by placing a camera, monitor, and speaker in the place that would otherwise be held by each remote participant. Using this technique, each person is presented with a unique view of each remote participant, and that view and its accompanying voice emanate from a distinct location in space. Figure 3 shows Hydra in use in a four-way conversation.

Figure 3. **A meeting among four people using three Hydra units. Each Hydra unit contains a video monitor, camera, and loudspeaker.**

The fact that each participant is represented by a separate camera/monitor pair means that the act of gazing toward someone can be conveyed. In other words, when Person A turns to look at Person B, B is able to see A turn to look toward B's camera. Looking away and gazing at someone else also can be conveyed, and the direction of head-turning indicates at whom one is looking.[3] Furthermore, because the voices come from distinct locations, the design is intended to make it easier to selectively attend to different speakers who may be speaking simultaneously.

LiveWire System

The third system differs from the other two primarily in that it displays only the current speaker to the group, rather than allowing all participants

3. Audio and video connections for Hydra can be configured by software that ensures that a consistent "around the table" mapping is made for each person. In other words, the switching network ensures that if Person A appears in the center unit for Person B, then Person B appears in the center unit for Person A. Similarly, if Person C appears to Person A's right, then Person C appears to Person B's left, and so on. In this way, head-turning and gaze cues deliver consistent and meaningful information.

to be continuously visually present. This system, called LiveWire,[4] is a version of a voice-switched video system. Voice-switched systems exploit the fact that, in a conversation, usually only one person speaks at a time.

This particular implementation, like the PIP system, involves a single camera and monitor for each participant but shows the current speaker on the full screen to everyone else, doing the switching automatically.[5] A further constraint in LiveWire is that whoever is currently speaking continues to see the last speaker; the assumption is that the current speaker is more likely to be responding to the previous speaker than to anyone else and therefore wants to see that person's reactions. Further, as in the PIP system, and unlike Hydra, all participants' voices are heard through a single audio speaker.

1.3. Comparing the Five Conversational Contexts

The comparison of these three systems with each other and with a same-room (i.e., face-to-face) and audio-only situation was designed to be, as much as possible, carried out in the tradition of controlled laboratory experimentation. In this way, it is different from other research that has examined multiparty video systems in real, work environments (e.g., O'Conaill et al., 1993). There are benefits and drawbacks to both approaches, and it must be said that, in adopting a laboratory approach, many factors that are without doubt important elude scrutiny or are sacrificed for the sake of experimental control. (For example, by recruiting participants as opposed to examining working groups "in situ," things like interpersonal dynamics, extended experience with a particular system, and organizational politics are either radically altered or precluded.) However, the aim in this study was to exert a reasonable amount of control in the design of the experiments so that comparisons could be systematic and some cause–effect relations could be inferred as a result of these comparisons.

Having said that, preserving the fundamental features of each of the video systems meant that, of necessity, they often differed from each other in more than one way. In this sense, the comparisons were not truly controlled. However, whenever possible, those features that could be were held constant. For example, each of the systems used full duplex audio and good quality images; none of the three systems had any perceptible audio

4. Implemented by Mike Sheasby of the Dynamic Graphics Project, University of Toronto.

5. This is accomplished using a pitch tracking device that captures voice signals, converts them into digital signals, and sends them via a Musical Instrument Digital Interface connection to a Macintosh II computer. The computer then controls the video switching network.

or video lag; and each system made use of similar camera angles, with one person per camera. Figure 4 summarizes the main similarities and differences among the five conditions.

2. EXPERIMENTS

The purpose of the experiments was twofold. First, they were designed to ask how (or even whether) video mediation affects conversation. Specifically, the question of interest was how are video-mediated conversations different from same-room or audio-only conversations, if at all? Same-room and audio-only contexts were used as baseline conditions in Experiments 1 and 2, respectively, for reasons outlined later.[6]

Second, these experiments were designed to investigate the ways in which the unique designs of three different videoconferencing systems affect conversational behavior. Although the three systems differ from each other in a variety of ways, two dimensions were of particular interest. The first is the support of selective gaze and listening, which Hydra was designed to support but the PIP system does not. The comparison of the PIP, Hydra, and same-room conditions is thus the basis for the first experiment; the same-room condition is the situation in which we know selective gaze and listening cues are used. The second dimension is the amount of visual access to other participants: the PIP system allows access to all others, the LiveWire system allows access to one person at a time, and the audio-only condition allows no visual access. This comparison forms the basis of a second experiment.

In both experiments, the primary measures of interest were the statistical properties of the resulting conversations, as measured by the on–off patterns of spontaneous speech in the various contexts. However, complete audio and video recordings of each conversation were made to capture any other kinds of conversational behaviors the different contexts might afford. In addition, participants' subjective opinions of the five conditions were collected and analyzed.

Because both the method and the analysis for the two experiments were essentially the same, they are described concurrently. There were three conditions in each experiment; the PIP condition appeared in both. Thus,

---

6. Although there were specific reasons for choosing the same-room and audio-only conditions as the control conditions in these particular experiments, in general it is difficult to know how to choose the right control against which to assess the effects of the mediating technology. One often gets the sense that conversations in the same room are put forth for comparison because they are thought, in some way, to be "ideal." No such assumptions were made in these experiments. Rather, the same-room context is used as a benchmark mainly because it is the situation about which we know the most, both from an empirical and personal point of view.

Figure 4. **Summary of some of the main features of each of the five conversational contexts compared in the experiments.**

Variable	Same Room	PIP	Hydra	LiveWire	Audio-Only
Video quality	Unmediated visual access	Good quality, color images, no lag	Good quality, color or black-and-white images, no lag	Good quality, color images, no lag	N/A
Number of participants simultaneously visually available	All	All	All	One other	None
Selective gaze supported	Yes	No	Yes	No	No
Number of images per monitor screen	N/A	4 continuously available	1 continuously available	1 voice switched	N/A
Image size	N/A	9.5 cm diagonal per quadrant	Color units (8 cm diagonal), black-and-white units (12 cm diagonal)	38 cm diagonal	N/A
Camera angle	N/A	Head-and-shoulders shot of one person	Head-and-shoulders shot of one person	Head-and-shoulders shot of one person	N/A
See own image?	No	Yes	No	No	No
Audio quality	Unmediated audio success	Full duplex, no lag	Full duplex, no lag	Full duplex, no lag	Full duplex, no lag
Number of participants who can be heard simultaneously	All	All	All	All	All
Selective listening supported?	Yes	No	Yes	No	No

the two experiments covered five different conditions in all, and the results for all five conditions are presented together. The reader is cautioned, however, that as each experiment used a within-subject design, direct comparisons of the data across the two experiments are confounded by the fact that the numbers come from different groups of participants. Nonetheless, the results can be interpreted relative to the PIP condition in each of the experiments.

2.1. Experiment 1 Predictions

The overall expectation was that the system that supports selective gaze and selective listening (Hydra) would give rise to behavior more like the same-room condition than the system that does not (PIP). If these kinds of cues are used in normal face-to-face conversation, then presumably their absence will affect the structure of the conversation and will be revealed in the PIP condition. Turn-taking behavior was expected to be one aspect of behavior affected, although the issue of what sorts of measures might indicate problems in turn-taking is a difficult one. As was earlier pointed out, research has failed to provide the expected results in terms of measures such as interruptions, pauses, and the like, when visual cues are reduced. Where possible, therefore, predictions were based on previous findings.

Turn Frequency, Duration, and Distribution. Same-room conversations would result in more shorter turns per session than video-mediated conversations. This is consistent with Cohen's (1982) finding that there were almost twice as many speaker switches in a same-room meeting than in a Picturephone meeting and is in line with some studies of audio-only versus same-room meetings (e.g., Argyle, Lalljee, & Cook, 1968; Rutter & Stephenson, 1977).

The distribution of turns among speakers would be more equal in same-room conversations than in video conversations. There was no previous literature on which to base this prediction. Rather, it was hypothesized that, if it is more difficult to take the floor or if speakers are more reluctant to try in the video-mediated conditions, floor changes would occur less often. Thus, in the video conditions, dominant speakers would hold the floor more, and nondominant speakers would have the floor less.

The overriding expectation was that the Hydra approach would tend to produce conversational patterns more similar to same-room conversations than the PIP approach would, in terms of these measures.

Simultaneous Speech and Speaker Switching. I expected that there would be more simultaneous speech in the same-room condition than in

the video conditions–participants in the same room would be more interruptive of each other. Previous research (Cohen, 1982; Rutter et al., 1981; Williams, 1978) found more simultaneous speech in same-room meetings than in video-mediated meetings, suggesting that same-room meetings are thus less polite, less orderly, and more interactive. Again, Hydra was expected to fall between the same-room and PIP conditions.

Handovers and Parallel Conversations. Being able to selectively attend to one person to the exclusion of others is presumably the basis for conversational acts such as implicitly signalling to whom one is paying attention or addressing a comment. Selective attention is presumably also necessary for making asides or holding a conversation with one person while another conversation is taking place among others. I therefore predicted that Hydra would afford various conversational acts that the PIP system would not, because it allows selective gaze and listening. One prediction was the occurrence of fewer formal or explicit handovers of the floor in the Hydra situation compared to the PIP system, because selective looking in Hydra is an implicit way of accomplishing handovers. The other prediction was that side and parallel conversations would occur in Hydra (where two conversational "threads" occur simultaneously) but would not occur in the PIP system.

2.2. Experiment 2 Predictions

In Experiment 2, the effects of selective gaze and listening were no longer the central issue, because the three conditions compared–PIP, LiveWire, and an audio-only condition–provide no explicit support for these cues. Rather, in comparing these three conditions, the amount of visual information was at issue, ranging from all participants being visually available (PIP) to one person at a time being visually available (LiveWire) to a complete absence of visual information (audio-only condition). Predictions for this experiment were formulated mainly by referring to previous research by Rutter and colleagues (summarized in Rutter, 1987). Rutter (1987) proposed that the finding of more interruptive and simultaneous speech in audio plus video conditions when compared with audio-only conditions indicates not a breakdown in synchronization when visual cues are reduced, but a breakdown in spontaneity. He argued that without visual access, there are fewer available social cues. Rutter (1987) proposed that "the role of social cues is to produce spontaneous, relaxed conversation" (p. 128). Thus, in audio-only conditions, people are less confident about interrupting. In his view, audio-only situations do not suffer from poor regulation but from increased formality and more social disengagement.

If this view is correct, we might expect the system with the most visual access (PIP) to give rise to the most spontaneous, interactive conversations and the audio-only system to give rise to the fewest, with the LiveWire condition falling somewhere in between. On this basis, and taking into account previous findings, the following predictions were made.

Turn Frequency, Duration, and Distribution. There is little directly comparable prior evidence on which to base predictions for turn frequency, duration, and distribution, so no specific hypotheses were put forth concerning these measures. Two studies found longer and fewer turns in comparing audio-only with face-to-face conversations in two-person conversations (Argyle et al., 1968; Rutter & Stephenson, 1977). However, Rutter et al. (1981) found no difference in the number of floor changes for an audio-only condition compared to a video-mediated condition and indeed no difference even with a face-to-face condition. In addition, Rutter's spontaneity argument makes no specific predictions about floor changes. Although people may be more interruptive when there are more social cues available, this does not necessarily mean the interruptions are intended to take the floor. Nor does it mean that the interruptions are likely to be more successful in taking the floor, with more cues available; having more cues available may also mean that it is easier for a speaker to hold on to the floor.

Simultaneous Speech and Speaker Switching. Taking simultaneous speech as an indication of the degree of interactivity of a conversation, I predicted that the lack of visual information, and hence social cues, in the audio-only condition would cause participants to be more formal and less interactive than when using the PIP system. The expectation was that the audio-only condition would therefore give rise to less simultaneous speech due to fewer interruptions occurring in the audio-only condition than in the PIP condition. The LiveWire condition was expected to fall between these two conditions.

Handovers and Parallel Conversations. In Experiment 2, no advantages in handing over the floor were expected to occur for any of the conditions. Similarly, no occurrences of side or parallel conversations were expected in any of the three conditions because participants do not occupy their own personal space in any of these contexts.

2.3. Method

Participants. Twelve groups of four adults participated in each experiment: 15 women and 33 men in Experiment 1 and 18 women and 30 men

in Experiment 2. With only two exceptions, none of the participants knew each other previously.[7]

Task and Experimental Design. Each group was asked to participate in a set of three informal debates lasting 16 min each. A debate situation was chosen, as such tasks have previously been shown to be more sensitive to the presence or absence of the visual communication channel (e.g., Short et al., 1976). Participants were randomly divided into teams of two, and each team was randomly assigned to either the "pro" or "con" side of the issue. Three different topics were introduced with the help of short newspaper clippings. The topics in Experiment 1 were the right to smoke in public, mandatory drug testing, and censorship in the news. In Experiment 2, the subject of the right to smoke in public was changed to something more topical at the time–whether bilingualism in Canada ought to be legislated. Each group discussed all three topics, one in each condition. Teams remained the same for all three topics, and topics were counterbalanced across conditions.

Both experiments used a simple one-factor, repeated measures design, comparing performance in three conditions. In Experiment 1, the three conditions were same-room, Hydra, and the PIP video system. In Experiment 2, the three conditions were the PIP video system, LiveWire, and an audio-only condition. In each experiment, order of condition was counterbalanced using a Latin square design.

Experimental Conditions and Apparatus. The five conditions are now described. Audio and video records of each conversation were made in a separate control room that also contained specialized speech tracking equipment (also described now). In all conditions, participants wore headset microphones.

1. *Same-room condition.* In this condition, all four participants met in the same room around a table. A video camera was set up in one corner of the room for recording purposes, and the video output was channeled through coaxial cable to a VHS video recorder in the experimental control room. In addition, audio output from the headset microphones was also fed through coaxial cable to the experimental control room. There, it went to both an audio mixer where all four voices were laid down on the audio track of the video cassette and the speech tracking equipment, also located in the control room.

---

7. I would have preferred to have used groups of people who knew each other and were used to working together, but it would have proved difficult to locate 24 such groups. Instead, I opted for recruiting strangers in some attempt at experimental control. Two participants "slipped through the net."

2. *Hydra condition.* Each participant was seated in a separate room containing three Hydra units. In one room, each of the Hydra units was constructed from a Sony Watchman color monitor (8 cm diagonal), a black and white camera from a Radio Shack surveillance unit mounted 4.5 cm below the screen, and a Sony Watchman speaker mounted just below the camera. The unit tilted back and forth for best viewing position. In the other three rooms, simulated Hydra units were used due to budget constraints. In these rooms, three Radio Shack black and white monitors were used (12 cm in diameter) along with two black and white Radio Shack surveillance cameras and one color camera. The color camera was used to feed the prototype Hydra units to take advantage of the color monitors in those units. Each camera was mounted directly on top of each monitor. In addition, each camera/monitor pair was mounted directly on top of a speaker. The speaker carried sound only from the room corresponding to the screen image. In all cases, the Hydra or simulated Hydra units were located 15 cm apart on the desk top and set back 38 cm from the edge of the desk.

As in the same-room condition, video and audio recordings were made of each conversation. In this and all other conditions that follow, the audio output from each microphone was mixed and laid down on the videotape in addition to being sent to the speech tracking equipment. In all video conditions and the audio-only condition, the same four rooms were used and equipment shared and reconfigured where possible.

3. *PIP conditions.* Each participant was outfitted with a color video monitor, video camera, a speaker, and a headset microphone. The camera was mounted on top of each monitor and the speaker was located immediately adjacent to each monitor. The speaker broadcast mixed audio from all four microphones. A video board allowed the display of four composite images. This configuration allowed each participant to see the other three participants as well as an image of themselves. Each participant saw exactly the same configuration of images as the other participants.

4. *LiveWire condition.* The same camera, monitor, and speaker were used as were used in the PIP condition, positioned in the same way. The difference was that, in the LiveWire system, the current speaker was shown on a full screen. Each participant who was not speaking viewed the person who "had the floor" at that moment (i.e., the current speaker), with the exception of the current speaker, who viewed the person who spoke previously. Speech input from the headset microphones was fed into the experimental control room that was used to accomplish the automatic voice switching. However, each participant heard mixed audio from all four microphones during the conversation.

5. *Audio condition.* In this condition, participants spoke to each other through their headset microphones and heard each other through speakers carrying mixed audio from all four microphones. No video images were available to the participants, although they were informed

that video recordings were being made by the experimenters throughout the conversation.

In the Speech Tracking System, the conversion of speech into digital on–off patterns was accomplished by obtaining audio output from each of the four participants using unidirectional, dynamic, headset microphones. Each microphone output controlled its own externally keyed audio noise gate. When a participant spoke louder than a preset threshold, the corresponding audio noise gate would open, allowing a fixed pitch generated by a Yamaha TX802 synthesizer to pass through. When a participant fell silent, the gate would close, cutting off the pitch. Each of the output signals from the four noise gates was fed into four input channels of an IVL Pitchrider 7000 Mark II pitch tracking device. The pitch tracker converted the pitch on–off signals into digital on–off signals and sent them, via a MIDI connection, to a Macintosh II computer. These on–off events were stored in the computer, and each event stamped with a time code. This time code was simultaneously laid down on the videotape so speaker events could be synchronized later when playing back the videotape.

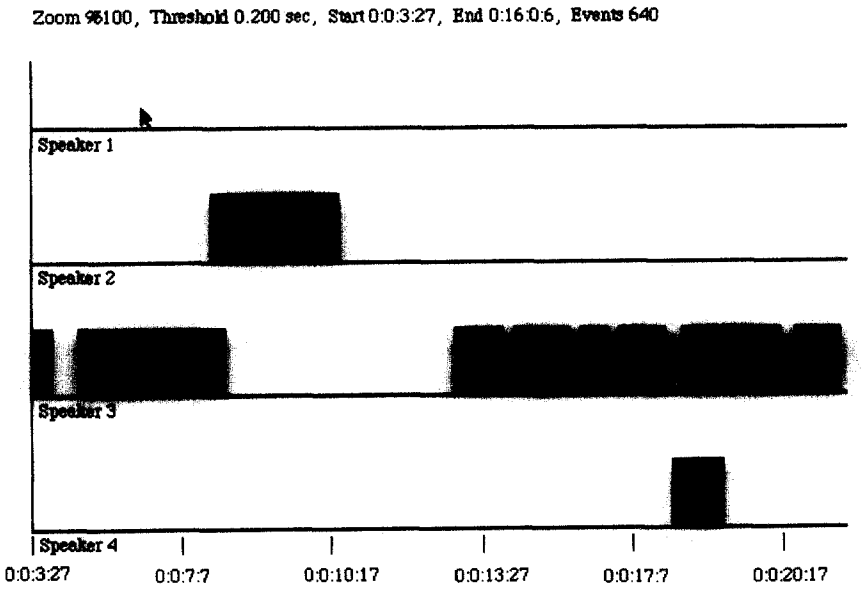
Procedure. The procedure was the same for both experiments. On arrival, participants on the same team were introduced to each other and given approximately 15 min to get acquainted, while completing the experimental consent forms. Following this, they were introduced to the members of the other team and were instructed to read the first topic for debate. Participants were then placed in separate rooms (or sometimes in the same room in the case of Experiment 1), and were instructed in the wearing of the headset microphones.

In all conditions, a similar procedure was used. Participants discussed the prescribed topic for 16 min. At the end of the conversation, they were asked to complete a questionnaire about the conversation they had just experienced, independently of each other. Each questionnaire consisted of nine statements about different aspects of the conversation (listed in Figure 10). Participants were asked to indicate their answers on a 7-point, Likert-type scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

While participants were completing the questionnaire, the technology was reconfigured for the next condition. Participants were then asked to move to a new room if moving between a video condition and the same-room condition. The new experimental condition was then introduced and explained.

At the end of the experiment, a group discussion was held specifically to discuss the three different conditions and to solicit participants' opinions about the videoconferencing systems they had used. This was done by asking participants to rank the three conditions in order of preference and to explain their rationale for the ordering. Notes were taken during these interviews, and the discussion was also videotaped.

Figure 5. **Approximately the first 20 sec of a conversation shown in the form of a speech time line, with time (in sec) along the horizontal axis and each of the four speakers represented on the vertical axis. Black rectangles correspond to "talkspurts" by each participant and are laid out in time showing when the on-off patterns of speech took place with respect to each other.**



2.4. Results

Analysis of Speech Data. The time lines for each 16 min conversation were checked for accuracy against the videotape data, edited where necessary, and coded using specialized software designed for this purpose. A sample time line is shown in Figure 5. Despite the impressive accuracy of the speech tracking system, some sporadic crosstalk caused by the equipment did occur and had to be deleted. In addition, 200 msec pauses within a speaker's utterance were automatically filled in to account for stop consonants (a procedure also used by Brady, 1968). Laughter and backchannel responses were coded so as to differentiate these data from speaker turns or attempts to take turns. (*Backchannel responses* are vocalizations such as "mmm-hmm," often used to show attentiveness, which do not constitute turns or attempts to take turns [Duncan, 1972]).

Definitions. The data were analyzed using definitions taken from both Jaffe and Feldstein (1970) and Dabbs and Ruback (1984, 1987) and slightly modified. Dabbs and Ruback's (1984, 1987) scheme is an exten-

sion of that of Jaffe and Feldstein (1970) to better account for groups larger than dyads. The following definitions were used:

1. *Turn.* A turn consists of the sequence of talkspurts and pauses by a speaker who has the floor. A speaker gains the floor when he or she begins speaking to the exclusion of everyone else and when he or she is not interrupted by anyone else for at least 1.5 sec.[8] The duration of a turn begins with the first unilateral sound and ends when another individual turn or *group turn* begins (see later). Note that turns therefore include periods of mutual silence at the end of utterances when no one else has yet taken the floor.
2. *Group turn.* Using Dabbs and Ruback's (1984) definition: "A group turn begins the moment an individual turn taker has fallen silent and two or more others are speaking together; the group turn ends the moment any individual is again speaking alone" (p. 519). Dabbs and Ruback (1984) proposed the group turn to cover instances where individual turn takers are effectively "drowned out" by the group.
3. *Speaker switch.* A speaker switch occurs whenever one person or group loses the floor, and another person or group gains it.
4. *Switch time.* Switch time consists of *switch pauses* and *switch overlaps.* A switch pause is a period of mutual silence bounded by different turn takers (individuals or groups). Unlike existing definitions, I also include as a related measure the concept of switch overlap. A switch overlap is a period of simultaneous speech immediately before and leading to the person who utters it taking a turn. The two measures can be conceptualized as a single continuous metric that measures the relationship between one person ending a turn and another starting one. A negative switch time is thus a switch overlap, and a positive switch time is a switch pause.
5. *Simultaneous speech.* Simultaneous speech is speech by speakers who do not have the floor (i.e., when two, three, or four people speak simultaneously). There are various different situations in which simultaneous speech occurs. Some simultaneous speech consists of switch overlaps (previously described), which can be distinguished in that they lead to a speaker switch. However, simultaneous speech also results when speakers simultaneously begin to speak, when they unsuccessfully attempt to take the floor, and so on.

Turn Analysis. The means and standard deviations for the number and duration of individual turns, number of group turns, and turn distribu-

8. Without this criterion, even the shortest unilateral sound would be designated as a turn. The criterion of 1.5 sec was chosen because this is estimated to be the mean duration of a phonemic clause and there is evidence for the phonemic clause as a basic unit in the encoding and decoding of speech (as discussed in Jaffe & Feldstein, 1970).

Figure 6. **Means and standard deviations for the number, duration, and distribution of speaker turns per 16-min session.**

	Experiment 1			Experiment 2		
Variable	SR[a]	H[b]	P1[c]	P2[c]	LW[d]	AO[e]
Number of turns per session	62.60 (17.40)	68.70 (24.30)	64.10 (19.50)	66.00 (27.80)	55.60 (20.20)	59.70 (21.50)
Turn duration (sec)	15.90 (3.70)	16.60 (10.30)	16.50 (6.70)	17.20 (8.00)	19.40 (7.10)	19.20 (10.90)
Number of group turns per session	3.80 (3.80)	3.80 (6.90)	4.60 (5.00)	2.60 (2.50)	1.70 (2.40)	1.80 (2.60)
Distribution of turns (*H* value)	1.83 (0.10)	1.83 (0.17)	1.82 (0.17)	1.88 (0.17)	1.87 (0.17)	1.88 (0.11)

Note. Standard deviations in parentheses. No significant differences in results for Experiment 1 or Experiment 2.
[a]Same-room condition. [b]Hydra condition. [c]PIP condition. [d]LiveWire condition. [e]Audio-only condition.

tion are shown in Figure 6. Turn distribution among speakers was calculated after Dabbs and Ruback (1984), who used Shannon and Weaver's (1949) equation for calculating information (in Information Theory terms). This equation–defining H, or amount of information–is essentially a way of calculating the average amount of uncertainty about who has the floor at any given time. H is defined by:

$$H = -\Sigma\, pi\, log(pi)$$

where pi is the proportion of the total number of turns in the session taken by person i.

Analyses of variance (ANOVAs) were carried out, evaluated throughout this article at $\alpha = .05$ level. For both experiments, these ANOVAs found no difference across conditions for any of these turn measures. In addition, no significant practice or order effects were found on any of these measures.

Simultaneous Speech Analysis. Figure 7 presents the data summary for simultaneous speech and switching measures. ANOVAs[9] and, where appropriate, post hoc analyses of the differences between means were carried out using the Peritz procedure (Toothaker, 1991).

9. The *F* values reported here are, in some cases, slightly different from those reported in Sellen (1992). This is because, unlike the analysis reported in Sellen (1992), both practice and order effects were extracted from the residual mean square error and tested separately. The overall conclusions, however, remain unaffected by these differences.

Figure 7. **Means and standard deviations for the simultaneous speech and switching measures.**

	Experiment 1				Experiment 2			
Variable	SR[a]	H[b]	P1[c]	Results	P2[c]	LW[d]	AO[e]	Results[f]
Percentage of simultaneous speech	9.70 (7.40)	5.40 (7.10)	7.10 (6.90)	SR>H=P1	3.10 (2.40)	2.00 (1.50)	3.00 (2.60)	*
Switch time (sec)	−0.46 (0.66)	0.25 (0.67)	0.04 (0.79)	SR<H=P1	0.32 (0.33)	0.77 (0.68)	0.48 (0.62)	*
Percentage of overlaps in speaker switches	54.10 (18.90)	43.50 (21.10)	46.30 (24.10)	SR>H=P1	38.50 (10.90)	28.90 (10.50)	38.30 (12.50)	LW<P2=AO

Note. Standard deviations in parentheses.
[a]Same-room condition. [b]Hydra condition. [c]PIP condition. [d]LiveWire condition. [e]Audio-only condition. [f]Asterisk = no significant differences.

These analyses found:

• Percentage of simultaneous speech was significantly different across conditions in Experiment 1, $F(2, 20) = 4.61$, $p < .03$. Planned comparisons showed same-room conversations to contain more simultaneous speech than video conditions, $F(1, 18) = 8.09$, $p < .01$, but showed no difference between video conditions. In Experiment 2, there was no difference across conditions. Percent of simultaneous speech refers to the proportion of time during which two, three, or four people were speaking simultaneously.

• Switch time was significantly different across conditions in Experiment 1, $F(2, 20) = 14.48$, $p < .001$. Furthermore, the same-room condition gave rise to a mean negative switch time value, and the video conditions gave rise to a mean positive switch time value, or switch pause. The difference between same-room and both video conditions was significant ($p < .05$, Peritz), but no difference between video conditions was found. In Experiment 2, no difference was found among conditions. Switch time is an average of switch pauses (positive values) and switch overlaps (negative values).

• Percent of speaker switches consisting of switch overlaps (as opposed to switch pauses) did differ across conditions in both experiments, $F(2, 20) = 4.03$, $p < .04$, Experiment 1; $F(2, 20) = 4.31$, $p < .04$, Experiment 2. In Experiment 1, more speaker switches consisted of overlaps in the same-room condition than the video conditions ($p < .05$, Peritz), with no difference between video conditions. In Experiment 2, the LiveWire condition gave rise to a lower percentage of switch overlaps than both the PIP and audio-only conditions ($p < .05$, Peritz). There were no differences between

the audio-only and PIP conditions. Percent of overlaps in speaker switches is the percentage of speaker switches that occur with a negative rather than a positive switch time.

In Experiment 1, there were no significant practice or order effects found, $F(2, 18) = 2.78$, $p < .09$, although percentage of simultaneous speech did increase from 5.6% in Session 1 to 7.9% in Session 2 to 8.7% in Session 3. In Experiment 2, a significant practice effect was found for percentage of simultaneous speech, increasing across conditions from 1.7% in Session 1 to 2.6% in Session 2 to 3.8% in the Session 3, $F(2, 12) = 5.86$, $p < .02$. There were no significant practice effects for either of the switching measures.

Simultaneous Speech Categorization. Simultaneous speech occurs for a variety of reasons. To better understand the source of the differences among conditions, every instance of simultaneous speech except back-channels[10] was classified into one of four categories using a taxonomy constructed by O'Conaill et al. (1993). The following categories were used:

1. *Interruptions.* Cases where another speaker attempted to take control of the floor when there was no indication that the current speaker was about to relinquish the floor.
2. *Projections/completions.* Cases where the next speaker appeared to predict that the current speaker was about to finish, and thus "jumped in" early. Sometimes, it was clear that the person taking the floor thought the current speaker was finished (because of a long pause or because the speaker asked a rhetorical question). Other times, it appeared that the current speaker had finished his or her main point, causing the next speaker to start his or her turn early. This category also includes instances when the next speaker helped the current speaker to finish an utterance.
3. *Simultaneous starts.* Instances where two or more speakers attempted to take the floor concurrently, usually some time after the previous speaker had finished. This included cases where the original speaker tried to resume speaking after a pause.
4. *Floorholding.* Cases where the next speaker tried to take the floor while the current speaker attempted to hold it by producing utterances that did not contain any information (e.g., by repeating a word, such as "so, so, ... ").

---

10. ANOVA was performed on the frequency of backchannels as coded in the initial pass through the speech data. No differences across conditions were found in either experiment. However, these results must be treated with caution because many backchannel responses, such as "mm-hmm" are not fully voiced and thus are difficult to detect with headset microphones.

Figure 8. **Mean frequency and standard deviation per 16-min session for four kinds of simultaneous speech.**

	Experiment 1				Experiment 2			
Variable	SR[a]	H[b]	P1[c]	Results	P2[c]	LW[d]	AO[e]	Results
Interruptions	31.0 (14.6)	20.4 (17.0)	24.3 (20.2)	SR > H SR = P1 H = P1	13.5 (8.6)	11.80 (8.40)	13.60 (9.00)	*
Projections/ completions	15.4 (5.8)	12.6 (9.3)	13.6 (8.1)	*	10.3 (8.3)	6.80 (4.20)	9.30 (5.30)	*
Simultaneous starts	9.6 (5.3)	7.8 (6.7)	9.2 (6.6)	*	6.5 (5.6)	4.20 (3.40)	7.40 (5.30)	*
Floorholding	0.6 (0.7)	0.1 (0.3)	0.0 (0.0)	Not tested	0.0 (0.0)	0.08 (0.29)	0.08 (0.29)	Not tested

Note. Standard deviations in parentheses. Asterisk = no significant differences. [a]Same-room condition. [b]Hydra condition. [c]PIP condition. [d]LiveWire condition. [e]Audio-only condition.

The results are shown in Figure 8.

A research assistant who was blind to the experimental predictions carried out the classification. However, the reliability of this categorization was assessed by comparing it to classification by an independent judge of six experimental sessions selected at random, one from each condition, in both experiments. Kappa statistics were then carried out to measure the extent to which there was consensus between the two judges in assigning each instance of simultaneous speech to one of the four categories. The resulting statistics for Experiment 1 were .74 for same-room, .82 for Hydra, and .78 for PIP. For Experiment 2, they were .77 for PIP, .84 for LiveWire, and .72 for audio-only. These measures of kappa were converted into z statistics and tested for significance. All were significant at the .05 level.

ANOVAs showed that the only significant difference occurred in Experiment 1, where the number of interruptions differed across conditions,[11] $F(2, 20) = 4.23$, $p < .03$. Multiple comparisons of the means showed a significant difference between the same-room and Hydra conditions ($p < .05$, Peritz) with no difference between the same-room and the PIP conditions or between video conditions.

This more in-depth analysis of simultaneous speech suggests that differences between the same-room conditions and the two video conditions are almost entirely due to the occurrence of more interruptions in the same-room condition. This difference is most pronounced when comparing the same-room condition to the Hydra condition.

11. Because of the low total frequency of the occurrence of floorholding, this measure was not subjected to statistical test.

Handover Categorization. A final analysis was carried out to discover whether the process of handing over the floor to another speaker was qualitatively different across conditions. For example, it is of interest to know whether the ability to use selective gaze toward other participants reduces the need for more explicit or formal ways of handing over the floor. Conditions where this ability is presumed to be present (e.g., same-room and Hydra) are thought to provide implicit cues as to who is to be the next speaker.

The videotapes were analyzed to look for instances of explicit handovers and cases when verbal devices were used by the current speaker either to indicate who the next speaker should be or simply that the turn was ended. A modification of categories used by O'Conaill et al. (1993) was again used, resulting in three categories:

1. *Naming.* When the current speaker indicated the next speaker by naming him or her.
2. *Tagging.* When the current speaker indicated that another speaker could take over by attaching a stereotyped question or phrase to the end of an utterance (e.g., "isn't it?", "you know").
3. *Directing.* When the current speaker specifically directed someone else to be the next speaker without using his or her name but by using statements such as "You go ahead" or "What do you think?"

The mean frequencies and standard deviations of formal handovers falling into these categories are shown in Figure 9.

According to ANOVA, the total number of formal handovers in Experiment 1 differed across conditions, $F(2, 20) = 10.11$, $p < .001$. Planned comparisons failed to find a difference between the PIP and Hydra systems, but found that the same-room condition had fewer handovers than both Hydra, $F(1, 18) = 18.45$, $p < .001$, and PIP, $F(1, 18) = 5.65$, $p < .03$. Analysis by category found the frequency of naming was higher in the two video conditions than in the same-room condition, with no difference between the PIP and Hydra conditions ($p < .05$, Peritz). It also found that the frequency of tagging was different across conditions–higher in the Hydra condition than in the same-room condition ($p < .05$, Peritz). No difference was found between the two video conditions or between the same-room and PIP conditions (though this difference was close to significance).

In Experiment 2, ANOVAs revealed that there were no significant differences across conditions in the total number of formal handovers or for any of the individual handover categories.

Parallel Conversations. In Experiment 1, analysis of the videotape data revealed that instances of side or parallel conversations occurred five

Figure 9. **Mean number and standard deviation of occurrences per 16-min session of three kinds of handovers.**

	Experiment 1				Experiment 2			
Variable	SR[a]	H[b]	P1[c]	Results	P2[c]	LW[d]	AO[e]	Results
Naming	0.10	1.30	1.30	SR < H = P1	1.30	0.70	0.80	*
	(0.30)	(1.50)	(1.90)		(2.30)	(0.80)	(1.00)	
Tagging	2.27	4.42	3.08	SR < H	2.80	2.60	3.40	*
	(1.29)	(3.09)	(1.68)	H = P1	(2.20)	(2.10)	(3.10)	
				SR = P1				
Directing	0.36	1.17	0.58	*	2.10	1.40	2.60	*
	(0.64)	(1.34)	(0.90)		(0.90)	(1.10)	(2.30)	
Total	2.73	6.83	5.00	SR < H = P1	5.70	4.30	6.80	*
	(1.54)	(3.13)	(2.89)		(4.20)	(3.20)	(5.20)	

Note. Standard deviations in parentheses. Asterisk = no significant differences. [a]Same-room condition. [b]Hydra condition. [c]PIP condition. [d]LiveWire condition. [e]Audio-only condition.

times in the same-room condition, six times in the Hydra condition, and never in the PIP condition. In Experiment 2, no parallel or side conversations were seen to occur in any of the conditions.

Questionnaire Data. Scores from the questionnaires averaged across all 48 participants in each experiment are shown in Figure 10. ANOVAs were carried out, and when significant, post hoc tests were used to assess differences among means (Peritz procedure). These results are also shown in Figure 10.

There were four significant results in Experiment 1. Participants assigned the same-room meetings significantly higher ratings than the video meetings in: (a) allowing them to take control of the conversation (Question 2), $F(2, 94) = 6.45$, $p < .002$; (b) interactivity (Question 5), $F(2, 94) = 5.35$, $p < .006$; (c) allowing them to selectively attend to one person at a time (Question 7), $F(2, 94) = 10.71$, $p < .001$; and (d) knowing when others were listening or attending to them (Question 8), $F(2, 94) = 29.65$, $p < .001$. No differences were found between video conditions, with the exception of the fourth finding, where Hydra was rated significantly better than the PIP system.

There were six statistically significant differences in Experiment 2. Participants rated the PIP condition as having fewer inappropriate interruptions (Question 3), $F(2, 94) = 6.50$, $p < .002$; as being less unnatural than the LiveWire and audio-only conditions (Question 4), $F(2, 94) = 4.20$, $p < .02$; and as best for knowing when others were listening and attending to them (Question 8), $F(2, 94) = 27.40$, $p < .001$. There were no differences between LiveWire and audio-only on these questions. Participants also

Figure 10. **Means and standard deviations for the nine questions in the questionnaires on a 7-point, Likert-type scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).**

	Experiment 1				Experiment 2			
Question	SR[a]	H[b]	P1[c]	Results	P2[c]	LW[d]	AO[e]	Results
1. I was able to talk and express myself freely.	6.1 (0.9)	5.7 (1.4)	5.8 (1.4)	*	5.8 (1.2)	5.7 (1.3)	5.5 (1.5)	*
2. I was able to take control of the conversation when I wanted to.	6.0 (1.0)	5.5 (1.4)	5.4 (1.5)	SR > H = P1	5.5 (1.1)	5.4 (1.2)	5.1 (1.6)	*
3. There were too many inappropriate interruptions.	2.4 (1.5)	2.2 (1.4)	2.4 (1.4)	*	1.9 (1.1)	2.9 (1.6)	2.6 (1.5)	P2 < LW = AO
4. This was an unnatural conver sation.	2.7 (1.7)	3.0 (1.6)	3.1 (1.7)	*	3.0 (1.7)	3.6 (1.7)	3.8 (1.8)	P2 < LW = AO
5. The conversation seemed highly interactive.	5.9 (1.3)	5.3 (1.2)	5.4 (1.3)	SR > H = P1	5.4 (1.4)	5.0 (1.4)	4.6 (1.7)	P2 > AO LW = AO LW = P2
6. There were many unnatural and uncomfortable pauses.	2.5 (1.7)	2.6 (1.4)	2.3 (1.1)	*	3.0 (1.7)	3.3 (1.6)	2.9 (1.5)	*
7. I could selectively attend to one person at a time.	6.1 (1.1)	5.3 (1.8)	4.8 (1.8)	SR > H = P1	5.7 (1.4)	5.2 (1.9)	4.6 (1.8)	P2 > AO LW = AO LW = P2
8. I knew when people were listening or paying attention to me.	6.3 (0.9)	5.3 (1.5)	4.3 (1.9)	SR > H > P1	5.8 (1.3)	3.9 (1.8)	3.8 (1.7)	P2 > LW = AO
9. I found it difficult to keep track of the conversation.	2.0 (1.5)	2.2 (1.2)	2.6 (1.7)	*	2.2 (1.6)	2.4 (1.3)	2.9 (1.6)	AO > P2 LW = P2 LW = AO

Note. Standard deviations in parentheses. Asterisk = no significant differences.
[a]Same-room condition. [b]Hydra condition. [c]PIP condition. [d]LiveWire condition. [e]Audio-only condition.

rated the PIP condition as the most interactive of the three (Question 5), $F(2, 94) = 3.22$, $p < .05$, and as the best for selectively listening and attending to others (Question 7), $F(2, 94) = 5.81$, $p < .004$. But in both cases, the ratings were only significantly higher than the audio-only condition: the LiveWire condition fell in between these two conditions and was not significantly different from either. Finally, participants rated the audio-only condition as being significantly more difficult to keep track of than the PIP condition (Question 9), $F(2, 94) = 3.31$, $p < .05$. There were no other significant differences among the mean pairs.

3. DISCUSSION

One of the main goals of Experiment 1 was to examine the effects of video mediation on conversation by comparison with same-room conver-

sations. A number of specific predictions were made with respect to a range of speech measures. The results of this analysis are discussed first. The goal of Experiment 2 was to assess the impact of visual access by comparing videoconferencing situations with audio-only conversations in terms of these same speech measures. This comparison is discussed next. Finally, these objective measures are contrasted with the results of discussions with the participants and their responses in the questionnaires.

3.1. Same-Room Versus Video-Mediated Conversations (Experiment 1)

To a large extent, the results of Experiment 1 ran counter to expectation. In terms of the speech measures, there were fewer differences between same-room and video-mediated conversations than originally predicted and a complete absence of differences between the two different video systems. No problems in regulation of conversation in the video conditions and no significant effects of selective gaze were evident in these measures. However, as is discussed later, in-depth analysis of the videotaped data as well as participants' comments revealed differences between the PIP and Hydra systems that these measures did not capture.

Turn Frequency, Duration, and Distribution. Mediating conversations with video technology had no discernible effects on the number of turns taken per session, the average length of those turns, or the distribution of turns among speakers, in comparison to conversations held in the same room.

The results for number and length of turns were unexpected, especially in light of Cohen's (1982) previous research, which found longer and fewer turns in a video-mediated condition. However, Cohen introduced a 705 msec audio transmission delay to simulate round-trip satellite conditions—a fact that very likely accounts for the disruption in turn-taking. That poor audio may be the critical factor is backed up by more recent experimental evidence (O'Conaill et al., 1993) that also found fewer, longer turns in a video-mediated condition with lagging, half-duplex audio.

The failure to find a difference in turn distribution was also counter to prediction. The expectation was that difficulties in floor control in the video conditions (especially the PIP condition) would lead to more unequally distributed turns over the speakers in comparison with the same-room condition. This suggests that there was no overwhelming difficulty in managing the floor in either of the video conditions. Rather, within-group characteristics such as the amount of talking by each participant remained relatively stable across conditions. Dominant speakers tended to dominate in all conditions, and quiet people tended to remain quiet, regardless of the condition. These within-group similarities across conditions can be contrasted with pronounced between-group variability on virtually all measures.

Simultaneous Speech and Floor Control. Where differences between conditions do emerge, however, is in the amount of simultaneous speech that occurred and in the time between switching speakers. As predicted, a higher percentage of time was occupied by simultaneous speech in the same-room condition relative to the two video conditions. This result is in line with the majority of previous work involving both audio-only and video-mediated conditions. That this difference is at least partly due to overlapping speech at the end of speaker turns is indicated by the analysis of switch time. The average switch time was negative in the same-room condition (as compared to a positive mean switch time in the video conditions), and over half of all speaker switches occurred with an overlap, significantly more often than in the two video conditions.

A more informative analysis comes from categorization of occurrences of simultaneous speech. Both interruptions and simultaneous starts might be assumed, a priori, to be indications that there are problems in regulating a conversation. In other words, that people interrupt each other or try to take the floor inappropriately might be symptomatic of problems in timing bids for the floor. Presumably, the more turn-taking cues there are, the more this kind of problem could be avoided, and one would expect that being physically present with others would offer the widest range of cues of this sort.

That people were actually being more interruptive in the same-room condition than in the video conditions and that simultaneous starts were no different across conditions calls into question the assumption that these measures indicate problems in floor control. Rather it suggests that what they may indicate, at least in this experiment, is the degree of interactivity and spontaneity of the conversation, rather than its synchronization. More simultaneous speech may be due to participants who feel more engaged in the conversation and who are more willing to attempt to take the floor. This is in line with many participants' responses in the questionnaire as well as their spontaneous comments. Participants rated the same-room condition as being more interactive than the video conditions. Many reported feeling "distanced" by the video systems and less a part of the conversation. Perhaps they felt that bids for floor control would be less effective in video-mediated conversations. This is in line with Rutter's (1987) suggestion that the attenuation of cues through mediation leads to increasingly less spontaneous, more formal, and more socially distant discussions.

If participants were more willing to bid for the floor in the same-room condition, why did this not result in more speaker switches in this condition? One possible explanation is that the increased interactivity and engagement in the conversation also meant that speakers were more tenacious about holding the floor. Thus, many of these attempted bids

simply failed to succeed. In the video conditions, speakers may have been more polite, allowing opportunities for people to take over without the need for interruption or yielding the floor when interruptions did occur.

Finally, the finding of increasing amounts of simultaneous speech over the course of the experiment (significant in Experiment 2) could be interpreted as further evidence that this measure reflects the degree of formality of the conversations. Presumably, as participants became more familiar with each other, they also became less formal toward each other–this fact becoming manifest in the amount of simultaneous speech that took place. It should also be added that there was no condition by practice interaction. This is fortunate from a statistical perspective, as it means that the effect does not confound interpretation of the means for the different conditions. But the other implication of the fact that the effects of condition and practice are independent is that it offers hope that familiarity among participants could eventually override any distancing effects of the mediating technology. However, this is an issue outside the realm of this experimental design.

Formal Handovers. Participants were more likely to resort to formal ways of handing over the floor in both video conditions. In particular, naming the next speaker or tagging the end of a turn to indicate it had finished was more frequent in the video conditions than in same-room conversations. Participants therefore must have felt the need to be more careful and unambiguous about when their turns were ended and who was to speak next.

This effect could not have been due to the absence of selective gaze and head-turning cues, because, as is discussed later, these were used to some effect in the Hydra system. Rather, it may be that participants, feeling somewhat disconnected from the situation, felt that these kinds of gestures would not carry the same impact, thus leading them to compensate in more explicit ways. Indeed, Heath and Luff (1991) presented evidence that supports this view, showing that video-mediation reduces the effectiveness of gesture, orientation, and gaze for their recipients. Therefore, this finding does not necessarily speak to the failure of the Hydra system, but instead suggests that the need for explicit handovers is due to the nature of video-mediated interaction in general; it suggests that implicit cues normally used when people are copresent are perceived to be relatively powerless in the remote situation, resulting in the requirement for more explicit devices for interaction.

3.2. Video-Mediated Versus Audio-Only Conversations (Experiment 2)

One might conclude from Experiment 1 that the increased formality of video-mediated conversations is due to the fact that videoconferencing

systems attenuate the visual cues available in a conversation. The comparison of video-mediated conversations with the audio-only condition in Experiment 2 provides a test of whether the amount of available visual information is the variable to which these effects can be attributed.

The results offer no support for the hypothesis that it is visual information per se that affects these speech measures. The presence or absence of a visual channel had no effect in terms of the number, duration, or distribution of turns in a conversation. One inference is that, as expected, a lack of visual cues caused no significant problems in the regulation and synchronization of conversation. This does not necessarily mean that visual cues did not play a role in the regulation of conversation–the cues carried linguistically or paralinguistically in this experimental situation may have been sufficient for managing the flow of the conversation.

The more important point, though, is that audio-only conversations showed no signs of being more formal and less interactive than when the video channel was present: Lack of the visual channel did not cause a reduction in the percentage of simultaneous speech (as predicted), and it did not affect the switch time in speaker transitions. With regard to the percentage of overlaps in speaker switches, the audio-only condition was no different from the PIP condition. In fact, this measure was lowest in the LiveWire condition, (a result no doubt due to the peculiarities of its design, which will be discussed later). Breaking down the simultaneous speech into different categories also confirmed that there were no differences between audio-only and video-mediated conversations. Participants in the audio-only condition were just as likely to try to interrupt one another or otherwise overlap with the talk of others than when video was present.

In terms of the number of explicit or formal handovers present in the three conditions, again, there were no significant differences, even when subdivided into categories. The absence of visual cues did not increase the need to use explicit methods of regulating turn-taking, offering further evidence that the audio-only conversations were no more formal or remote than the video conversations.

3.3. Participants' Opinions and Preferences

The analysis of the speech measures for the two experiments combined[12] indicates that the only differences that existed were between

12. Note that the numbers cannot be compared absolutely, as two different participant populations were involved. Indeed, the means for the PIP condition across the two experiments were very different, largely due to the fact that the 12 groups of participants in Experiment 1 were, on average, much more talkative and interactive than the 12 groups in Experiment 2. Nonetheless, the results (if not the numbers) can be compared because the PIP condition was identical in each of the two experiments.

unmediated (same-room) and mediated conversations. In these two experiments, the *nature* of the mediating technology, whether it was a videoconferencing facility or an audio-only system, had no significant effects on these kinds of measures. This suggests that, with respect to this set of dependent variables, if a conversation takes place via a technological system, the way in which the visual information is presented, or even whether visual information is present, makes no difference to the conversational process.

Although the speech measures failed to differentiate between the various systems for mediating remote conversations, the questionnaire and interview data, as well as a closer analysis of the video data, indicate that participants were affected by differences between systems, and most had strong opinions about which system they preferred. For example, according to the questionnaire, participants in Experiment 1 felt it was more difficult to take control of the conversation in the video conditions than in the same-room condition. Participants' perceptions were, therefore, not always paralleled by the speech measures.

One possibility is that the speech measures simply failed to capture important behavioral phenomena–that the measures were not the relevant or interesting ones. Another possibility is that such perceptions were not manifested in, or derived from, any outward behavior (i.e., perhaps perceived problems in taking control were independent of the existence of any real problems occurring). It is difficult to say why the subjective and objective data were sometimes inconsistent, but whether participants' perceptions were grounded in outward behavior, those perceptions and opinions are behavioral phenomena in themselves and are thus important to understand in relation to the design of the various systems.

The next two sections examine the responses to the different systems in more detail. These observations were made on the basis of notes initially taken during the interviews, which were completed after the experiment by referring back to videotapes of the interviews. These notes were used to summarize participants' comments and also to record participants' rankings of each of the three systems they had experienced, in order of preference.

PIP Versus Hydra

Because the speech measures did not reveal any problems in the regulation of conversation between same-room and video-mediated conversations, it is perhaps not surprising that these measures also failed to show the predicted effects of selective gaze and listening in the Hydra system when compared with the PIP system. Despite the fact that these measures did not distinguish between PIP and Hydra, during the final interview, the majority (two thirds) of the participants preferred the Hydra to the PIP system.

For the minority that preferred the PIP system, reasons given include the interesting comment by one participant that seeing herself on the same screen as the others made her feel more "part of the group." Others cited the presence of their own image on the screen as a desirable feature. These participants said they liked to see themselves to know how they were seen by others, even though this could sometimes be distracting.

However, for participants who preferred Hydra (i.e., most of them), one of the most frequently reported reasons was the fact that they could selectively attend to people, and could tell when people were attending to them. The questionnaire data support this in that participants rated it better than the PIP system in knowing when people were listening or attending to them. Also, a frequent comment was that participants liked the multiple sources of audio in the Hydra system and that this helped them keep track of one thread of the conversation when people talked simultaneously. These comments suggest that Hydra was successful in facilitating selective listening and selective gaze, in line with the original intent behind its design.

Further evidence for the effectiveness of Hydra's selective gaze cues comes from a closer analysis of the videotape data that showed that Hydra was indeed successful in affording side and parallel conversations. In Experiment 1, instances of side or parallel conversations were quite infrequent in all conditions, yet occurred about an equal number of times in the same-room (five times) and the Hydra condition (six times), demonstrating that the Hydra system is indeed capable of supporting this kind of behavior. With the Hydra system, participants exhibited behavior much like they did while conversing in the same room, in that they tended to lean toward one of the video units, securing that person's attention in order to carry on a conversation separate from the one in progress. Asides also tended to be made this way, as they would when people are physically copresent.

That such behaviors never occurred in the PIP system does not constitute proof that they are not possible, but strongly suggests that they are less likely to occur when selective gaze is not supported. Analysis of the videotapes also revealed at least three instances of conversational breakdown occurring in the PIP condition due to an ambiguity about who was being addressed in a conversation. All of these examples involved the current speaker using the word *you* in a way that the other participants found ambiguous. For example, in one case the current speaker said, "The point you were trying to make was ..." The other participants, being puzzled as to whom he was addressing, lost the thread of their argument as they tried to establish the identity of "you." Although this kind of breakdown was infrequent, it was never observed to occur in the Hydra system (nor indeed in the same-room situation). Further indication that it is symptomatic of systems without selective gaze comes from analysis of the second experiment, which found that such breakdowns occurred twice

in the audio-only condition, once in the LiveWire condition, and twice in the PIP condition.

Again, though, these findings contrast with the speech measures: though participants appeared to be able to use selective gaze in the Hydra system, they used just as many explicit handovers in the Hydra condition as in the PIP condition. Why the selective gaze and head-turning cues had no effect on such measures is an interesting issue. Head-turning and directional gaze could be readily observed in the Hydra conversations, but speakers may have faced difficulties in knowing how their gestures were received. Indeed, many participants commented that in the Hydra condition they wanted a mirror to see how they were framed from the point of view of others. Without this feedback, participants might perceive that they have less control and that their gestures have less effect. Thus, although head-turning and gaze cues conferred some of the predicted advantages in terms of the behaviors they afforded, it appears that disengagement or disempowerment caused by the technology may again provide an explanation for this set of results.

PIP Versus Audio-Only Versus LiveWire

In Experiment 2, as in Experiment 1, participants were interviewed and asked to rank the three conditions in order, putting the system they liked the best first and the system they liked the least last. The majority (73%) of the participants said they preferred the PIP system, 17% preferred LiveWire and only 10% liked the audio-only condition best.

PIP. Of the participants preferring the PIP system, most of them were of the opinion that video was valuable for gauging other people's reactions through facial expressions and body language so they could know who was paying attention and how their arguments were being received. This was reflected in the questionnaire data, with participants rating the PIP condition as best for knowing when people were listening and attending to them. It was also rated as the best for selectively listening and attending to others, although this was only in comparison with the audio-only condition. These people also mentioned that they liked having choice and control over at whom they were looking. Many commented that they felt this was the most natural, relaxed, and interactive situation of the three. Again, this was confirmed in the questionnaire data, with participants rating the PIP condition, on average, as the most interactive and the most natural of the three conditions.

On the negative side, some people who preferred the other two systems commented that they felt self-conscious and distracted seeing themselves on the screen. On the other hand, many participants said they found it useful to know how they were framed and to confirm how others were

seeing them. Thus, it appears that even though having to view oneself in conversation is unnatural, this feature was useful in helping to overcome the problem in video-mediated conduct of knowing what others are perceiving.

Audio Only. Of the people who preferred the audio-only condition, there were two main reasons given. One was that they were used to this method of conversing, and so felt comfortable with it. The second reason given was that they liked the anonymity and felt less self-conscious than when the video was present.

Participants who liked the audio-only condition the least complained that without seeing the other participants, it was sometimes difficult to follow the conversation or to distinguish which person was talking. After all, none of the participants knew each other prior to this experiment, and, for a third of the groups, this condition was experienced first. The questionnaire data confirm that participants felt that audio-only conversations were harder to keep track of than PIP conversations. Another common complaint was that the audio-only condition made them feel "dislocated" from the conversation–it was as if they were listening to a radio, the members of one group said. Finally, some participants said that they felt the audio-only conversation was less coordinated and that it was harder to know who was about to speak. This is supported by the questionnaire data, which showed that this condition was ranked as having more inappropriate interruptions than the PIP condition.

LiveWire. For those who liked the LiveWire condition the best, the reason most frequently given was that they felt it helped them to concentrate on the thread of the argument. One person said that he felt the LiveWire condition was the best for focussed conversations, but not for when the conversation "drifted around," or when there was some sort of group response. Two people said they liked the LiveWire conversations because they thought they were the most polite–people were more likely to wait their turn.

Nonetheless, the LiveWire condition was ranked second or third by most of the participants. The most commonly mentioned reason for ranking it second was that "some video was better than no video." One could at least monitor the reactions of one person, if not all of them.

However, both these people and people who ranked it last had a variety of complaints about the system. Some participants said they felt more self-conscious about speaking because they knew that speaking or interrupting would cause them to appear on the full screen for everyone else.

Another major criticism was that the automatic switching was often distracting and inappropriate, especially when people in the group coughed or laughed, causing the screen to quickly switch from one person

to the next. This is probably why participants rated LiveWire as having more inappropriate interruptions than the PIP system, interruptions in this case referring to visual distractions. This, taken together with people's concerns about taking over from others, probably explains why the LiveWire system had significantly fewer overlaps in speaker switches than both the audio-only and PIP conditions. It is likely that conversants were trying to avoid these distracting visual effects.

Other criticisms included participants' statements that they disliked the fact that they had no control or choice over whom to monitor. Sometimes participants wanted to monitor other people's reactions to what was being said. Also, the rule that the speaker sees the last person who spoke was inappropriate when that speaker wished to address another member of the group. A related point raised was that the design of LiveWire caused the participants to lose a sense of the larger group–people who were not speaking had virtually no persistent presence. Finally, the fact that there was no feedback to confirm who was viewing whom was said to be very problematic. When people are copresent, what other people are looking at can be easily deduced by the participants. However, in the LiveWire system, speakers got no feedback that they were being seen by others, especially because any speaker continues to see the previous speaker. There is, thus, no indication that they have successfully "taken over" the screen. Participants also said they wondered at whom the current speaker was looking, because they often forgot who spoke last.

Does Visual Access Benefit Conversation?

Participants' comments painted a more complex picture of the benefits and drawbacks of visual access during conversation than the speech measures suggest. Despite being inexperienced with videoconferencing technology, visual access was perceived to be beneficial by most of the participants. Participants' comments, combined with the results of the questionnaires from the two experiments, suggest that visual access in conversation may confer benefits by:

1. Providing nonverbal feedback from others (e.g., facial expressions, orientation, and gestures). This helps participants to gauge the reactions of others and helps them to know whether others are paying attention.
2. Helping to identify and discriminate among speakers, as well as helping to keep track of the conversational thread. This becomes increasingly important as the size of the group becomes larger.
3. Making participants feel more part of the group and less remote from the other conversants. Especially for systems that present all remote participants simultaneously, others in the group are given a presence they would not otherwise have.

4. Increasing the interactivity, or naturalness, of the conversation. However, this opinion must be contrasted with some participants who felt ill at ease with video technology and liked the anonymity of the audio-only system. Presumably, experience with video technology would affect this factor.

5. Affording parallel and side conversations for systems that enable selective gaze. Selective gaze also appears to be effective in preventing some kinds of conversational breakdown by allowing participants to use visual cues to indicate the person they are referring to or addressing.

The various systems investigated in this experiment were different in the extent to which they provided these benefits, and consideration of these differences raises important design issues. For example, consider the issue of providing nonverbal feedback through video. Interpreting the reactions of others is dependent upon any one person knowing something about what the other participants are perceiving. When people share the same physical space, they also share a common frame of reference with which to deduce what others are perceiving. Video systems provide remote participants with disjunct, individual frames of reference that make these inferences difficult. This was a general issue of concern that was raised by participants in both experiments: Participants wanted to know if they were being perceived, how they were perceived, and who was monitoring whom.

The video systems were different in the extent to which they provided this information. The PIP system provided each participant with the outgoing picture of themselves, so they could be reassured that they were on camera and see how they were framed. Being able to see everyone also gave participants information about who was attending in a general way, even if they could not tell specifically who was attending to whom. The Hydra system provided better information about who was attending to whom, but the lack of an outgoing picture led to concerns about whether one was properly framed and, thus, whether one's gestures and visual conduct could be perceived. The LiveWire system did not provide information for any of these aspects. If anything, it seems to have increased participants' uncertainty about whether they were being viewed and about who was looking at and attending to whom.

The LiveWire situation also points out that design can be such that the particular problems it creates can outweigh its potential benefits. The particular design of LiveWire took choice and control over who one monitors away from the participants. This was problematic in itself, and the fact that participants did not want to be limited to viewing only the speaker emphasizes the fact that video should not be seen merely as supplemental to speech, but that the nonverbal behavior of others was an important source of information that participants wanted actively to access.

In summary, from the results of the questionnaires and interviews, we can infer that being able to see each other in conversation impacts that conversation in a range of ways that are difficult, and probably in some cases impossible, to capture in surface-level, structural measures of speech. Further, these results suggest that adding a visual channel to communication systems cannot be based on simplistic assumptions about the role of visual information but must account for other aspects of interaction, such as the need for feedback and control of access–qualities inherent in conversations when people occupy the same physical space.

4. CONCLUSION AND IMPLICATIONS FOR DESIGN

At the beginning of this investigation, a question was posed, namely, how are video-mediated conversations different from same-room and audio-only conversations, if at all? One focus was on the regulation and synchronization of conversation, and a variety of different measures were analyzed to find out if differences among the conditions existed. Far from discovering that visual cues such as selective gaze benefitted the regulation of conversation, it was discovered that measures that might indicate turn-taking behavior were unaffected by a complete absence of visual information. Doubtless, had other, poorer quality systems been compared (such as those with degraded audio), problems in regulation would have arisen (as they did in Cohen, 1982 and O'Conaill et al., 1993). However, the speech measures indicated that conversational regulation is not problematic given that a good quality audio channel is available. Perhaps this is because audio cues for managing turn-taking were substituted for the visual cues that usually serve this function in the audio-only condition. Whatever the mechanism, remote participants appear to manage perfectly well without video.

Having said that, some interesting differences among conditions did emerge. Specifically, these experiments found that, rather than occurring more when visual cues were reduced, interruptions were more likely to occur when people occupied the same physical space. Further, phenomena such as simultaneous starts, which might otherwise be taken to be indicative of problems in floor control, were just as likely to occur when talking in the same room as when conversing via mediating technology. An argument was made, therefore, that these measures can be taken to be an indication of the formality and interactivity of a situation rather than of problems in turn-taking. These findings are in line with many other findings in this area, and their interpretation owes much to Rutter's (1987) inferences in a similar line of investigation.

Contrary to Rutter's (1987) view, however, the results also showed that the finding of increased disengagement and, hence, increased formality in video-mediated conversations cannot be attributed to simple notions of reduced communicational bandwidth in technology-mediated conversations. Rutter's (1987) model of cuelessness proposes that it is the aggregate

of cues that determines a feeling of psychological distance. The fewer cues there are in the aggregate, the greater the psychological distance one feels from other participants, and thus the more task-oriented and less spontaneous the style of conversation. Experiment 2 showed that in the audio-only context, where cues are limited to the linguistic and paralinguistic domains, people were no more formal and less spontaneous in style than when both audio and video cues were available. Conversants in the audio-only condition were equally uninterruptive of each other, were just as likely to use formal methods of handing over the floor, and produced comparable switch times between speaker transitions as when visual information was present.

According to the findings of these two experiments, therefore, what appears to be important is not the number and type of cues present, but whether the conversations were mediated by technology. Mediated conversations, using any of three different video systems or using only an audio connection, exhibited all the symptoms of depersonalization, psychological distance, and formality that Rutter (1987) had previously commented on. This suggests that the notion of cues as something generic and as entities that can be aggregated is too simplistic. There appears to be something critically different about sharing the same physical space that needs to be examined more carefully.

It is perhaps not surprising that the results show the inadequacy of a theoretical approach that aims to quantify or measure communicational bandwidth. We know that in many aspects of behavior, people are nonlinear creatures. That is to say, increases or decreases in the number or intensity of stimuli are typically not accompanied by incremental changes in behavioral response. So, too, in these experiments, predictions based on notions of increasing or decreasing cues simply fail to account for the data. With regard to conversational behavior, there is no simple, predictable way in which behavior can be seen to "degrade" with a reduction in the "number" of available cues.

Although this in itself may be no real revelation, the finding of the relative equivalence of different kinds of technology-mediated conversation in terms of some sort of notion of psychological distance offers some guidance in the development of a new model of video-mediated communication. Such a model must attempt a deeper understanding of the relation between the mediating technology and behavior. What aspects of interaction are fundamentally altered when people no longer share the same physical space? Why might these distortions cause people to feel less in control or less a part of the interaction? Why does the presence of a video channel fail to compensate? The findings of this research suggest that these are the important questions.

The foregoing argument makes no distinction between audio and audio plus video systems. In this sense, the results could be seen as rediscovering conclusions reached twenty years ago–that adding a video channel doesn't

much matter, and we might just as well settle for the telephone for remote conversations. Here again, though, the perceptions and opinions of the participants provide important insights: Access to visual information was considered by most to be both important and beneficial in conversational interaction, for a variety of reasons. Problems arose in the particular ways the technology provided visual access, and these problems varied from system to system. The implication is that researchers need to address ways of providing the benefits of visual access that can override such problems. Effective design may be able to compensate for the shortcomings of video systems as they are currently configured. It may even be possible that in doing so, the effects of the mediating technology, like those of depersonalization and disengagement, can be overcome.

Several design issues surfaced throughout these two experiments, but there were three that were emphasized. First, the issue of personal space is one that arose with reference to Hydra. It seems that the attempt to recreate personal space for remote participants through something akin to "video surrogates" may help people keep track of conversational threads through separation of audio sources and may facilitate processes such as parallel conversations.

Reciprocity and feedback is a second important, underlying issue. When conversants share the same space, they know if they can be seen by others, know when others are attending to them, and can use the reactions of others to modify their own behavior appropriately. Video systems make it difficult to know what others perceive. The use of confidence monitors is a step in the right direction, but it is a solution that tends to make people self-conscious. The Hydra design gave people more clues about who is attending to whom, but did not solve the other problem of helping people to know how they were framed. It appears that there are often cost–benefit trade-offs involved in providing effective feedback, but it is important enough to warrant careful thought.

Finally, control and amount of visual access is a third design issue that must be considered. Conversants want to be able to exercise some control over where they look. Groups also give rise to responses en masse (e.g., laughter, disagreement, and so on), emphasizing the need for participants to have access to everyone. Further, if the view of the group is limited, individuals lose their visual presence, which presents an unnatural situation. Systems that choose what to view automatically and, in doing so, restrict users' freedom to monitor the group, thus suffer by denying control and choice of what to monitor.

Many other issues arise when one goes beyond the narrow conversational context examined in these experiments. For example, most meetings and discussions involve shared visual workspaces and physical artifacts, and researchers have found that, given a choice, people will center on views of documents and artifacts more frequently than on face-to-face views (Gaver et al., 1993). Providing people with the ability to

access these task-centered views, to more fully explore the remote space, and to see people in relation to their activities are among other potential benefits that video can bring to these sorts of situations. But, in turn, providing these benefits will necessarily invoke a new set of problems for consideration. Design of systems for remote collaboration, therefore, presents a complex challenge, but experimental analysis of existing systems can provide valuable help in understanding how different design variables may act to shape the nature of interaction through technology.

NOTES

Background. The first experiment reported in this article was carried out while Abigail Sellen was a member of the CAVECAT Project at the University of Toronto. The second experiment was carried out in collaboration with the Toronto Telepresence Project. This article builds on earlier work reported in Sellen (1992).

Acknowledgments. These experiments were only made possible by relying heavily on the skills and good nature of my colleagues in the Computer Systems Research Institute at the University of Toronto. I am grateful to all of them. Bill Buxton was responsible for much of the conceptualization behind the design of the various systems reported in this article. Gordon Kurtenbach configured the hardware for the speech tracking equipment and wrote ConEd, the software for recording and analyzing the speech time lines. Garry Beirne, Gary Hardock, and Tom Milligan helped to set up and run the experiments. Finally, the patience of all the people in the CAVECAT project and the Dynamic Graphics Project at the University of Toronto is also much appreciated. I also gratefully acknowledge the contribution of the Arnott Design Group of Toronto for the design and fabrication of the Hydra models. For help with the data analysis, I thank Rebecca Jeanes, Tracy Narine, and Ian Nimmo-Smith. For valuable comments on the article itself, I thank Bill Buxton, Marge Eldridge, Gifford Louie, Allan Maclean, Judy Olson, and three anonymous reviewers.

Support. This work was supported by Rank Xerox Research Center (EuroPARC), the Toronto Telepresence Project, Xerox PARC, the Ontario Information Technology Research Centre, the Natural Sciences and Engineering Research Council of Canada, and The Arnott Design Group (Toronto).

Author's Present Address. Abigail J. Sellen, Rank Xerox EuroPARC, 61 Regent St., Cambridge, CB2 1AB, U.K. E-mail: `sellen@europarc.xerox.com`.

HCI Editorial Record. First manuscript received March 22, 1993. Revision received March 4, 1994. Accepted by Judith S. Olson. Final manuscript received October 17, 1994. – *Editor*

REFERENCES

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze.* London: Cambridge University Press.

Argyle, M., Lalljee, M., & Cook, M. (1968). The effects of visibility on interaction in a dyad. *Human Relations, 21,* 3–17.

Beattie, G. W. (1983). *Talk: An analysis of speech and non-verbal behavior in conversation.* Milton Keynes, England: Open University Press.

Bly, S. A., Harrison, S. R., & Irwin, S. (1993). Media spaces: Bringing people together in a video, audio, and computing environment. *Communications of the ACM, 36(1),* 28–46.

Brady, P. T. (1968, January). A statistical analysis of on–off patterns in 16 conversations. *The Bell System Technical Journal, 47,* 73–91.

Cohen, K. M. (1982). Speaker interaction: Video teleconferences versus face-to-face meetings. *Proceedings of Teleconferencing and Electronic Communications,* 189–199. Madison: University of Wisconsin Press.

Cook, M., & Lalljee, M. (1972). Verbal substitutes for visual signals in interaction. *Semiotica, 3,* 212–221.

Dabbs, J. M., Jr., & Ruback, R. B. (1984). Vocal patterns in male and female groups. *Personality and Social Psychology Bulletin, 10,* 518–525.

Dabbs, J. M., Jr., & Ruback, R. B. (1987). Dimensions of group process: Amount and structure of vocal interaction. *Advances in Experimental Social Psychology, 20,* 123–169.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23,* 283–292.

Fish, R., Kraut, R., Root, R., & Rice, R. (1992). Evaluating video as a technology for informal communication. *Proceedings of the CHI '92 Conference on Human Factors in Computing Systems,* 37–48. New York: ACM.

Gaver, W. (1992). The affordances of media spaces for collaboration. *Proceedings of Computer-Supported Cooperative Work '92,* 17–24. New York: ACM.

Gaver, W., Moran, T., Maclean, A., Lovstrand, L., Dourish, P., Carter, K., & Buxton, W. (1992). Realizing a video environment: EuroPARC's RAVE system. *Proceedings of the CHI '92 Conference on Human Factors in Computing Systems,* 27–35. New York: ACM.

Gaver, W., Sellen, A. J., Heath, C., & Luff, P. (1993). Multiple views in a media space: One is not enough. *Proceedings of the INTERCHI '93 Conference on Human Factors in Computing Systems,* 335–341. New York: ACM.

Heath, C., & Luff, P. (1991). Disembodied conduct: Communication through video in a multi-media office environment. *Proceedings of the CHI '91 Conference on Human Factors in Software,* 99–103. New York: ACM.

Jaffe, J., & Feldstein, S. (1970). *Rhythms of dialogue.* New York: Academic.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 32,* 1–25.

Mantei, M., Baecker, R., Sellen, A., Buxton, W., Milligan, T., & Wellman, B. (1991). Experiences in the use of a media space. *Proceedings of the CHI '91 Conference on Human Factors in Software,* 203–208. New York: ACM.

McGrath, J. E. (1984). *Groups: Interaction and performance.* Englewood Cliffs, NJ: Prentice-Hall.

Nardi, B., Schwarz, H., Kuchinsky, A., Leichner, R., Whittaker, S., & Sclabassi, R. (1993). Turning away from talking heads: The use of video-as-data in neurosurgery. *Proceedings of the INTERCHI '93 Conference on Human Factors in Computing Systems,* 327–334. New York: ACM.

Ochsman, R. B., & Chapanis, A. (1974). The effects of 10 communication modes on the behaviour of teams during co-operative problem-solving. *International Journal of Man-Machine Studies, 6,* 579–619.

O'Conaill, B., Whittaker, S., & Wilbur, S. (1993). Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human–Computer Interaction, 8,* 389–428.

Rutter, D. R. (1987). *Communicating by telephone.* Elmsford, NY: Pergamon.

Rutter, D. R., & Stephenson, G. M. (1977). The role of visual communication in synchronizing conversation. *European Journal of Social Psychology, 2,* 29–37.

Rutter, D. R., Stephenson, G. M., Ayling, K., & White, P. A. (1978). The timing of looks in dyadic conversation. *British Journal of Social and Clinical Psychology, 17,* 17–21.

Rutter, D. R., Stephenson, G. M., & Dewey, M. E. (1981). Visual communication and the content and style of conversation. *British Journal of Social Psychology, 20,* 41–52.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking in conversation. *Language, 50,* 696–735.

Sellen, A. J. (1992). Speech patterns in video-mediated conversations. *Proceedings of the CHI '92 Conference on Human Factors in Computing Systems,* 49–59. New York: ACM.

Sellen, A. J., Buxton, W., & Arnott, J. (1992). *Using spatial cues to improve desktop video conferencing* [Videotape presented at CHI '92 Conference on Human Factors in Computing Systems]. New York: ACM.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana: University of Illinois Press.

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

Toothaker, L. (1991). *Multiple comparisons for researchers.* Newbury Park, CA: Sage.

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