Chapter 2

The Role of Vision in Face-to-Face and Mediated Communication

Steve Whittaker
AT & T Laboratories

Brid O'Conaill
Hewlett Packard Laboratories

This chapter uses a general communication framework to predict where visible information supplied by video might be critical in mediated communication. We first identify fundamental features of communication that have to be supported in any conversation, regardless of the available communication modalities. We characterize the different types of visible information that play a role in face-to-face interaction and the communication features they support. We then use the framework to evaluate three predictions about the benefits of video in mediated communication: (a) Video supports visible behaviors and hence supplies nonverbal information that is missing from the speech channel; (b) video provides visible information about the availability of other people, and hence supports connection for unplanned communications; (c) video provides dynamic visual information about objects and events that is important for certain collaborative tasks, an application that we refer to as "video-as-data." We evaluate existing work on VMC in terms of these three hypotheses. Current evidence suggests weak support for video as communicating one type of nonverbal information, that is, affective information. We conclude that there is insufficient data to evaluate the connection and video-as-data hypotheses. We discuss what refinements are needed in our theories about how visible information operates in mediated communication, and briefly examine the technology and design implications of these results.

INTRODUCTION

Face-to-face communication is a multimodal process. It involves complex interactions between verbal and visual behaviors. As people speak, they gesture for emphasis and illustration, they gaze at listeners and visually monitor their environment, their facial expressions change, and their body posture and orientation shift as they talk. Likewise, listeners look at speakers, as the speakers talk. Listeners monitor speaker's facial expressions and gestures, they nod their heads to show assent, and their facial expressions and physical posture change depending on their interest in and attitude to the speaker's utterance. Furthermore as people interact, they orient to, gesture at, and manipulate physical objects in the environment they share.

Despite the multimodal nature of face-to-face communication, the most pervasive and successful technology for communicating at distance is the telephone, which relies solely on the voice modality. Early attempts to supplement the voice modality by adding visible information about other conversational participants have not led to the expected improvements in remote communication. Laboratory studies to demonstrate the benefits of adding a visual communication modality to voice have in general shown few objective improvements (Chapanis, 1975; Chapanis, Ochsman, Parrish, & Weeks, 1972; Reid, 1977; Rutter & Robinson, 1981; Sellen, 1995, this volume; Short, Williams, & Christie, 1976; Williams, 1977). Furthermore, technologies premised on the advantages of supplementing voice with visible information, such as the videophone and the videoconference, have not yet proved greatly successful in the marketplace (Egido, 1990; Noll, 1992).

This prior work shows that the role of visible information in communication is both complex and subtle. From a theoretical viewpoint, it indicates that we need more detailed understanding of the precise functions that visible information plays in communication. From a practical viewpoint, we need to understand when and how visible information is vital for communication. This understanding is important in the design of technologies that exploit visible information to provide more effective remote communication than is currently supported by the telephone. Much of the prior technology-oriented work on video-mediated communication (VMC) has been based on the intuition that visible information will necessarily benefit interaction, without having specific hypotheses about how those benefits will come about.

The aim of this chapter is to use a general communication framework to predict and evaluate where visible information might be critical in mediated communication. The structure of this chapter is as follows. We begin by presenting a communication framework, specifying fundamental features of communication that have to be supported in any conversation, regardless of the communication modalities available. We then describe the different types of visible information that play a role in face-to-face interaction and the conversational features they support. We distinguish between two main classes of visible information: visible behaviors produced by the participants, and information about the visible environment. Visible behaviors include the gaze, gesture, facial expressions, and posture of the participants. The visible environment includes information about objects and events that have been observed by the conversational participants, as well as information about the availability of other participants. On the basis of this analysis of the communicative function of visible information, we evaluate three predictions about how video might improve communication. The three hypotheses are:

1. Video supports visible behaviors and hence supplies important nonverbal information.
2. Video provides visible information about the environment, specifically the availability of other people; this in turn facilitates connection for unplanned communications.
3. Video provides dynamic visual information about objects and events in a shared visual environment. This is important for certain collaborative tasks, an application that we refer to as *video-as-data.*

We evaluate existing work on VMC in terms of these three hypotheses. We conclude by discussing the impact of these results for theories about how visible information operates in mediated communication, and exploring the implications for technology designers.

COMMUNICATION FRAMEWORK

There are a number of key aspects of communication that have to be supported regardless of what communication modes are available (see Table 2.1). Communication is a joint activity that requires coordination of both process and content between speakers and listeners (Clark & Brennan, 1991; Whittaker, Brennan, & Clark, 1991). Content is the subject matter of conversation, and concerns what participants talk about. Content coordination addresses how participants build up common beliefs and understanding about that subject matter. Process is concerned with the mechanisms and management of conversation. The key issues for process coordination are (a) the set of procedures by which participants agree to begin and end entire conversations and (b) the rules that allow participants to switch roles between speaking and listening.

Coordination of Process

We address two aspects of process coordination: turn-taking and availability. Turn-taking is concerned with how participants jointly determine who will speak, who will listen, and how transitions are made between these roles

TABLE 2.1
A Taxonomy of Conversational Mechanisms and Their Communication Functions

	Conversational Mechanism	*Communication Function*
Process coordination	Turn-taking cues	Determine who will speak, who will listen, and how transitions are made between these roles
	Availability cues	Determine when to initiate or end a communication episode
Content coordination	Reference	Allow participants to identify the objects and events they jointly want to talk about
	Feedback cues	Inform speaker of listener's understanding; contribute to maintenance of mutual beliefs and common knowledge between speaker and listeners
	Interpersonal cues	Allow participants to infer the emotional stance, affect, and motivations of other people to what is being discussed, and to other conversational participants

(Sacks, Schegloff, & Jefferson, 1974; Walker & Whittaker, 1990; Whittaker & Stenton, 1988). A striking feature of conversations is that in general only one person speaks at any point. Less than 5% of speech is delivered in overlap (two people speaking at the same time), and yet gaps between one speaker finishing and the next starting are frequently measurable in milliseconds (Levinson, 1983). Turn-taking addresses how these transitions are successfully achieved (Sacks et al., 1974). Process coordination also addresses how entire conversations are initiated and concluded (Schegloff & Sacks, 1973). Many communications are unplanned. They therefore require participants to establish precisely when other people are available for communication and when it is opportune to initiate such spontaneous interactions, based on awareness of others' movements and activities (Heath & Luff, 1991; Kendon & Ferber, 1973; Kraut, Fish, Root, & Chalfonte, 1993; Whittaker, Frohlich, & Daly-Jones, 1994).

Coordination of Content

Coordination of content is concerned with how participants arrive at, and maintain, common understanding in conversation (Clark & Brennan, 1991; Clark & Marshall, 1981; Grosz & Sidner, 1986). Coordinating content presents problems for both speakers and listeners. From the listener's perspective, one of the fundamental problems with human communication is that the literal meaning of an individual utterance underspecifies the speaker's intended meaning (Clark & Marshall, 1981; Grice, 1975). Listeners have to infer the speaker's intended meaning by supplementing what was said with contextual information external to the utterance. In

general, listeners are able to generate such external inferences quickly and accurately, and hence determine the speaker's intentions (Levinson, 1983). How are listeners able to do this and how do they know which information not directly stated in the utterance should be used in deriving inferences? Clark and Marshall (1981) argued that this external information is a restricted set, common knowledge, which is shared between participants.

Common knowledge is crucial for achieving an important aspect of communication, namely, reference. Reference enables conversational participants to jointly identify the objects and events that they want to talk about. Speakers have multiple communication choices when they wish to refer to an object or event (Clark & Marshall, 1981). A speaker may therefore verbally refer to a dog as "the dog." Alternatively, there may be more precise or specific ways to describe the animal, such as "Marvin," "the dachshund," or "whoever chewed my slipper." If the animal is present they may point at it, or gesture and say "that animal" or even "that." Speakers make these choices based on the common knowledge they share with listeners: It would be of little use to describe the dog as "Marvin" to a complete stranger. Likewise, the term *dachshund* is almost certain to fail with a very young child. When speakers make these reference choices they must therefore balance the additional precision these terms provide, against the likelihood their audience may not understand (Clark & Marshall, 1981).

A second problem for speakers in coordinating content is to determine whether their utterance had the intended effect, that is, whether the listener drew the correct set of inferences from what they said. Listeners' knowledge and beliefs are usually not directly accessible to the speaker (Walker, 1993), so feedback mechanisms are crucial for the maintenance of common knowledge. Speakers provide listeners with frequent opportunities to offer feedback about what was just said (Kraut, Lewis, & Swezey, 1982; O'Conaill, Whittaker, & Wilbur, 1993; Yngve, 1970)—to show acceptance (Clark & Schaefer, 1989) or to clarify their level of understanding (Walker & Whittaker, 1990; Whittaker & Stenton, 1988). These feedback processes take place on a moment-by-moment basis in conversation, so that misunderstandings can be quickly identified and rectified (Clark & Brennan, 1991; Kraut et al., 1982; O'Conaill et al., 1993).

Communication is not restricted to the exchange of propositional information, however, and a final aspect of content coordination concerns the affective state or interpersonal attitude of the participants. This is social information about participants' feelings, emotions, and attitudes to the other conversants and to what is being discussed. As with conversational intentions, participants generally do not make this information verbally explicit, so it usually has to be inferred. Access to affective information is important: It can change the outcome of conversations in situations where emotion plays a critical role, such as negotiation (Short et al., 1976).

THE ROLE AND FUNCTION OF VISIBLE INFORMATION IN COMMUNICATION

We now turn to an analysis of the types of visible information that are used to support some of these features of face-to-face communication. In face-to-face conversation, there are two different types of visible information. The first is information about the behaviors of other conversational participants, that is, the set of communicative actions that they perform with their eyes (gaze), faces (facial expressions), hands and arms (gestures), and the movements and orientation of their bodies (posture). The second type of visible information is about the visible environment that conversational participants share, with its set of shared objects, access to shared events, and information about the movements and activities of other people.

It is important to note that the correspondence between visible information and communication function is not direct: The same type of visible information can support multiple communication functions, and likewise one communication function may be supported by different types of visible information. For example, one type of visible behavior, gaze, supports multiple functions. It can coordinate reference, give feedback of understanding the utterance, and aid turn-taking. Similarly, one communication function may be mediated by multiple different types of visible behavior. Thus turn-taking can be supported by gaze, gesture, and posture (Beattie, 1978, 1981; Duncan, 1972; Kendon, 1967).

Visible Behaviors

Gaze. Gaze is the way people extract visible information from their environment. The direction in which a person looks, the amount of time looking in that direction, and the patterns of gaze are all important aspects of this visible behavior. The communicative functions of gaze are shown in Table 2.2.

Gaze is a general indicator of attention and can be directed at other conversational participants, as well as at features of the physical environment. By gazing at the speaker, listeners derive important information from the speaker's facial expressions, posture, and gesture. In this way, the speaker's visible behaviors help clarify the content of what is being said. Speakers interpret listeners' attentional behavior as feedback for determining how well their message is being understood and coordinate their content accordingly (Beattie, 1978, 1981; Clark & Brennan, 1991; Goodwin, 1981). On some occasions speakers also try and elicit this type of feedback: They may pause in their speech until they detect visible attention from listeners (Goodwin, 1981).

TABLE 2.2
Communicative Functions of Gaze

	Conversational Mechanism	*Gaze Behaviors*	
		Speaker	*Listener*
Process coordination	Turn-taking cues	Speaker predominantly looks away from listener while talking—negotiated mutual gaze used as "turn-yielding" signal	Listener predominantly looks at speaker while speaker is talking; negotiated mutual gaze used as "turn-accepting" cue
	Availability cues		
Content coordination	Reference	Gaze at an object indicates person's interest and attention to that object; joint attention allows pointing	Gaze at an object indicates person's interest and attention to that object; joint attention allows pointing
	Feedback cues	Gaze at listener can be an attention-eliciting device	Gaze at the speaker indicates interest in what the speaker is saying
	Interpersonal information cues	Patterns of gaze interpreted as indicating sincerity, trustworthiness, friendliness; indicate speakers' affective attitude to utterance	Patterns of gaze interpreted as indicating sincerity, trustworthiness, friendliness; indicate listeners' response to utterance

Gaze may also be coordinated between conversational participants to achieve mutual gaze or joint attention. Mutual gaze occurs when two participants are concurrently looking at each other. Joint attention is when participants are mutually oriented to a common part of their shared visible environment and are aware that their conversational partners are also looking at it. People are very good at determining where others are looking. This facilitates joint attention, which allows greater flexibility in referring to objects, because the speaker can infer which objects are highly salient to listeners (Clark & Marshall, 1981).

Gaze is also an indicator of interpersonal attitude or affect. Speakers tend to gaze at a listener's face more when they are being more persuasive, deceptive, ingratiating, or assertive (Kleinke, 1986), and are more likely to look at conversants whom they like (Exline & Winters, 1965). In addition, people tend to evaluate others by their patterns of gaze: People who look at their interlocutor only a small part of the time are judged as "defensive" or "evasive," whereas those who look a lot of the time are "friendly," "mature," and "sincere" (Kleck & Nuessle, 1968).

Gaze is also a key mechanism in the co-ordination of the conversational process of turn-taking (Kendon, 1967). As speakers draw to the end of a natural phase in their utterance, they are likely to gaze directly at their listener. By looking at the listener, speakers signal that they are ready to finish speaking. They then wait for a gaze from the listener confirming that the listener is ready to continue. When the listener and speaker achieve mutual gaze this serves as a confirmation of the transition, and the listener then takes over as next speaker.

Finally, it is important to note that participants spend relatively small amounts of time gazing at others while conversing. The amounts of gaze directed at others can be as low as 3% to 7% of conversational time, in the presence of relevant visible objects (Argyle, 1990; Argyle & Graham, 1977). Mutual gaze is even lower with recent studies reporting levels of below 5% (Anderson, Bard, Sotillo, Newland, & Doherty-Sneddon, in press). Given these low frequencies, it appears that conversational participants have restricted opportunities for eliciting visual information about others and neither speakers nor listeners have access to all the visible behaviors of others. This work also shows that conversationalists are often more visually focused on their environment than on other people (Argyle, 1990; Argyle & Graham, 1977).

Gesture. Gesture is the set of dynamic movements and shapes formed by a person's hands and arms during communication. Like gaze, it also supports multiple communication functions, shown in Table 2.3. It is used to coordinate conversational content, achieve reference, and assist in turn-taking. Some gestures, known as emblems, have conventionalized meanings, such as the "thumbs up," and "V for victory" (Efron, 1982). These can substitute for spoken words or phrases. Other iconic gestures can express propositions with a spatial or dynamic aspect: "It was this big," or "it moved like this," where the word "this" is illustrated by an appropriate gesture. A final class of gesture (known variously as "beats," "batons," or "speech primacy movements"), allows speakers to emphasize, evaluate, or "comment on" the information they are conveying verbally. In some cases "beats" serve the function of coordinating talk with the listener (Cassell, McNeill, & McCullough, in press; McNeill, 1992).

Given joint attention, pointing gestures can be used to achieve reference. Pointing gestures can also be used to manipulate or direct the attention of others, either by pointing and using speech, such as "look at that" or by pointing alone (Goodwin, 1981). Gesture can also be used to communicate more abstract characteristics of the space the speaker is talking about, such as the relative positions of the objects it contains, and their relative orientations (Cassell et al., in press; McNeill, 1992). Finally, gesture can serve to coordinate turn-taking transitions, and hence serve to coordinate process. The continuation of any speaker hand movement—of whatever gestural type—acts as a cue that the current speaker wishes to hold the conversational floor. Similarly, the termination of gesture—again regardless of gestural type—acts as a signal that the speaker is ready to hand over the conversational floor, and is therefore referred to as a "turn-yielding" cue. Hand gestures can also be used by listeners to signal that they want to say something (Goodwin, 1981).

TABLE 2.3
Communicative Functions of Gesture

	Conversational Mechanism	*Gestural Behaviors*	
		Speaker	*Listener*
Process coordination	Turn-taking	Termination of speaker gesture interpreted as "turn-yielding" cue	Listener gestures signal desire to speak
	Availability		
Content coordination	Reference	Pointing facilitated by joint attention	Pointing facilitated by joint attention
	Feedback		
	Interpersonal information		

Facial Expression. Facial expressions are conveyed by the eyes, eyebrows, nose, mouth, and forehead. Information from the eyebrows and mouth is of prime importance in facial expressions (Ekman & Friesen, 1975). Facial expressions play a role in coordinating content: They provide listener feedback and serve as general indicators of emotional state. Information from the speakers' lips can also serve to disambiguate spoken content. These functions are shown in Table 2.4.

Facial expression offers speakers information about listeners' level of understanding. There are two methods by which the face provides feedback. The first, head nods, provides the speaker with concurrent feedback about what has just been said, and modification of the normal frequency and duration of head nods disrupts speakers' ability to communicate (Birdwhistell, 1970). Listener's facial expressions also reveal interest, puzzlement, or disbelief about what they are being told (Ekman & Friesen, 1975).

The face is also a rich source of information about the affective state of the conversational participants. The eyes, mouth, and eyebrows are highly expressive. Ekman and Friesen (1975) have shown that people across a number of cultures are able to recognize seven distinct facial expressions from posed photographs (happiness, sadness, surprise, anger, disgust, fear, and interest). Affective expressions allow listeners to infer speakers' current

TABLE 2.4
Communicative Functions of Facial Expressions

	Conversational Mechanism	*Facial Expressions*	
		Speaker	*Listener*
Process coordination	Turn-taking		
	Availability		
Content coordination	Reference	Visual information from reading the speaker's lips decreases the ambiguity of speech	
	Feedback		Head nods indicate assent or dissent; expressions indicate interest, understanding, puzzlement, or disbelief
	Interpersonal information	Expressions indicate happiness, fear, interest, surprise, sadness	Expressions indicate happiness, fear, interest, surprise, sadness

emotional state, and expressions allow their audience's emotional reaction to what is being said.

Finally, facial expressions, especially lip and teeth movements, can help decipher speech. The lip shape, teeth, and to a lesser extent tongue position give listeners visual information about the phonemes that the speaker is producing. Unintelligible speech can be rendered interpretable by visual information about lip shape. When lip shape information is available, listeners can interpret an additional 4–6 dB of noise and achieve the same level of intelligibility (Summerfield, 1992). The effect of visual information on speech perception is also demonstrated by the "McGurk effect," in which conflicting information from face and voice is "heard" in a way that combines both modalities. If the lips say "ga" and the voice "ba," then people hear it as "da" (McGurk & MacDonald, 1976).

Posture. This is the information supplied by the inclination and orientation of a conversational participant's body, in particular their trunk and upper body. The positions of both the arms and legs are also important here. Posture is less dynamic than other visible behaviors, with variations occurring less frequently. The communicative functions of posture are shown in Table 2.5. Posture is another cue as to the degree of interest or engagement of a conversational participant. It therefore provides feedback to the speaker about how the message is being received. Interest is signaled in listeners by leaning forward, and by speakers in leaning forward and drawing back their legs. In contrast, boredom is signaled by head lowering, or turning the head to one side, supporting the head on one hand, leaning back and stretching out one's legs (Bull, 1978). Body position and orientation can also be used by the speaker to include or exclude people from the conversation (Goodwin, 1981).

TABLE 2.5
Communicative Functions of Posture

	Conversational Mechanism	*Postural Behaviors*	
		Speaker	*Listener*
Process coordination	Turn-taking		Listener activity can signal a desire to interrupt
	Availability		
Content coordination	Reference		
	Feedback		Attention, interest in what speaker is saying
	Interpersonal information	Reveals speaker's attitude to utterance	Reveals listener's affective reaction to utterance

Sharing a Visible Environment: Information About Shared Events, Objects, and People

We have so far focused on the role of visible behavior, namely, the gaze, gestures, facial expressions, and posture of other conversational participants. In face-to-face conversation, however, the fact that participants have access to a shared physical environment means that other types of visible information are available, such as information about physical objects, events, and people. For the purpose of coordinating content, sharing the same physical environment enables people to make inferences about the set of objects and events that others in the same environment are likely to know about and want to talk about (Clark & Marshall, 1981; Whittaker et al., 1991, Whittaker, Geelhoed, & Robinson, 1993). Listeners can reduce the ambiguity of an incoming message, by using this physical information to infer what the message is likely to be about. Similarly, speakers can make inferences about what their audience might expect them to converse or know about, based on this shared physical information. Finally, people can make inferences about the availability of others for communication, based on visible information about those people. This availability information helps the process of initiating and terminating conversations.

Using Visible Information About Objects and Events in Collaboration. The visible environment includes information about the objects and events in the participants' shared environment, as well as their spatial configuration and interrelations. Information about the visible environment often interacts in important ways with verbal and visible behaviors, such as when participants gesture at, orient toward, and manipulate aspects of their environment. As we have seen, the visible environment provides crucial contextual information that can help participants determine what things are likely to be communicated about, and what entities are likely to be salient to others (Argyle & Graham, 1977; Cooper, 1973; Whittaker et al., 1993; Whittaker et al., 1994).

For tasks that require participants to jointly manipulate or modify complex objects, it is crucial to have access to a shared environment containing these objects to help coordinate content (Nardi, Schwarz, Kuchinsky, Leichner, Whittaker & Sclabassi, 1993; Nardi, Kuchinsky, Whittaker, Leichner, & Schwarz, 1996; see chapter 23, this volume; Gaver, Sellen, Heath, & Luff, 1993; Whittaker et al., 1993). Having access to a shared environment can be beneficial in multiple ways: Not only can participants mutually directly observe changes to that environment, but the shared environment also provides straightforward methods for people to exchange and simultaneously look at objects. Thus many workplace interactions involve documents, and a shared environment enables people to easily hand over a document, or for both participants to mutually orient to the document (Whittaker et al., 1994; Whittaker, Swanson, Kucan, & Sidner, in press). Furthermore, objects such as documents can be used as "context-holders" for intermittent workplace communications between colleagues. There are often long delays between different fragments of workplace conversations about ongoing shared tasks. Workers sometimes leave documents relating to current ongoing communications on their desktops as reminders of tasks in progress (Whittaker et al., in press). Similar "context-holding" functions were observed in intermittent interactions around different types of shared objects (Nardi, Kuchinsky, Whittaker, Leichner, & Schwartz, 1996; chapter 3, this volume).

Using Visible Information About People for Availability. Most simply, one can infer the presence of another person if the person is visible.[1] Furthermore, information about the proximity, current activities, and movements of other people has been shown to influence certain aspects of communication, such as its initiation and termination, as well as how interruptions are handled (Heath & Luff, 1991; Kendon & Ferber, 1973; Kraut et al., 1993; Tang, Isaacs, & Rua, 1994; Whittaker et al., 1994; Isaacs & Tang, chapter 9, this volume; Isaacs, Whittaker, Frohlich, & O'Conaill, chapter 22, this volume). This information is easily derived from the visible mode. In addition, other information about physical appearance is conveyed visibly and can support inferences about other participants' gender, age, and possibly dispositions.

Workplace interactions are generally unplanned (Isaacs et al., chapter 22, this volume; Kraut et al., 1993; Whittaker et al., 1994), and visible information provides mechanisms for initiating those types of communication. First, sightings of others can lead one to fall into spontaneous conversation, such as in public areas like coffee areas (Kendon & Ferber, 1973). In addition, seeing a colleague may remind one of an issue that needs to be discussed, so that the sighting serves as a reminder that a conversation needs to take place (Kraut et al., 1993; Whittaker et al., 1994). Visible information is also helpful in determining whether a colleague is receptive to an unplanned conversation, offering vital clues as to how available or interruptible they are. For example, passing by a colleague's office, it is possible to determine whether they are present and, if so, infer whether they can be interrupted and in some cases, how long such an interruption should last (Frohlich, 1995; Isaacs et al., chapter 22, this volume; Whittaker et al., 1994). Finally, this class of availability information can influence the termination and character of conversation. A substantial number of dyadic workplace communications are ended or changed by the arrival of a third party. Often this person indicates a desire to interrupt or join the conversation by "hovering," waiting for the current conversation to reach a point where the person can break in (Whittaker et al., 1994). Again, the information indicating to the conversationalists that another person wishes to interrupt or join them, is available in the visual channel.

EVALUATING THE EFFECTS OF ADDING VISIBLE INFORMATION TO AUDIO-ONLY COMMUNICATION

Given the analysis of critical communication features and the different functions of visible information, where might we expect to find benefits to supplementing audio with visible information? The preceding analysis suggests three distinct hypotheses about how visible information provided by video might improve audio-only communication. Extensive reviews of these hypotheses are also provided in Whittaker (1995, 1996).

Nonverbal Communication Hypothesis

The nonverbal communication hypothesis is that visible behaviors such as gaze, gesture, facial expressions, and posture provide information that is absent from audio-only communication. One evaluation of the hypothesis has been to conduct short-term laboratory studies comparing video-mediated with (a) audio or (b) face-to-face interaction, in the context of a particular communication task. The comparison with audio reveals how and when video information enhances speech-only communication, and the comparison with face-to-face communication about how effectively video/audio mimics face-to-face conversation. Another technique has involved longer term field studies, installing video systems to evaluate their use (Abel, 1990; Bly, Harrison, & Irwin, 1993; Fish, Kraut, & Chalfonte, 1990; Fish, Kraut, Root, & Rice, 1993; Gaver et al., 1992; Mantei et al., 1991; Tang et al., 1994). There are methodological difficulties with the two types of data from field evaluations, however. The first type of data is from workgroups at the same physical location using high-quality video systems. Here, the ready availability of face-to-face communication may reduce the incidence and use of video technology. There is also data from geographically distributed workgroups that have higher incentives to use the system but are using inferior video technology because of the constraints imposed by wide-area networking bandwidths.

There are three distinct versions of the nonverbal communication hypothesis, each of which addresses different features of communication (Whittaker, 1995, 1996): (a) video provides cognitive cues that facilitate shared understanding; (b) video offers process cues to support turn-taking; (c) video provides social cues and access to emotional information. Cognitive and social cues address the issue of coordinating conversational content, whereas turn-taking addresses process coordination. The prototypical systems here are the videoconferencing suite or videophone. We review evaluations of each of the three subhypotheses about the role of nonverbal communication, first for high-quality and then for low-quality systems.

Chapanis and colleagues (Chapanis, 1975; Chapanis et al., 1972) conducted a series of laboratory experiments testing the cognitive cuing hypothesis, namely, that visual cues such as head nods and gaze help speakers to evaluate listener's understanding and attention. They compared the effectiveness of a variety of different media combinations for different cognitive problem-solving tasks, by looking at task outcome measures such as time to solution and quality of solution. The tasks involved complex instruction giving and route planning. In one task, subjects had to jointly construct a mechanical object where one person had the physical components and the other had the instructions. In another task, one person was given a map and the other given a copy of the *Yellow Pages.* They were asked to identify a map location satisfying a number of criteria, such as the nearest dentist to a given street address. The research compared two media conditions: audio-only communication, and high-quality video/audio, where the video showed the head and shoulders of the remote participant. However, the studies showed that adding visual information in tasks where it is important to track the understanding of remote participants did not increase the efficiency of problem solving, or produce higher quality problem solving.

Furthermore, other experiments comparing different combinations of media indicated that speech was the critical medium for interpersonal communication in collaborative problem solving: Removing the speech channel had huge effects on the outcome of communication. If participants could use the speech channel, then the addition or removal of video, text, or writing media had little effect on task outcome or quality of solution.

These results showing little impact of visual information on cognitive problem solving have been replicated by several other laboratory studies (Reid, 1977; Short et al., 1976; Williams, 1977). Most importantly, this is not an issue of video quality: Even face-to-face interaction is no better than speech only communication for this class of task (Williams, 1977). Similar negative results are suggested by field study research. A study of high quality local area videophone conducted over several months in a research laboratory showed few objective usage differences compared with the telephone (Fish et al., 1993).[2] Phone and videophone calls have similar durations, and are used for the same set of communication tasks. The researchers also administered a questionnaire asking people to state the tasks for which they felt that different communication techniques (e.g., videophone, telephone, face-to-face) were appropriate. Multidimensional scaling techniques applied to people's answers indicated that videophone is viewed by users as more similar to the telephone than face-to-face communication.

There is some counterevidence to these negative results, however. For a design task, Olson, Olson, and Meader (1995) showed that groups communicating face-to-face generated higher quality designs than audio-only groups, when those groups had access to a shared workspace. There were also differences in mutual understanding. Audio-only groups spent more time in stating and clarifying issues than groups that also had a high-quality video link.

The results are more mixed for the turn-taking hypothesis. Sellen (1995) investigated this in a series of laboratory studies of negotiation tasks, in which groups discussed contentious issues and tried to reach consensus. There was little evidence to support the claim that high-quality video information improves conversation management and turn-taking, when compared with audio-only conversations. For objective conversation process measures such as pausing, overlapping speech, and interruption management, there were no process differences between the video/audio systems and speech-only communication. Furthermore, none of the video/audio systems replicated face-to-face conversational processes. The video/audio systems reduced the ability of listeners to spontaneously take the conversational floor, as measured by number of interruptions.[3] Video/audio systems led speakers to use more formal techniques for handing over conversational initiative, such as naming a possible next speaker or using "tag" questions,[4] when compared with face-to-face interaction. Similar data are reported by O'Conaill et al. (1993), who also found speakers holding real meetings using high-quality videoconferencing used more formal turn-taking techniques than were observed in face-to-face interaction. One explanation of the failure of even high-quality videoconferencing to replicate face-to-face communication processes is that most videoconferencing systems do not support directional sound or visual cues. They tend to present sound and picture from a single monitor and speaker, which may compromise sound direction, head turning, and gaze cues in group interactions. Testing his hypothesis remains an outstanding research issue.

However, there are some differences in subjective data about turn-taking gathered from questionnaires. These differences concerned subject's impressions of the impact of video on conversational processes (Sellen, 1995). Video/audio is perceived to be better than speech in a number of ways. It is perceived to support interruptions; lead to more natural conversations that are more interactive; increase the ability to listen selectively to particular speakers; allow one to determine whether one is being attended to; and to generally keep track of the conversation. People also believe that they are better able to track the attention of others when they have video. Similar qualitative data are reported by Isaacs and Tang (1993), who found that video seemed to allow participants to manage pauses better than in speech-only communication. Despite this, Tang and Isaacs (1993) also found that high-quality video was not perceived as equivalent to face-to-face interaction: Subjective data showed that video was seen as less effective in supporting interactivity, selective attention, and the ability to take initiative in the conversation.

There is stronger evidence for the claim that video supports the transmission of social cues and affective information. Adding video information to the speech channel changes the outcome and character of communication tasks that require access to affect or emotional factors. Example tasks here include negotiation, bargaining, and conflict resolution. Participants focus more on the motives of others when they have access to visual information, and video/audio conversations are more personalized, less argumentative, more polite, and broader in focus. They are also less likely to end in deadlock than speech-only communications (Reid, 1977; Short et al., 1976; Williams, 1977). These results can be explained in terms of affective cues: Providing visual access to facial expressions, posture, and gesture allows people to make inferences about other participants' affective or emotional state. There are also subjective benefits to providing visual information: Participants believe that video/audio and face-to-face interaction are better than audio only for tasks requiring affect, such as getting to know other people, or person perception tasks. In addition, groups conversing using audio and video tend to like each other more (Reid, 1977; Short et al., 1976; Williams, 1977).

The preceding evaluations all used high-quality audio and video. Current technology limitations and restricted networking bandwidth mean, however, that high-quality systems will not be available for some years. It is therefore crucial for design and implementation that we understand the utility of low-quality video. One key finding from studies of low-quality video systems is that in certain circumstances adding visual information can detract from the interaction processes, if the video is implemented in a way that interferes with audio. There are two ways that audio can be affected in low bandwidth systems. First, certain commercial systems delay audio transmission, to allow time for video compression and decompression over wide-area networks, in order to present synchronized audio and video.[5] Second, some videoconferencing systems enforce half-duplex[6] audio to preserve bandwidth for video.

There is evidence that reducing audio quality to incorporate video is highly disruptive of turn-taking processes. In a naturalistic study, O'Conaill et al. (1993, chapter 6, this volume) compared face-to-face and video-mediated interaction in a low-quality wide-area system. The system had one-way half-duplex audio with one-way lags of between 410 and 780 msec and poor picture quality. The study measured a number of characteristics of conversation processes. Interactive aspects of conversation that required precise timing such as giving feedback, switching speakers, and asking clarifying questions were much reduced in the low-quality system compared with face-to-face interaction. Given the half-duplex audio and lags, speakers were unable to time their conversational contributions, with the result that backchannels or interruptions arrived too late, or at inappropriate points in the conversation. As a consequence, people had to explicitly manage speaker switches and there was increased formality in handing over the conversational floor, using devices such as selecting the next speaker by name. The result of both decreased interactivity and increased formality was a "lecture-like" style of interaction, with conversational turns in the videoconference being three times as long as face-to-face ones, making the system only suitable for certain types of conversational tasks, such as information exchange, that do not require quickfire exchanges.

Similar results showing the impact of audio lags on conversational processes are reported elsewhere. Cohen (1982) compared communication processes in face-to-face communication with low-quality videoconferencing for a series of laboratory tasks. The system she investigated had a 705-msec lag in both video and audio to simulate the performance of the AT & T PicturePhone. Participants found it hard to switch speakers and hard to ask clarifying questions in videoconferences. There were twice as many speaker switches in face-to-face communication compared with the videoconferencing system, and many more interruptions. Tang and Isaacs also evaluated low-quality videophone and videoconferencing systems (Tang & Isaacs, 1993; Isaacs & Tang, 1993). They found that lagged audio is highly disruptive of turn-taking, producing many fewer, longer turns. Their study also provides strong subjective support for the importance of low-lag audio. Participants preferred to use a separate half-duplex speakerphone to reduce delays in audio, even though it meant that synchronization between audio and video information was lost.

Connection Hypothesis: Using Video to Provide Availability Information

The second hypothesis is that video provides availability information about the movements and interruptibility of coworkers. This visible environment information can facilitate connection for unplanned interaction. Two separate classes of video application have systematically tested this hypothesis: (a) glance, which enables a user to briefly "look into" the office of a coworker to assess their communication availability, and (b) open links, in which persistent video/audio channels are maintained between two separate physical locations. There are again methodological problems in drawing conclusions about the video for connection hypothesis. In wide-area connection applications, video quality is poor. There may therefore be less motivation for using video for achieving wide-area connection, when the ensuing conversation will be over low-quality video. In local-area applications, visual connection information about coworkers' availability may already be accessible, as people move around their workplace. These confounding factors may lead to reduced use of video for connection. Nevertheless, when people do choose to use the technology for assessing availability, we can still ask how successful the technology is in achieving connection, and we now turn to this data.

For a local-area system, Fish et al. (1993) tested the use of different types of glance and their differential success in promoting opportunistic interactions. A brief glance at a user-selected recipient was the most frequently chosen type of glance: 81% of user initiated interactions were of this type, with 54% of these leading to an extended conversation. All other modes of glances were much less frequent and had much reduced likelihood of resulting in conversations. One type of glance was intended to simulate chance meetings such as "bumping into" another person in a hallway. In face-to-face settings neither participant normally intends such encounters, but they can promote extended work-related conversations. These types of chance encounters were implemented as a system-initiated connection between two arbitrary participants. These system-initiated connections showed very high failure rates, with 97% being terminated immediately without conversation. Overall, the glance options that callers chose indicate that they want direct control over who they connect to, and when they connect, rather than having the system do this. Furthermore, people wanted to use the "glance" as a preparation for communication, not merely to know "who is around." Glances that allowed "looking into" another office without the option of communicating were an infrequent user choice, accounting for only 12% of user selected glances.

The relationship between glances and opportunistic communication was also explored by Tang et al. (1994) for a system operating across multiple sites in a local area. Participants could first "look into" the office of a remote coworker, with the option of converting this into an extended conversation. Altogether, only 25% of glances were converted into conversations. This is no better than connection rates using only the phone (Whittaker et al., in press). Why was successful connection so infrequent? A significant proportion of failures (38%) occurred when recipients were out of their offices, but the reasons for the remainder are unclear; only 4% were when the recipients explicitly signaled that they were unavailable for communication. Many of the other failed connection attempts may occur when the recipient is in the office but busy with another activity, or another person. Tang et al. did not report this data, however.

Video and audio can also be used to support continuously "open links" between the offices of remote collaborators (Fish et al., 1993; Heath & Luff, 1991; Mantei et al., 1991). This is intended to approximate sharing the same physical office, so opportunistic communications can be started with minimal effort between connected participants, and visual and auditory information about communication availability is persistently available. However, Fish et al. (1993) reported that only 5% of connections lasted more than 30 min, and Tang et al. (1994) reported that only five interactions (out of a possible 233) lasted more than 30 min. Thus, both sets of usage data suggest brief interactions, rather than open links, are the main uses of the system. Open links can also be constructed between public areas of geographically separated sites (Abel, 1990; Bly et al., 1993; Fish et al., 1990). Cameras can be installed in common areas, transmitting images of people at remote sites, so that people can see, for example, who happens to be in the coffee area of a remote site. This is intended to promote opportunistic conversations of the type that can occur when people meet in public areas of the same site. Field trials report frequent use of open links for social greetings or "drop-ins'' between remote sites, with 70% of open link usage being of this type (Abel, 1990; Bly et al., 1993). Clearly, these brief interactions would have been unlikely to occur in the absence of the system. The use of the open link was mainly limited to these brief social exchanges, however, and the link was seen by the users as being ineffective in supporting work (Fish et al., 1993). Fish et al. (1990) also examined how often extended verbal communications resulted from sighting someone over the videolink. They compared this with the likelihood of interaction following face-to-face sightings, and found that sightings over a videolink were less likely to convert to extended conversation than face-to-face sightings.

Taken together, these preliminary results on glance and open links indicate a lack of evidence for the utility of video for connection: (a) Failure rates with glancing are as high as with phone alone; (b) open links are an infrequently chosen user option; (c) open links are less likely to promote conversation than face-to-face sightings; and (d) open links between common areas are not adequate to support work. These failures may, however, be due to confounding factors in the evaluations, or to implementation problems, such as lack of support for overriding or interrupting an existing open link (Whittaker, 1996).

Video-as-Data: Video Provides Information About the Visible Environment

An alternative hypothesis is that a major benefit of video lies in its ability to depict complex information about dynamic 3D shared work objects, rather than images of the participants themselves. This approach is partially motivated by finding that participants spend more time looking at relevant work objects than other people (Argyle & Graham, 1977). Thus, the video image can be used to transmit real-time information about work objects, and this can then be used to coordinate conversational content among distributed teams, by creating a shared physical context. The example discussed here is remote surgery, but other tasks such as concurrent engineering, or training also have similar requirements (Egido, 1990; Nardi et al., 1993; Nardi et al., 1996; chapter 23, this volume).

This work is discussed in detail in chapter 23, but to summarize, four different types of communicative use of the image were found. First, the dynamic image of the surgeon's actions allowed detailed coordination of interleaved physical action between the assisting nurse and the surgeon in the operating theater. By monitoring the surgeon's actions, via a shared video image viewed through the microscope, the nurse could anticipate the surgeon's requirements and provide the correct surgical instrument, often without it being directly requested. A second communicative function of the video image was that it served to disambiguate other types of surgical data that were supplied to remote consultants, such as neurophysiological monitoring data. The interpretation of these neurophysiological data depends critically on precise information about the physical actions that the surgeon is currently executing, such as the exact placement of a surgical clamp or the angle and direction of entry of a surgical instrument. Without the video image depicting these actions, the remote consultant had to rely on verbal reports from those who were present in the operating theater, and the inadequacy of the descriptions meant that the consultant often had to resort to physically visiting the operating theater to observe the actions directly. Third the video image served as a physical embodiment of progress through the operation. Members of the team who were involved in multiple operations at different locations and also those within the operating theater could see the current stage of each operation by inspecting the physical image, and observing what stage of the procedure the surgeon was at. The remote consultants could thus coordinate their visits to each operating theater accordingly, so as to arrive at times when their physical presence was critical.[7] Finally, the image was used for learning and education. The application was installed in a teaching hospital that undertook innovative surgical procedures. Academic visitors and trainees would often come to the operating theater to observe the novel procedures on the large monitor in the operating theater as they occurred. Some surgeons also recorded these procedures, to use them as aids in teaching classes.

Similar arguments for this use of "video-as-data" were made by Gaver et al. (1993). They looked at the use of images of 3D objects in design tasks. Users could choose between a number of different images, including between an image of the other participant, and various views of the object under study. Participants rarely chose facial views of their coparticipant (11% of the time), and "mutual gaze" (where both participants were simultaneously viewing each other) occurred only 2% of the time. Instead, people were much more likely to choose an image of the object, spending 49% of their time with the object views. This shows that for this class of design application, information about gaze and gesture of the other conversational participant seems to be less important than information about the shared physical context. An extensive research program has also been executed by Ishii, who has built a series of prototypes that use video to combine a semireflective writing surfaces with images of the participant's upper bodies. This enables the fusing of an image of another participant onto the work surface itself, making it possible to see both participant and object simultaneously and hence accurately track visual attention, while writing or manipulating the object (Ishii & Kobayashi, 1992). Again, a major focus of the work is that crucial collaborative information is embodied in the work object, although systematic evaluation of the benefits of adding this attentional information has not yet been conducted.

CONCLUSIONS

We have provided a framework for identifying potential functions of visible information in communication, reviewed evidence for three different hypotheses about the role of video in interpersonal communications, and identified outstanding research and design issues. With the exception of tasks that require access to affective information, we found that evidence for the nonverbal communication hypothesis is not strong, with few task outcome and process differences being found between audio and video-enhanced communications.[8] Furthermore, despite the absence of compelling evidence for the nonverbal communication hypothesis, certain current implementations may have compromised overall system utility by focusing on video at the expense of providing full duplex, low-lag audio. Failing to provide this type of audio information disrupts conversational processes that require precise timing and bidirectionality (O'Conaill et al., 1993).

Nevertheless, methodological and theoretical questions remain about the nonverbal communication hypothesis. We need to refine the hypothesis, so that more specific predictions can be tested and better systems designed. Visible information changes the outcome of tasks depending on affect or emotion, supporting the social cuing hypothesis. Neither process nor cognitive cuing accounts are well supported, however. For cognitive cuing, even face-to-face communication is no better than speech only, and even high-quality video cannot replicate the conversational processes of face-to-face communication (O'Conaill et al., 1993; chapter 6, this volume; Sellen, 1995). We therefore need to understand why even high- quality audio and video do not replicate face-to-face processes. One possibility is that current systems do not accurately simulate the presentational aspects of face-to-face interaction; spatial audio and video may therefore be needed to replicate conversational processes (O'Conaill et al., 1993; Sellen, 1995). However, there are other possible explanations that also need to be tested.

Another possible explanation of the results on the nonverbal communication hypothesis is that certain types of information are substitutable across different conversational media, whereas others are not. Thus in face-to-face communication, cognitive and process information is partially transmitted by head nods, eye gaze, and head turning. However, data on the efficacy of speech-only communication indicate that cognitive and process information can also be communicated effectively by other nonvisual cues (Walker, 1993; Walker & Whittaker, 1990). In contrast, the removal of the visual channel changes the outcome of tasks that require access to affect suggesting affective information is not substitutable. Part of the reason might be that affective cues are often not generated intentionally, so that although speech can signal affect, speakers omit the full range of affective cues when using audio-only communication. Future theoretical work should address this issue of the substitutability of different media and information types, and the role of intentional cuing. Another unresolved problem concerns inconsistencies between subjective and objective measures: Although outcome and process show few differences between audio and video conversations, people reliably prefer video-mediated communications (Fish et al., 1993; Isaacs & Tang, 1993; Tang & Isaacs, 1993). One possibility is that subjective preferences are an aspect of social cuing, but the social cuing account must be clarified for this argument to be sustained.

The connection hypothesis has yet to be systematically tested. The putative connection function of providing availability information for the process of conversation initiation is therefore undemonstrated. Although workplace studies show the importance of opportunistic communications, it is currently unclear how well video can support their initiation. One problem is the methodological limitations of current studies. Evaluation work needs to focus more on situations in which there is a critical mass of users who are geographically remote: Early evaluations have suffered from only investigating small user populations who often share the same physical space. Other design factors such as long delays in initiating communication, style of initiation, and, most importantly, privacy issues also have to be addressed before we know about the effectiveness of video for connection (Tang et al., 1994; Whittaker, 1996; Whittaker et al., 1994). Work should also be done to investigate whether alternative technologies, such as active badges (Pier, 1991), could also supply availability information and hence substitute for visual information. There is also the question of the extent to which other asynchronous technologies can partially substitute for opportunistic meetings. Can a brief e-mail or voicemail message replace a short synchronous discussion and hence reduce the need for remote opportunistic meetings (Whittaker et al., in press)?

Finally, video-as-data is a promising area, where more applications should be built and evaluations conducted. Much early work on video has neglected the importance of shared objects as part of a shared context. Given the lack of clear support for nonverbal communication, video-as-data may be a more successful use of video if we can identify tasks that are focused on complex dynamic 3D objects. Recent work on the nonverbal communication hypothesis also indirectly offers support for shared objects and a shared environment. For desktop videoconferencing applications, the presence of a shared workspace improves cognitive problem solving (Olson et al., 1995). However, as with opportunistic connection, there are also outstanding social issues about privacy and access that have yet to be addressed for "video-as-data."

Overall, this chapter suggests that the role of visible information and the successful application of video technology for interpersonal communications still require extensive research. Rather than the single function of broadening communication bandwidth implied by the nonverbal communication hypothesis, we need to extend the set of hypotheses we entertain about video, to think about video for initiating opportunistic communication and representing shared objects. The work reviewed here also suggests that the benefits of video are task and situation specific. Future research must explain when and why this technology brings benefits to interpersonal communication.

NOTES

1. Of course there are other indicators of presence. One can often infer presence from hearing another person, or from hearing others talking to them.
2. Many other recent field trials have investigated videophones, open links, and media spaces (Abel, 1990; Bly et al., 1993; Gaver et al., 1992; Mantei et al., 1991; Tang et al., 1994), but few of these studies have explicitly addressed the enhanced audio hypothesis. Instead, their focus has either been on the technical feasibility of building distributed video systems or alternatively on discovering novel uses of video applications such as video for connection. We review these novel applications in the next section.
3. The ability to interrupt the speaker at any point of the conversation, such as to ask a clarifying question, is regarded as a positive aspect of conversation, indicating spontaneous speaker switching (O'Conaill et al., 1994; Rutter & Robinson, 1981; Sellen, in press; Walker & Whittaker, 1990; Whittaker & Stenton, 1988).
4. Examples are "isn't it?," "aren't they?," "couldn't you?," and involve an auxiliary verb and question syntax, at the end of a sentence.
5. Exact lags depend on the system and network, but typical figures for one-way lags are 705 msec for the PicturePhone system (Cohen, 1982), between 410 and 780 msec for an ISDN system operating between the United States and the United Kingdom (O'Conaill et al., 1993; Whittaker & O'Conaill, 1993), and 570 msec for an ISDN system operating from coast to coast in the United States (Tang & Isaacs, 1993; Isaacs & Tang, 1994).
6. Half-duplex audio only allows unidirectional transmission of audio. This prevents certain key conversational processes that depend on multiple participants at different ends of an audio link being able to speak simultaneously, for example, backchannels to provide feedback to the speaker, or interruptive clarifying questions.
7. This function is similar to using video for connection, in that video information is used to coordinate a communication episode between people at remote locations.
8. Although more recent work with very-high-quality directional video may show small differences (Olson et al., 1995).

REFERENCES

Abel, M. (1990). Experiences in an exploratory distributed organization. In J. Galegher, R. Kraut, & C. Egido (Eds.), *Intellectual teamwork* (pp. 489–510). Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, A., Bard, E., Sotillo, C., Newland, A., & Doherty-Sneddon, C. T. (in press). Limited visual control of the intelligibility of speech in face-to-face interaction. *Perception and Psychophysics.*

Argyle, M. (1990). *Bodily communication* London: Routledge.

Argyle, M., & Graham, J. (1977). The Central European experiment: Looking at persons and looking at things. *Journal of Environmental Psychology and Nonverbal Behaviour, 1,* 6–16

Beattie, G. (1978). Sequential patterns of speech and gaze in dialogue. *Semiotica, 23,* 29–52.

Beattie, G. (1981). A further investigation of the cognitive interference hypothesis of gaze patterns in conversation. *British Journal of Social Psychology, 20,* 243–248.

Birdwhistell, R. (1970). *Kinesics and context: Essays in body motion communication.* Harmondsworth: Penguin.

Bly, S., Harrison, S., & Irwin, S. (1993). Media spaces: Bringing people together in a video, audio and computing environment. *Communications of the ACM, 36,* 28–45.

Bull, P. (1978). The interpretation of posture through an alternative method to role play. *British Journal of Social and Clinical Psychology, 17,* 1–6.

Cassell, J., McNeill, D. & McCullough, K. E. (in press). Speech gesture mismatches: Evidence for one underlying representation of linguistic and non-linguistic information. *Cognition.*

Chapanis, A. (1975). Interactive human communication. *Scientific American, 232,* 34–42.

Chapanis, A., Ochsman, R., Parrish, R., & Weeks, G. (1972). Studies in interactive communication: The effects of four communication modes on the behavior of teams during cooperative problem solving. *Human Factors, 14,* 487–509.

Clark, H., & Brennan, S. (1991). Grounding in communication. In L. B. Resnick, J. Levine, & S. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: APA.

Clark, H., & Marshall, C. (1981). Definite reference and mutual knowledge. In A. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge: Cambridge University Press.

Clark, H., & Schaefer, E. (1989). Contributing to discourse. *Cognitive Science, 13,* 259–292.

Cohen, K. (1982). Speaker interaction: Video teleconferences versus face-to-face meetings. *Proceedings of teleconferencing and electronic communications* (pp. 189–199). Madison: University of Wisconsin Press.

Cooper, R. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology, 6,* 84–107.

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

Efron, D. (1972). *Gesture, race and culture.* The Hague: Mouton.

Egido, C. (1990). Teleconferencing as a technology to support co-operative work: A review of its failures. In J. Galegher, R. Kraut, & C. Egido (Eds.), *Intellectual teamwork* (pp. 351–372). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ekman, P., & Friesen, W. (1975). *Unmasking the face.* Englewood Cliffs, NJ: Prentice Hall.

Exline, R., & Winters, L. (1965, April). Effects of cognitive difficulty and cognitive style on eye contact in interviews. *Proceedings of the Eastern Psychological Association.* Atlantic City, NJ.

Fish, R., Kraut, R., & Chalfonte, B. (1990). The Videowindow system in informal communications. *Proceedings of Conference on Computer Supported Cooperative Work* (pp. 1–12). New York: ACM.

Fish, R., Kraut, R., Root, R., & Rice, R. (1993). Video as a technology for informal communication. *Communications of the ACM, 36,* 48–61.

Frohlich, D. Requirements for interpersonal information management. In P. Thomas (Ed.), *Mobile personal communication and co-operative working* (pp. 35–65). London: Alfred Waller.

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

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

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 225–242). New York: Academic Press.

Grosz, B., & Sidner, C. (1986). Attentions, intentions and the structure of discourse. *Computational Linguistics, 12,* 175–204.

Heath, C., & Luff, P. (1991). Disembodied conduct: Communication through video in a multimedia environment. *Proceedings of CHI '91 Human Factors in Computing Systems* (pp. 99–103). New York: ACM.

Isaacs, E., & Tang, J. (1993). What video can and can't do for collaboration: A case study. In *Proceedings of the ACM Multimedia 93 Conference* (pp. 199–206). Anaheim, CA.

Ishii, H., & Kobayashi, M. (1992). Clearboard: A seamless medium for shared drawing and conversation with eye contact. In *Proceedings of CHI '92 Human Factors in Computing Systems* (pp. 525–532). New York: ACM.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26,* 1–47.

Kendon, A., & Ferber, A. (1973). A description of some human greetings. In R. Michael & J. Crook (Eds.)., *Comparative ecology and behaviour of primates* (pp. 591–668). London: Academic Press.

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

Kleinke, C. (1986). Gaze and eye contact: A research review. *Psychological Bulletin, 100,* 78–100.

Kraut, R., Fish, R., Root, B., & Chalfonte, B. (1993). Informal communication in organizations. In R. Baecker (Ed.), *Groupware and computer supported co-operative work* (pp. 287–314). San Mateo, CA: Morgan Kaufman.

Kraut, R., Lewis, S., & Swezey, L. (1982). Listener responsiveness and the co-ordination of conversation. *Journal of Personality and Social Psychology, 43,* 718–731

Levinson, S. (1983). *Pragmatics.* Cambridge: Cambridge University Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 126–130.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* Chicago: University of Chicago Press.

Mantei, M., Baecker, R., Sellen, A., Buxton, W., Milligan, T., & Wellman, B. (1991). Experiences in the use of a media space. *Proceedings of CHI '91 Human Factors in Computing Systems* (pp. 203–209). New York: ACM.

Nardi, B., Schwarz, H., Kuchinsky, A., Leichner, R., Whittaker, S., & Sclabassi, R. (1993). Turning away from talking heads: An analysis of "video-as-data." *Proceedings of CHI '93 Human Factors in Computing Systems* (pp. 327–334). New York: ACM.

Nardi, B., Kuchinsky, A., Whittaker, S., Leichner, R., & Schwarz, H. (1996). "Video-as-data": Technical and social aspects of a collaborative multimedia application. *Computer Supported Co-operative Work, 4,* 73–100.

Noll, M. (1992). Anatomy of a failure: PicturePhone revisited. *Telecommunications Policy,* May/June, 307–316.

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.

Olson, J., Olson, G., & Meader, D. (1995). What mix of video and audio is useful for small groups doing remote design work? *Proceedings of CHI '95 Human Factors in Computing Systems* (pp. 362–368). New York: ACM.

Pier, K. (1991). Active badge panel. In *Proceedings of Conference on Organisational Systems,* Atlanta, GA.

Reid, A. (1977). Comparing the telephone with face-to-face interaction. In I. Pool (Ed.), *The social impact of the telephone* (pp. 386–414). Cambridge, MA: MIT Press.

Rutter, R., & Robinson, R. (1981). An experimental analysis of teaching by telephone. In G. Stephenson & J. Davies (Eds.), *Progress in applied social psychology* (pp. 143–178). London: Wiley.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language, 50,* 696–753.

Schegloff, E., & Sacks, H. (1973). Opening up closings. *Semiotica, 7,* 289–327.

Sellen, A. J. (1995). Remote conversations: The effects of mediating talk with technology. *Human–Computer Interaction, 10*(4), 401–444.

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

Summerfield, Q. (1992). Lipreading and audiovisual speech perception. *Philosophical Transactions of the Royal Society of London, B335,* 71–78.

Tang, J., & Isaacs, E. (1993). Why do users like video: Studies of multimedia-supported collaboration. *Computer Supported Cooperative Work, 1,* 163–196.

Tang, J., Isaacs, E., & Rua, M. (1994). Supporting distributed groups with a montage of lightweight interactions. *Proceedings of Conference on Computer Supported Cooperative Work* (pp. 23–34). New York: ACM.

Walker, M. (1993). *Information redundancy in dialogue.* Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia.

Walker, M., & Whittaker, S. (1990). Mixed initiative in dialogue. In *Proceedings of 28th Annual Meeting of the Conference on Computational Linguisitics* (pp. 70–78). Morristown, NJ: ACM.

Whittaker, S. (1995). Video as a technology for interpersonal communication: A new perspective. *IS&T SPIE Symposium on electronic imaging science and technology, 2417,* 294–304.

Whittaker, S. (1996). Rethinking video as a technology for interpersonal communication. *International Journal of Human–Computer Studies, 42,* 501–529.

Whittaker, S., Brennan, S., & Clark, H. (1991). Co-ordinating activity: An analysis of computer supported co-operative work. In *Proceedings of CHI '91 Human Factors in Computing Systems* (pp. 361–367). New York: ACM.

Whittaker, S., Frohlich, D., & Daly-Jones, O. (1994). Informal workplace communication: What is it like and how might we support it? In *Proceedings of CHI '94 Human Factors in Computing Systems* (pp. 130–137). New York: ACM.

Whittaker, S., Geelhoed, E., & Robinson, E. (1993). Shared workspaces: How do they work and when are they useful? *International Journal of Man-Machine Studies, 39,* 813–842.

Whittaker, S., & Stenton, P. (1988). Cues and control in expert client dialogues. In *Proceedings of the Conference for the Association for Computational Linguisitics* (pp. 123–130). Cambridge, MA: MIT Press.

Whittaker, S., Swanson, J., Kucan, J., & Sidner, C. (in press). Telenotes: Managing lightweight interactions in the desktop. *Transactions on Computer–Human Interaction.*

Williams, E. (1977). Experimental comparsions of face-to-face and mediated communication. *Psychological Bulletin, 84,* 963–976.

Yngve, V. (1970). Getting a word in edgewise. In *Proceedings of the Sixth Meeting of the Chicago Linguistics Society* (pp. 567–577). Chicago, IL: Chicago Linguistics Society.