

6 ACKNOWLEDGEMENTS

Thanks to Erik Geelhoed for assistance in statistical analysis, to Hewlett Packard Laboratories Bristol for providing facilities and a grant to support this work, to Richard Beckwith of University College, London for providing a grant, and to the people who participated in the meetings and kindly allowed themselves to be recorded and interviewed. Thanks also to Lyn Walker, Herb Clark and David Frohlich for discussions of these ideas.

needed for this type of solution, most specifically to identify the audio requirements and the effects of lack of synchronisation (CCITT, 1988; Shah, Staddon, Rubin, & Ratkovic, 1992).

Finally this work contributes to a developing theory of mediated communication. Other work has shown that the organisation of mediated communication is critically dependent on the properties of the communication channels (Whittaker, Brennan, & Clark, 1991; Whittaker, 1992). Previous research has explained this class of result in terms of concepts that are difficult to operationalise such as “social presence” (Short, et al., 1976), “media richness” (Daft & Lengel, 1984) or “cuelessness” (Rutter & Robinson, 1981). Here we were able to test predictions derived from an analysis of face-to-face interaction, about how certain channel properties influenced specific characteristics of speaker and listener behaviour. Although our results were not entirely consistent with our initial hypotheses, this helped us both to identify a further potentially important channel property as well as refining our understanding of the relationship between channel properties and communication characteristics. The new channel property is that of directionality, which we argue mainly impacts **speaker** behaviours. In its absence, speakers tend to show increased formality and explicitness in managing turn-switching. A combination of our initial channel properties of lags, half-duplex audio and poor quality video contribute to what seemed to be reductions in the spontaneity of **listener** behaviour, with lower levels of listener participation, and more “lecture-like” speech resulting. Further work should test these more specific hypotheses about the relationships between these channel properties and communication characteristics.

to choose what they want to see and hear rather than have these choices made for them. They also suggested the use of several monitors. One monitor could be used to provide a high quality image of the speaker or object of interest, and other monitors could then present lower quality panoramic images of the remaining remote participants for visual context.

Another approach to improving video systems is to examine the use of the video image for things other than pictures of participants' head and shoulders. Elsewhere we suggest an alternative novel application of video in the notion of "video as data" for remote microsurgery (Nardi, Schwarz, Kuchinsky, Leichner, Whittaker, & Sciabassi, 1992). We claim that there may be many situations in which video is best used to transmit images of the work itself, rather than of the participants who are carrying out the work. Another application might be the "Open Distributed Office" in which video is used to give people in distributed teams monitoring information about whether remote collaborators are present or absent (Dourish & Bly, 1991; Fish et al., 1992; Mantei et al., 1991). This contrasts with other applications of video, because it stresses the benefits of video for background awareness instead of solely for direct communication.

Another short-term way to improve communication is to allow different trade-offs between audio and video in the limited bandwidth. We might improve communication by relaxing the requirement for **synchronised** audio and video, and reducing the bandwidth allocated to video. Studies have reported the importance of audio as compared with visual information in this type of application (Chapanis, 1975; Reid, 1977; Williams, 1977). Consequently removing the requirement for synchronisation would allow the reduction of audio lag, because there is less audio than video data to be compressed. User studies are required to determine just what quality of audio and video are

negotiation, where the ambiguity of the information and the requirement for rapid clarification and feedback are critical for the success of the interaction (Daft and Lengel, 1984; Whittaker, 1992).

If ISDN cannot effectively support these tasks, this may contribute to the lack of success of this quality of video-conferencing. It may be that future remote collaborators have to choose appropriate communication technologies for the task at hand and ensure that certain types of task, e.g. conflict resolution and negotiation, are resolved in face-to-face situations. Naturalistic studies of remote collaborators who are using multiple technologies should be conducted to determine how people currently allocate technologies to communication tasks, and more theoretical work is needed to specify the relationship between communication task requirements and the basic communication processes that are needed to support them.

What are the practical implications of these results? First it would seem that introducing low lag, full duplex channels will lead to improvements in communication, as evidenced by the superiority of LN over ISDN. This suggests that we should continue to work on high speed wide-area networks and compression technology to reduce the disruptions to communication described above. However the LN results also suggest that improving these properties alone will not exactly reproduce face-to-face interaction.

How might we improve video systems, in addition to improving networks and compression? One possibility is the implementation of directional audio and video which might address the outstanding differences between LN and FTF (Sellen, 1992). A second strategy would be to modify existing video-conferencing systems by acting upon our users' comments. They suggested providing remote audio and video controls, so that remote participants are able

visual information, lags and half-duplex audio can all independently produce these types of effects (Cohen, 1982; Krauss & Bricker, 1967; Rutter & Stephenson, 1977).

A study of low lag, full duplex video-conferencing by Sellen (1992) is similar to our LN and FTF comparisons. Sellen stressed the effects of video mediation on listener behaviour and concluded: "it is as if conversants in video-mediated conversations were more opportunistic or polite, waiting for a pause or for a speaker to finish before attempting to take the floor" (p57). We found some changes in listener behaviour in LN: the number of backchannels was reduced, although there were equal numbers of interruptions in FTF and LN. However we found that in our study the main differences between LN and FTF seemed to be attributable to changes in **speaker** behaviour with greater use of handovers and reduced floorholding.

The current work shows that certain basic communication processes are disrupted by the channel properties of the two mediated communication systems. Due to the fact that we observed real meetings with naturalistic data, we were unable to measure directly the overall effectiveness of communication in the different conditions. Other research has shown that characteristics such as backchannels and interruptions are related to task outcome (Kraut et al., 1982; Oviatt and Cohen, 1991). Laboratory studies are needed to measure the effects on task outcome of disrupting the different conversation characteristics under more controlled conditions.

Although we cannot judge overall quality, there may be implications about the kinds of tasks for which the current ISDN quality is appropriate. The "lecture-like" character and the inability to support quickfire exchanges could mean that ISDN is unsuitable for tasks such as conflict resolution, planning or

For other characteristics ISDN was different from both LN and FTF, as we expected. It therefore seems that changes in interruptions, projections, simultaneous starts, and turn size and length, resulted from lagged half-duplex audio and poor picture quality. These channel properties seemed to lead to changes in listener behaviours. Being conscious of the disruptive effects of lag and half-duplex audio, listeners wait for the previous speaker to finish before taking the floor. The effect of reduced listener participation is to decrease the number of speaker switches and hence produce fewer but longer turns.

Backchannels seem to be affected by all the channel properties, with FTF being different from LN, which in turn differs from ISDN. It may be that listeners are aware of the disruptive effects of backchannels with half-duplex, lagged audio and this may account for differences between ISDN and LN. The difference between LN and FTF may arise because speakers rely on directional gaze in FTF to elicit backchannels, and this cue is removed in both LN and ISDN.

Finally, there were conversational characteristics that seemed to be unaffected by channel properties, such as tagging and turn distribution. It may be that these are reflexive conversational behaviours produced independently of the communication situation.

We could not, however isolate whether audio lag, half-duplex audio or video quality was mainly responsible for the disruptions in ISDN. This was because we attempted to gather data for real systems for which these properties were not independent. Other laboratory work should be done to confirm which of the channel properties of the ISDN system was most disruptive of these conversation characteristics. Currently we cannot rule out any of these channel properties, and other research has independently shown that poor quality

responsible. Our predictions about differences were only met for certain characteristics, namely interruptions, projections, simultaneous starts, turn size and turn frequency. Here we observed differences between ISDN and both LN and FTF, with LN and FTF equivalent. The implication is therefore that differences in these characteristics are attributable to lags, half-duplex audio and poor video quality. In contrast, other characteristics such as floorholding, and handovers showed equivalence between LN and ISDN.

Figure 11 about here

How can we explain these similarities between ISDN and LN and what channel properties are responsible? ISDN and LN are similar because they both have non-directional sound and vision from a restricted number of sources, i.e. one or two monitors and loudspeakers. Both systems contrast with FTF where sound and visual behaviour are directional, because they emanate from the different participants.

Other research has shown that head-turning and eye-gaze play an important role in speaker switching (Duncan, 1972), and both these behaviours are reliant on directionality. Its absence in ISDN and LN may lead to changes in speaker behaviour, with speakers having to use the verbal channel to signal turn transitions explicitly and carefully manage speaker switches. This may explain the increased incidence of questions in ISDN and LN, and the reductions in floorholding. One study has directly addressed the impact of directionality on conversations in video-conferences (Sellen, 1992). She found few objective effects for directionality, but this may have been due to the small image size employed, and more work should be done to test this.

“lecture” style of interaction, with long turns, handed over by a very deliberate process.

In LN, even where there is a full duplex line, immediate transmission and broadcast quality image, the properties of the spoken communication still differ from face-to-face interaction:

- Although listeners interrupt as frequently as in FTF, they are less likely to give backchannels.
- Speakers use questions to formally hand over the floor more frequently, and they are also less likely to hold the conversation floor with redundant information.

Thus although LN was similar to FTF it was still characterised by highly formal conversational behaviours.

How can we explain these findings? Our initial claim was that certain key channel properties of ISDN disrupt basic communication processes. Face-to-face interaction has full duplex, almost instantaneous transmission of audio as well as high quality visual information. As we expected, when we change these channel properties, to those of the ISDN system, we produce a style of interaction that is lecture-like and lacks spontaneity. However the argument that these channel properties are **solely** responsible for communication disruption must be re-examined in the light of the LN data, because for several conversation characteristics, LN was more like ISDN than FTF. This suggests that other channel properties are also critical here and the account should be extended to include these properties and determine which conversation characteristics they impact.

Figure 11 addresses this question. It depicts where differences occurred between ISDN, LN and FTF and which channel properties might be

People also complained that in ISDN small movements are not picked up and that sudden movements appear jerky and blurred. The movements were described as “puppet like” by one user. Some participants also reported attempting to compensate for poor quality image by using exaggerated gestures like nodding and the shaking of heads to substitute for their inability to provide verbal feedback.

5 CONCLUSIONS

Many reasons have been put forward for the failure of video-conferencing to gain widespread acceptance, including cost, incorrect marketing and the questionable value of a video channel. There have been few detailed empirical studies of the actual communication that occurs over real implementations of wide area video-conferencing systems. By examining how the characteristics of two such systems affect the nature of spoken conversation, we aimed to identify possible reasons for the lack of success of video-conferencing technology. We also sought to explain **why** channel properties affected conversational processes in the way they did.

Our results showed that compared with FTF, spoken conversation patterns are disrupted over ISDN with its half duplex line, transmission lags and poor quality image:

- Listeners produced fewer backchannels and interrupted less.
- Listeners were also less likely to anticipate turn endings.
- Speakers also alter their behaviour, being more likely to hand over turns formally using a question or naming the next speaker. They were also less likely to hold the floor with redundant phrases.
- The result of listeners reducing interruptions and speaker feedback, combined with the general difficulty of switching speakers, was a formal

speak. Then when the opportunity does arise, I don't take it because my comment often isn't relevant anymore..." In contrast one LN user acknowledged the greater formality of LN meetings compared with face-to-face but said that she sometimes exploited this to hold the channel for longer periods.

On the other hand, despite the problems with the video-conferencing systems, people preferred these to audio-conferencing. The main stated advantages of video-conferencing were knowing who was at the remote location, and knowing who was speaking, although users' behaviour suggested this information was not always available in ISDN due to poor image quality. Another stated advantage was the feeling of "not talking into a void". Our users also commented that they found video-conferencing appropriate for only certain types of meeting such as information exchange or project updates.

There were other problems that were specific to ISDN alone. It was at times difficult to identify the speaker at the remote location in ISDN; the quality of the visual information was poor and this seemed to reduce the impact of visual speaker cues such as leaning forward, increased gesturing and posture changes. Informal observations suggest that speaker identification often took several seconds, although there was one ISDN meeting where a participant spoke for several minutes and was still misidentified. Some groups in ISDN attempted to resolve this problem by panning the camera and focussing exclusively on the local speaker which solved the identification problem for the remote participants. Unfortunately this produced awkward transitions and further panning and focussing when someone at the same location spoke next. It also meant a narrowing of the visual field for the remote participants with the result that they only had visual information about the current speaker and not the other remote participants.

achieve in both video systems because participants look at the image of the remote participants and not directly into the camera. Furthermore, gaze behaviour in both video-conferences differs from FTF in extremely obvious ways. Participants tend to stare fixedly at the screen displaying the remote participants even when the speaker is local, and they therefore show none of the normal modulation of gaze behaviour and local speaker monitoring that is characteristic of face-to-face interaction (Duncan, 1972). Similar effects of “monitor capture” or “TV watching” are reported in Abel (1990). The result is that the speakers are presented with an array of remote people staring relentlessly almost directly at them. Speakers report finding this situation confrontational. It also means that speakers receive little local attentional feedback, because local listeners are staring only at the remote site.

Further problems relate to the control of the cameras. In both video-conferencing systems, camera control is local. This led to a number of interchanges which result from attempted changes to camera angles, because the local participants were unable to tell whether the displayed image they were presenting was adequate for the remote participants. This was a particular problem with the document camera, when the issue concerned the resolution of the presented document. Here participants would have discussions along the lines (“can you see it yet?”, “back a bit”, “is that okay?”, “back a bit more”). This type of fine tuning of the image was further hindered by the lags in ISDN which meant that feedback about the acceptability of the image was not timely.

People stated that video-conferences involved more “effort”. For example participants complained about the difficulty of assuming control of the conversation in both video-conferences. One participant reported for ISDN: “I have the feeling that I want to say something, but there’s no opportunity to

Finally we expected that the different conditions would lead to unequal distribution of turns between participants. We expected that in ISDN, participants would rely on two people, one at either end of the link, to manage interactions across the connecting link, and they would channel their responses through these people. However, when we examined the data for dominance by two speakers this was not the case (see Figure 10). We measured the number of turns that were produced by the two most frequent speakers in the three conditions⁶. There was no overall difference either in the percentage of turns taken by these people or in the number of words that they spoke. (Turns: $F(2,11) = 0.59$, $p > 0.05$; Words: $F(2,11) = 0.55$, $p > 0.05$). We also investigated whether ISDN served to **exclude** certain speakers: the fact that they were less able to interrupt might prevent participants who are not “chairpeople” from having the opportunity to speak. Again this hypothesis was not born out by our results. We looked at the number of words and the number of turns for the two people who spoke least. Again there were no differences: Turns ($F(2,11) = 0.75$, $p > 0.05$), Words ($F(2,11) = 1.32$, $p > 0.05$). This result is interesting because it runs contrary to the perceptions of the people using ISDN and LN. They report feeling both that certain participants are able to dominate the meeting and that others are less able to contribute to it.

Figure 10 about here

4.7 User comments and Informal Observations

No objective measures were taken of the use and effectiveness of non-verbal behaviour such as gaze and gesture in ISDN and LN, although there do appear to be differences from FTF. Our informal observations and comments made by the users show several apparent effects. First, mutual gaze is difficult to

6. These scores were normalised to allow for the fact that there were different numbers of people in each meeting

A: It's a bug fix
B: yes yes
A: Not a new functionality
B: I don't think so no There's also a new version of of
meta software (Etches) available
A: yes I know

The pattern is similar in FTF.

A: then this point that's the order
B: no [because some of the points are are implied
C: [you only give it two
A: [ah okay
B: [cos cos you know you're drawing a rectangle obviously you
only give [the two=
A:
[ahh right=
B: =corners so you don't give all four corners
A: =okay
A so so it's because, yea, so something like a circle

Both FTF and LN conversations have a “quickfire” character with clarifications taking place (LN, lines 3 and 5) and also disagreements (FTF, line 2), showing that participants are able to react quickly to incoming information when they do not understand or when they disagree.

4.6 Turn distribution

These results strongly support our prediction that ISDN would produce a “lecture-like” interaction with speakers holding the floor for lengthy uninterrupted monologues. In contrast in both LN and face-to-face we see many more short turns with higher frequency of interruptions and backchannels.

The following examples show typical interactions for ISDN, LN and FTF. The first clearly shows the “lecture-like” style in ISDN. Here speakers supply large amounts of uninterrupted information, with transitions often being accompanied by pauses, and there is little evidence of incremental checking of listener understanding.

A: ...and ah essentially what they are doing is they're ah comparing preoperative waves with with the actual interoperative ones they're looking at what the guy was like before they did anything to him to what he's like now ahm and its kind of you know they sort of look at this thing and they sort of say its its a bit different isn't it type of thing and your thinking yeah it is I suppose and and then they sort of say well actually I think I'll tell him but I I don't quite I don't I haven't quite got a grip on what the algorithm was they were sort of saying well it looks similar and look its sort of kind of moved that that way a bit ahm and that's how they were doing delays it was it was very approximate.

((pause))

B: Yeah I mean the two things that they seem to be looking at predominantly are latency over the preoperative signal and also some characteristics which we couldn't fathom which were like the shape of the waves you know something to do with peaks and you know like when they hit or you know how their characteristics changed and you know in some way that related to ahm you know the particular nerve that was being tested but...

In contrast, LN has many more short turns with conversational exchanges being incremental and interactive.

A: Is there any significant difference?

B: ahm there was a problem there was a mouse problem on two point one which occurred intermittently

Participants had other methods of explicitly handing over control in video-conferences. They were observed raising their hands as an indication of a desire to speak. In one ISDN conference participants agreed to use their hands throughout the meeting to indicate they wished to speak.

4.5 Turn Size

We predicted that the problems encountered in speaker transition, coupled with listeners' reluctance to interrupt or provide backchannels would result in longer turns in ISDN. Figure 9 shows the number of turns taken and their average word length. We analysed both the total number of turns on a meeting by meeting basis and also for each participant. Typically, the meetings held over ISDN were characterised by fewer turns of greater length. There were significantly fewer turns per participant in ISDN compared with LN and face-to-face ($F(2,86) = 6.48, p < 0.002$). (ISDN vs. FTF, $F(1,8) = 13.02, p < 0.001$, ISDN vs. LN, $F(1,7) = 5.68, p < 0.05$, FTF vs. LN, $F(1,7) = 2.27, p > 0.05$). The complementary result was that the number of words/turn was significantly greater in ISDN than in the other two media ($F(2,21) = 60.15, p < 0.001$). (ISDN vs. FTF, $F(1,8) = 101.66, p < 0.0001$, ISDN vs. LN, $F(1,7) = 62.82, p < 0.0001$, FTF vs. LN, $F(1,7) = 1.76, p > 0.05$). It is possible that these effects are due to the reduction of brief turns in ISDN. To investigate this we repeated the analysis excluding all turns of less than five words, but both effects were still present.

Figure 9 about here

These differences in turn size were observed despite the fact that there were no overall differences in the total number of words/meeting in each condition ($F(2,11) = 0.39, p > 0.05$): While the total number of words remained constant across conditions, the differences between the conditions lay in how the words were distributed across turns.

Handovers by naming the next speaker were more frequent in ISDN than in face-to-face ($F(2,11) = 4.09, p < 0.05$). (FTF vs. ISDN, $F(1,8) = 6.57, p < 0.05$, FTF vs. LN, $F(1,7) = 2.21, p > 0.05$, ISDN vs. LN, $F(1,7) = 2.39, p > 0.05$).

In the ISDN condition participants used questions at the end of long turns to encourage speaker transition, for example:

A: ...there are only two possible choices either there is an input file or there is none or rather either it is empty or not If it is if there is data in it then the job runs correctly otherwise all the subsequent steps test the condition code and if it is different from zero then they don't run as simple as that any ah ((pause)) any counter indication on your end?

In some instances names were used to address the question to a particular individual as in the following two LN examples:

A: how much does that cost Mike?
B: Are we still on state of play Alan?

Tagging such as “is that okay?” or “you know” or redundant information were equally frequent in all media. Here a participant in an ISDN conference ends a turn with a tag question which both facilitates speaker transition and acts as a check for understanding.

A: ...The only thing you have to change is ahh the the step card and thats it its one line in this JCL Have I made myself clear?

A: the appearance of [that
 B [just out of curiosity what
 differ- ence ()
 C: go ahead

To summarise, there are no differences in the combined number of Overlaps but the subtypes of Overlaps are differently distributed in the three conditions. Overlaps occur in face-to-face mainly because of projections and floorholding, in LN because of projections, and in ISDN because of simultaneous starts.

4.4 Explicit Handovers

We predicted that speakers would try to remedy the problem of speaker transition in ISDN by explicitly handing over the floor. Figure 8 shows turns ending in questions, tagging and naming of the next speaker. Again this was measured in terms of frequency per turn because of the different numbers of turns across conditions. As we predicted, there was a greater number of each of these formal handovers in ISDN compared with FTF, because of the need to explicitly manage speaker transitions ($F(2,11) = 9.46, p < 0.004$). Contrary to our expectations, however, we found the same overall pattern of formal handovers in LN as in ISDN. (FTF vs. ISDN, $F(1,8) = 14.70, p < 0.01$, FTF vs. LN, $F(1,7) = 38.22, p < 0.001$, ISDN vs. LN, $F(1,7) = 2.06, p > 0.05$). Again we discuss this unexpected result in the Conclusions section.

Figure 8 about here

Further analysis of the different classes of handover indicated that handovers using direct questions were more frequent in both video conferences ($F(2,11) = 13.14, p < 0.001$). (FTF vs. ISDN, $F(1,8) = 21.42, p < 0.001$, FTF vs. LN, $F(1,7) = 30.47, p < 0.001$, ISDN vs. LN, $F(1,7) = 0.21, p > 0.05$). Tagging was equal in all three conditions.

B: [Sorry we missed that
from communication

A: okay for the the communication protocol that
be...

In contrast an interruption during LN causes no problems for the speakers. No information is lost, so this does not need to be repeated and A simply drops out leaving B to take the floor.

A: because we have people actively using Omega
we have Beta both of which we would lose
[(and we)

B: [that's
a lot of money just to pay for those packages

4.3 Overlapping Speech

Overlaps were analysed in terms of their frequency per turn. This was to allow for the fact that there were many fewer turns and speaker switches in ISDN, and the chance of generating an overlap is clearly dependent on the number of speaker transitions. Figure 7 shows that the overall number of overlaps per turn did not differ substantially, $F(2,11) = 1.22, p > 0.05$). However, the different types of overlaps showed different distributions in the three conditions.

Figure 7 about here

For projections we found as we predicted there were differences between the conditions ($F(2,11) = 11.90, p < 0.002$) with more overlaps following projections in the face-to-face and LN media. The combination of half duplex and lags seem to combine to reduce projections in ISDN, with listeners avoiding overlapping speech even when this could assist the speaker in composing their message. Projections were reduced in ISDN

Figure 6 shows the distribution of backchannels and interruptions in the three conditions. Mean levels of backchanneling were low in ISDN compared with FTF (7.00 vs. 60.80), confirming our prediction that people in ISDN would avoid backchannels. The finding that backchannels were also reduced in LN compared with FTF (30.50 vs. 60.80) was not predicted and we discuss reasons for this in the Conclusions. The differences were analysed in a one way ANOVA. The overall difference was significant ($F(2,11) = 18.16, p < 0.001$), with backchannels being more frequent in face-to-face than LN ($F(1,7) = 15.82, p < 0.01$), which are in turn more frequent than in ISDN ($F(1,7) = 6.77, p < 0.05$)⁵.

Figure 6 about here

The example below indicates why backchannels were reduced in ISDN as shown in Figure 6. Where backchannels do occur, they can lead to a disruption of the flow of the speaker. In this instance B responded with a backchannel to A's comment "...it would be interesting to see if ah we could marry that...". Locally the backchannel was placed after the suggestion overlapping A's "because". However, because of the lag, A does not receive the backchannel until some words later leading him to hesitate ("ahh").

A: ... portion of the interface that's been put
there it would be interesting to see if ah we
could marry that [because that was the intent
of the **ahh** an original interrogation=
B: [mm

5. Posthoc ANOVA tests have been administered to make pairwise comparisons between the conditions following the recommendations of Kirk (1982).

The transcripts did not replace the tapes for scoring purposes, but were used in conjunction with the tapes. We also conducted a reliability analysis with two judges independently scoring two meetings in each condition, a total of 1054 turns in total. Both judges tried to identify every instance of backchannels, interruptions, overlaps, and formal handovers. Reliability scores, were measured as:

$$\frac{(\text{number of agreements} - \text{number of disagreements})}{(\text{number of agreements} + \text{disagreements})}$$

These were as follows: backchannels (0.91), overlaps (0.74), interruptions (0.62) and handovers (0.92). We also compared reliability of coding across the three conditions and found coding was most reliable for ISDN (0.89), and LN (0.88), but slightly less reliable for face-to-face interaction (0.79). To evaluate the success of our coding we computed Kappa (Cohen, 1960) for each condition. The respective Kappas for each condition across the four categories were: ISDN = 0.92, $p < 0.001$, LN = 0.93, $p < 0.001$, FTF = 0.86, $p < 0.001$. This indicates the reliability of our coding scheme.

4 RESULTS

4.1 Overview

In what follows we present statistical analyses for each prediction followed by representative examples from the interactions to illustrate our claims. All analyses apply to the 20 minute segment we analysed for each meeting and not to the whole meeting.

4.2 Backchannels and Interruptions

B: Oh [yes

A: [Tim I [ga I gave to Timmy oh it's circulating is it yeah [it
seemed it was=

B: [it's circulating

B: [()

A: =quite interesting ah

[A single left square bracket indicates the point of overlap

= Equal signs, at the end of one line and the beginning of the next
indicates no gap between the two lines

(.) A dot in parentheses indicates a tiny gap within or between
utterances

() Empty parentheses indicate the transcriber's inability to hear
what was said.

(word) Parenthesized words are especially dubious hearings.

(()) Doubled parentheses contain transcribers' descriptions

We recorded at one location only, so that assessments of simultaneous speech were analysed only with respect to that location, although as we will see, it is sometimes possible to determine the points at which simultaneous speech occur at the remote location.

are shown in Figure 5. This shows that all the meetings were co-operative in nature with their main function being to exchange information. Secondary functions and activities such as problem solving and idea generation also took place.

Table 5 about here

The FTF and ISDN meetings were to report progress where participants described the work that they had recently been doing. In some cases this involved the demonstration of software. These meetings centred around project teams with one or two project managers being present. The LN meetings were the coming together of representatives from different colleges. Participants from the various colleges gave updates on the developments and progress made at their site.

In the majority of cases, the participants knew each other before the meetings, although in a few of the video-conferences the people at either end of the link had not all previously met face-to-face. We could not control for certain parameters of familiarity, e.g. participants at either end of a video-conference link are likely to know each other better and have a greater understanding of local work. Where possible however we tried to reduce this problem by our choice of face-to-face meetings: two of the face-to-face meetings were between collaborators from the U.S. who were visiting the U.K., and therefore had little day to day contact. All participants were familiar with using video-conferences. As they already had experience with the systems, we did not expect participants' conversational strategies to alter significantly during the meeting. We therefore did not analyse whether conversational behaviours changed in the course of each meeting.

In the FTF and ISDN conditions there was a mixture of agenda and non-agenda based meetings. All the LN meetings were agenda based³. Both the FTF and ISDN meetings

3. The participants of course chose whether or not the meeting was agenda based.

having more turns than the average group member. In contrast we expected turns to be more evenly distributed in the LN meetings.

Prediction: Turns should be unequally distributed in ISDN, with two speakers, one at each location, producing more turns than other group members. In contrast, turns should be equally distributed in LN and face-to-face interaction.

3 METHODOLOGY

3.1 Recording Method

The ISDN video-conferences were recorded by placing a video-camera next to the monitor and camera stack in the conference room. An additional monitor displaying the remote participants was placed beneath the table at which the participants sat. The video camera thus captured the local participants with the remote participants visible on the monitor under the table. The stills screen was not monitored. The LIVE-NET meetings were recorded at the central video switch site. The picture on each of the two quadrant monitors was recorded on to video tape. The face-to-face, (FTF), meetings were audio taped. An observer was present at each meeting who noted any events not picked up on tape.

3.2 The meetings

Five ISDN video-conferences, four LIVE-NET meetings and five face-to-face meetings were recorded and analysed. All meetings were scheduled for work-related reasons and were not arranged for the study. We attempted to identify analogous groups and meetings for the three conditions. Details of the functions and activities of the meetings, based on the DACOM classification of business meetings (Short et al., 1976),

(e) Total number of turns and turn length.

Turns are defined as attempts by speakers to gain the conversation floor. We expected a number of factors to combine to increase turn length in ISDN. Given the problems of switching speakers, we expected that switches would occur less frequently and hence produce a greater number of longer turns. In addition, the difficulty of backchannelling and interrupting also would reduce the number of “quickfire” interchanges serving to indicate or clarify understanding. The absence of feedback and clarifications may also lead speakers to overelaborate and supply redundant information. This should also increase turn length. We therefore expected the ISDN meetings to have more of the characteristics of formal presentations or lectures where speakers deliver large amounts of material as an uninterrupted monologue. In contrast, in LN there should be no problems with rapid speaker switching or quickfire exchanges and turn number and length should be comparable with face-to-face interaction.

Prediction: ISDN should have large turns and infrequent changes of speaker. In LN turn length and frequency of switching should be equivalent to face-to-face interaction.

(f) Turn distribution

Finally we expected that turns might be unequally distributed in the different technologies. In each video-conference, it is possible for people to communicate with people at the local site (via standard face-to-face interaction), as well as with the remote site using the system. Given the difficulty of interacting over ISDN, our informal observations suggested that groups attempt to manage this problem by channelling their responses to the remote location through one specific individual at each location. We therefore expected that these local co-ordinators would dominate their group's contributions: the overall distribution of turns would be unequal with these individuals

different in LN where low lag times and full duplex should allow equivalent numbers of simultaneous starts as in face-to-face interaction.

Prediction: More simultaneous starts in ISDN, and LN and face-to-face interaction should be equivalent .

(d) Explicit Handovers

These occur when speakers signal that they intend to relinquish the floor using explicit verbal cues such as: (i) the use of questions; (ii) tagging, using stereotyped questions such as “isn’t it?”, “arent they?”, or statements such as “you know” or by the addition of redundant information on the end of a turn, for example,

A: ...ahm now I don't have I don't have a prob-
 lem with that at all but it but it wouldn't
 it would not mean that we have at any one
 point one interface you know it would just be
 [you know

and finally (iii) naming the next speaker (Levinson, 1983, Sacks et al., 1974).

We expected that speakers would try and alleviate the problem of speaker switching in ISDN by explicitly signalling that they had finished their turn. In ISDN, we therefore expected more instances of questions, tagging and naming of the next speaker. This should not be true in LN where speaker switching should be unproblematic, and explicit handovers unnecessary.

Prediction: More formal handovers in ISDN, and LN should be equivalent to face-to-face interaction.

We expected less floorholding in ISDN, because of adjustments by speakers. Speakers should be less likely to hold the floor because they want to avoid the disruptive effects of the half-duplex line on simultaneous speech, in deleting one speaker. There should be no such constraints with LN, where floorholding should be possible without such disruptive effects.

Prediction: Less floorholding in ISDN, with equal levels in LN and FTF.

(iii) Simultaneous Starts: These are instances of simultaneous speech when two participants concurrently begin a new turn. These occur when two or more speakers compete for the floor when the previous speaker has just finished. In some instances this may include an attempt by the original speaker to resume. This can happen when the original speaker yields the floor and after some time has elapsed believes there to be no contenders and so begins a new turn (Sacks et al., 1974).

A: well they'd be better be quick cos the nine-
 teenth is next Wednesday

B: [next week isn't it

C: [That's right exactly

In ISDN, we should expect more simultaneous starts because of the problems that participants have in timing speaker switches. Because of the lag, and the desire not to overlap the end of the previous speaker's turn, listeners may deliberately wait to respond to ensure that the speaker has finished. Given the slow response, the original speaker may assume that no other person wants to speak and may then begin to speak again. Meanwhile at the remote location another participant may have already begun to speak. This situation conspires to produce simultaneous starts. The situation is

disruptions, for example if the interruption deletes material which then has to be repeated, and turntaking re-established.

Again we expect that in LN with full duplex audio and almost zero lag, the interruptions will be much easier to achieve successfully. They can be delivered in overlap with the speaker, and the absence of lag means that the conversation has not moved on by the time they are transmitted.

Prediction: In ISDN, half duplex audio combined with lags in audio should produce fewer interruptions. In LN, interruptions should occur as frequently as in face-to-face interaction.

(c) Overlaps

Overlaps are instances of simultaneous speech which follow signals speakers give indicating that they may relinquish the conversational floor (Levinson, 1983). We made predictions for three different types of overlaps:

(i) Projection/Completion: this type of overlap occurs when the next speaker anticipates that the current speaker is about to finish, or tries to help the “forward movement” of an ongoing utterance (Clark & Wilkes-Gibbs, 1986). In projecting the possible finish by the current speaker, the next speaker may recognise that the message of the current speaker is complete, although the utterance has not finished.

A: initially that’s true but I wonder how the
 market will shape up [over time

B: [well you have to have a second punch
 behind it of course...

The next speaker may overlap in an attempt to complete the current speaker’s utterance. This can occur when the next speaker perceives that the first is having some

An inherent limitation of our method is that we did not attempt to measure the effectiveness of the communication across different media as other laboratory studies have done (Chapanis 1975; Morley & Stephenson, 1969, 1970, 1977; Wichman 1970). With real life meetings it is not clear what is an appropriate objective measure of successful communication, nor how we can easily compare the success of the different meetings. However, other work suggests that the types of communication characteristics measured here have implications for task outcomes. Laboratory studies have shown that lack of support for interactive processes such as backchannels and interruptions has effects on outcome measures such as time to solution and participants' understanding (Kraut, Lewis, & Swezey, 1982; Oviatt and Cohen, 1991). The characteristics can therefore be seen as indirectly measuring communication effectiveness.

(a) Backchannels

For the purpose of this study, only auditory backchannels were measured and not head nods or gaze behaviour. Backchannels are short feedback utterances, produced by the listener to indicate several functions including attention, support or acceptance of the speaker's message (Yngve, 1970). Example of utterances serving as backchannels are "mm", "uhu", "right", "okay" and "yes", although these utterances can sometimes have other functions than those described. They are often delivered with split second timing, for example²:

```
A:          ... in the absence of a of of a task for any
           particular set of users if we take the gen-
           eral task,=

B:          =right=

A:          =personal information management,=
```

2. Transcription conventions are described in Section 3.3

of eye gaze and posture change, although the importance of these cues is the subject of much debate (Williams, 1977); finally listener's nonverbal reactions may offer the speaker information about the effects of what they are saying on their audience (Short et al., 1976).

In face-to-face interaction, all participants in principle have equal access to the conversational floor, although there are external factors such as knowledge which can influence participation levels (McGrath, 1984, 1990).

2.1 Predictions

From the above it can be seen that face-to-face conversation exhibits characteristics which depend on three properties of the communication channels: (1) low transmission lags, i.e. messages are received almost instantaneously by listeners; (2) two way, e.g. feedback can be produced at the same time as the speaker's utterances, (3) multiple modalities, i.e. both verbal and visual channels are used (Whittaker, 1992). How will the properties of spoken conversation be changed by communicating using technologies that do not have these channel properties?

The ISDN system introduces a transmission lag of between 410 and 780ms for both audio and video, and a half duplex (one-way) line for audio. In addition, the ISDN system allows only limited visual cues, because the picture quality is poor and subject to jitter and occasional frame loss. Figure 4 summarises our expected findings for a number of spoken conversational characteristics. These predictions are based on the differences between the system channel properties and the properties of face-to-face interaction. We will first define the characteristics and give explanations of our predictions.

interaction the flow of the speaker is not interrupted by backchannels because the audio channel is two way. There is a second sense in which communication is interactive: when a breakdown in understanding does arise, the listener can immediately interrupt the speaker for clarification or to register disagreement. Alternatively by withholding verbal or visual feedback, the listener can indicate to the speaker that understanding is not guaranteed and the speaker can then make the requisite modifications.

Central to the process of conversation is turn-taking. In order to achieve the level of interactivity described above, speaker switches must be smooth and not disruptive to the overall flow of the conversation (Sacks, Schegloff, & Jefferson, 1974). How is this process achieved? There are a number of intonational, syntactic, pragmatic and non-verbal devices that speakers use to indicate that they are about to finish their conversational turn. Listeners also use non-verbal devices to indicate that they wish to speak e.g. leaning forward or achieving mutual gaze. The fact that listeners are able to predict when speakers are about to finish, means that there can be very low latencies, with speaker switching pauses varying between 620 and 770ms (Jaffe & Feldstein, 1970). In some cases, there is no pause, and overlaps occur between speaker transitions. Speakers also sometimes give overt cues to select the next speaker, such as naming an individual or directing a question at them. On other occasions, turn-taking is not smooth. Speakers sometimes attempt to hold the conversational floor and listeners make unsolicited bids for the floor (Levinson, 1983). These lead to overlapping speech, although whenever overlaps do occur they are usually resolved quickly with one speaker dropping out rapidly.

Verbal messages are often accompanied by multiple non-verbal cues; gaze, facial expression, posture and physical proximity. These serve a number of possible functions: they may help the listener to identify the meaning of the utterance (Argyle et al., 1968; Jaffe & Feldstein, 1970); they may also support smooth speaker transition by the use

Figure 3 about here

Face to Face Meetings

The face to face meetings took place in the conference rooms available on site at Hewlett Packard Laboratories Bristol. The room layouts were very similar to the one containing the ISDN system. Participants sat around tables approximately six feet long and four feet wide. Documents were shared by passing them around the table. An overhead projector was available but was not used in the meetings we observed.

2 FACE-TO-FACE COMMUNICATION AND PREDICTIONS ABOUT COMMUNICATION IN THE TWO VIDEO-CONFERENCES

Communication is a joint activity which requires co-ordination of both process and content (Clark & Wilkes-Gibbs, 1986; Whittaker, 1992). To allow this co-ordination to take place, conversation is both incremental and interactive. A key aspect of interactivity is listener feedback. The speaker delivers utterances incrementally, while the listeners provide concurrent feedback that the conversation is on track, by giving both auditory backchannels (e.g. “mm”, “uhu”) and visual evidence in the form of head nods and eye-gaze. This positive concurrent feedback informs the speaker that s/he can rely on and build upon the listener's understanding (Clark & Schaefer, 1989; Duncan, 1972; Whittaker & Stenton, 1988; Yngve, 1970). Where this feedback is absent or even delayed, the speaker's ability to formulate efficient messages is reduced (Krauss & Bricker, 1967; Krauss & Fussell, 1990). Without feedback, speakers are unable to assume the message has been understood: they may therefore attempt to clarify or reiterate points, sometimes unnecessarily, to ensure that the listener has not misunderstood. Absence or delay of feedback can therefore encourage the speaker to take long turns (Krauss & Bricker, 1967; Oviatt & Cohen, 1991). In normal face-to-face

loudspeaker sound being reflected back into the microphones. Secondly, a Shure AMS automatic microphone system is used which has unidirectional microphones, which do not pick up sound from the rear, and a very fast switching system ensures that only one or two microphones in the group are active at any time. Thirdly, a frequency shifter (5Hz) is used between the audio mixer and the network to limit howl reinforcement.

The rooms used are typically lecture theatres or seminar rooms. An example layout is shown in Figure 2.

Figure 2 about here

The participants sit at a table and face a set of four 20 inch monitors and a CCD camera. A confidence monitor displays the outgoing picture. Figure 3 shows an example monitor set up. On the table is an overhead camera for the display of documents and a control panel for the cameras. The controls are used by the participants to select the camera to be used for output and to pan and tilt as necessary. As in ISDN, participants can directly select and control the images they transmit but not the images they receive. Where four or fewer sites are being connected the sites are shown in full on the four monitors in front of the participants. If more sites wish to take part a system called "chairman's control" is used. The sites are shown "picture in picture" format in quadrants on the monitors as depicted by ABCD and EFGH in Figure 3. The chairman of the session chooses and displays the active speaking site on a full monitor. As the physical layouts for the sites vary, the perceived distances between participants also vary. In addition, as the number of participants increases at any single location, a wider angle of image is required, increasing perceived distance between participants. The broadcast quality video means that head movements are easily discernable. However, the offset camera means mutual gaze is difficult to achieve.

indication of what the other side sees. Thus, users are unaware of quality losses that may have occurred.

The second stack contains another large monitor which is used to display “stills” from the remote site. Stills are individual static frames showing graphics and documents captured and transmitted using an overhead camera. It is not possible to gesture at these images. If gesturing is necessary, an alternative is to use the live channel for documents or graphics but this means that the remote participants will be unable to view the images of the local participants.

LIVE-NET (LN)

LIVE-NET is the London Interactive Video Education Network. The system has been in operation since 1987 and now connects eight sites to a central video switch. The longest link is 42 km. It is used for intercollegiate lectures, seminars and meetings. The colleges are dispersed over a large densely populated metropolitan area, making travel between the different colleges difficult and time-consuming. LIVE-NET is an analogue system. Each site is connected by a pair of optical fibres each carrying four full bandwidth video channels, with sound on a 6 MHz subcarrier, and a fifth lower bandwidth channel used for data up to 2 Mb/s using a switched star topology. These five channels are frequency modulated onto a carrier, which is then converted into a 250 MHz multiplex and used to intensity modulate a laser diode. The result is a full motion picture with none of the frozen picture motion that is associated with some digital video systems.

As there is no video or audio processing, the time lag is simply the propagation time at the speed of light. Delays can therefore be measured in microseconds. The audio subsystem is full duplex. Several measures have been taken to eliminate feedback problems. Firstly, there has been some acoustic treatment in the rooms to prevent

The audio channel is half duplex and so the voice of only one person can be transmitted at any time. This is necessary to eliminate problems caused by echo or feedback when the sound from the loudspeaker is picked up by the microphone and retransmitted across the line. There are also occasional transmission problems with the system causing brief disruptions both to the audio channel and the video picture.

The conference room is a converted meeting room containing a table at which three people can sit comfortably. Sitting at the table users can see two screens in front of them (See Figure 1). The first is directly in front of the table at a distance of approximately nine feet and contains a 26 inch colour monitor above which two cameras are located. The monitor displays the live picture of the remote location. Mutual gaze is not possible given the offset camera, and the distance and the video quality make remote eye-gaze and head movements unclear. The perceived distance of the remote participants is difficult to evaluate but it seems to depend on actual distance from the screen and the nature of the image of the remote participants, namely whether the shot is full face, head and shoulders, or full body.

Figure 1 about here

A small desktop control panel enables users to switch between cameras, focus, pan and zoom. Participants control their local cameras, choosing the view which they wish to transmit. The control panel allows users to switch between close up shots of a speaker and a view of the participants seated around the table. In practice, this alternating between views was rarely used. People tended to fix on a view showing all the remote participants seated at their table, displaying their head, arms and upper bodies. On top and to the right of the cameras is a small 9 inch colour monitor, (“confidence monitor”), which displays the live picture which is transmitted to the remote site. This image has not been compressed and decompressed and is not, therefore, a true

1.1 The Systems

ISDN System (ISDN)

The system is located at Hewlett Packard Laboratories Bristol and the majority of the conferences held in Bristol are to the USA. Conferencing takes place over two ISDN lines each at 64kb/s. Rate adaptation must take place because US installations use a public switched 56kb/s digital network. Thus the available bandwidth is reduced from 128kb/s to 112kb/s. Of this 16kb/s is used for audio with an additional amount for communication between CODECs. The amount of bandwidth available for video transmission is approximately 90kb/s.

The video signal is compressed by removing both spatially and temporally redundant data using a CLI Rembrandt CODEC. This process takes about 120ms with an equivalent time required for decompression at the other site. The audio signal is also compressed but has to be buffered to synchronise it with the video. In addition, there is the propagation delay of sending the data. This delay depends on whether a terrestrial or satellite link is used. For a terrestrial link a propagation delay of approximately 170ms in each direction, can be expected to the West Coast of the US, although this will vary depending on the route taken. A satellite link is much slower. The time taken for the signal to travel from the earth station to the satellite is 135ms and an equivalent time is taken to transmit the signal to the next earth station. To connect to the West Coast of the US two satellite jumps are required which means a delay of 540ms in one direction. Thus, allowing for compression and transmission, the lag between a person on one site speaking and the signal arriving at the other site can vary between 410ms to 780ms, depending on the propagation route.

Previous work has addressed the relationship between the properties of different communication media and the conversational characteristics they can support. This work has shown that the more closely a set of media approximates to face-to-face interaction in their properties, the closer the conversational style is to face-to-face interaction. This has been demonstrated for a number of different conversation characteristics, e.g. for number of turns, interruptions, overlapping speech and pausing (Argyle, Lalljee, & Cook, 1968; Cohen, 1982; Jaffe & Feldstein, 1970; Rutter & Stephenson, 1977). A major problem with this research however, has been to operationalise the theoretical constructs used for defining media traits, such as “social presence” (Short et al., 1976), “cuelessness” (Rutter & Robinson, 1981), and “media richness” (Daft & Lengel, 1986). In contrast, in this study, we select measurable variables based on channel properties: half-duplex versus full-duplex audio and lags in audio; broadcast versus low quality video. We study the effects of these variables on a number of spoken conversational characteristics that have independently been shown to be important in face-to-face interaction. Our analysis mainly focusses on the spoken aspects of conversation, although we include a brief discussion of the impact of channel properties on visual behaviour.

We address how these channel properties of the video-conferencing technology affect the nature of spoken conversation in real meetings by comparing interaction in two wide-area systems with face-to-face conversation. We outline the critical characteristics of conversation and then examine how these differ for the following interaction technologies: (a) a video-conferencing system with half-duplex audio, transmission lags and poor picture quality (b) a high quality video-conferencing system with duplex audio, no transmission lags and full bandwidth video (c) face-to-face communication. We first describe the two video-conferencing systems, motivate the conversational characteristics, and then derive predictions about how those characteristics differ across the respective media as a result of the different channel properties of these systems.

themselves, and reports sometimes anecdotal. Recent evaluation work has also failed to find strong benefits for this technology. Evaluation of desktop video indicates that interaction is more like phone conversations than face-to-face meetings, with conversations tending to be brief, and task focussed (Fish, Kraut, Root, & Rice, 1992). Attempts to use video for opportunistic meetings (“social browsing”) have similarly been unsuccessful when compared with face-to-face interaction: unplanned video contacts were less likely to lead to lengthy conversations than similar face-to-face contacts (Fish, Kraut, & Chalfonte, 1990).

There are two problems with the research into workplace video. Most of it examines systems that link local work environments, and the technology is local area networks or Cable TV, which support high quality audio and video. However, most commercially available systems are wide-area, where networking constraints do not allow high quality video and audio. We know from other work that reduced quality audio and video has strong impacts on communication properties (Cohen, 1982; Krauss & Bricker, 1967; Krauss & Fussell, 1990; Rutter & Stephenson, 1977; Sellen, 1992). Also there may be less incentive to converse over a local video system or to hold certain types of conversations, when there is the alternative of engaging in face-to-face communication. For these reasons we chose to examine real work meetings over two wide-area video-conferencing systems currently being used by organisations to support remote collaboration. We examine how the channel properties of the video-conferencing media affect the characteristics of the spoken conversations.

The aim of this research is therefore to identify possible reasons for the lack of success of video-conferencing technology. Our claim is that the properties of the communication channels in those wide area systems prevent the execution of certain basic communication processes which may be crucial for certain collaborative interaction tasks.

telephone are limited to the auditory medium. The hypothesis is that by adding a visual channel to the phone, the added benefits of gaze, gesture and the ability to monitor people's reactions will improve the quality of the communication. In addition, by facilitating frequent high quality interaction between distant sites, these technologies will increase the number of potential co-workers, and hence improve the quality of remote collaboration. There should also be less need to travel because people can replace face-to-face meetings with video-conferences.

However an examination of the history of video-conferencing and video-phone reveals a lack of success. Despite promising past market forecasts, video technology has not gained widespread acceptance (Egido, 1988; Noll, 1992). In part this is due to inadequate analysis of user needs, in particular with regard to travel substitution (Johansen, 1984; Johansen & Bullen, 1984; Panko, 1992).

Other work has raised questions about the value of the video channel. Laboratory studies assessed the impact of different channels on the efficiency of solving various tasks. Results indicated little value to adding a visual channel for task-based communication such as information transmission or collaborative problem-solving (Chapanis, 1975; Reid, 1977; Williams, 1977). The time taken to solve problems, and the quality of solution is almost equivalent, whether or not a visual channel is available. Other studies have found some tasks for which video does influence outcome, but this is highly dependent on task type (Short, Williams, & Christie, 1976; Williams, 1977).

Another approach has investigated video as a technology in the workplace and from this perspective, studies of desktop video and persistent video links have been undertaken (Abel, 1990; Dourish & Bly, 1992; Mantei, Becker, Sellen, Buxton, Milligan, & Wellman, 1991). Although early reports of this work were mainly favourable, the systems have often been used by the developers or researchers

improve but not remove all the problems of video-conferencing as an inter-personal communications tool and we describe possible solutions to the outstanding problems.

KEYWORDS: Mediated communication, video, audio, multimedia, conversation.

1 INTRODUCTION

In most working environments people collaborate in groups to undertake collective tasks. Face-to-face communication plays an important role in the development and maintenance of these collaborations and it is also critical for certain classes of workplace communication task such as project definition, initiation and planning (Finholt, Sproull, & Kiesler, 1990; Galegher & Kraut, 1990; Kraut, Egidio, & Galegher, 1990). A number of changes in work practice mean, however, that physical proximity and hence informal face-to-face interaction may not always be possible for future work groups. There are three main trends here: telecommuting, (Harkness, 1977; Kraut, 1989); mobile work, e.g. from customer sites or “on the road” (Sproull & Kiesler, 1991); and concurrent engineering, with designers, suppliers, and manufacturers increasingly co-ordinating over widely dispersed geographical locations (Johansen, 1984).

Given these trends, it is imperative that some means be found to support informal interaction remotely. This has led to an increased interest in technologies that attempt to replicate face-to-face interaction, such as video-conferencing and the video-phone. These technologies are premised on the hypothesis that the more closely they mimic face-to-face communication, the more effective the communication that will take place. Current pervasive technologies for remote synchronous communication such as the

ABSTRACT

Recent trends towards telecommuting, mobile work and wider distribution of the workforce, combined with reduced technology costs, have made video communications more attractive as a means of supporting informal remote interaction. In the past, however, video communications have never gained widespread acceptance. Here we identify possible reasons for this by examining how the spoken characteristics of video-mediated communication differ from face-to-face interaction, for a series of real meetings. We evaluate two wide area systems. One uses readily available ISDN lines but suffers the limitations of transmission lags, a half duplex line, and poor quality video. The other uses optical transmission and video-switching technology with negligible delays, full duplex audio and broadcast quality video.

To analyse the effects of video systems on conversation, we begin with a series of conversational characteristics that have been shown to be important in face-to-face interaction. We identify properties of the communication channel in face-to-face interaction that are necessary to support these characteristics, namely that it has low transmission lags, it is two way, and uses multiple modalities. We compare these channel properties with those of the two video-conferencing systems and predict how their different channel properties will affect spoken conversation. As expected, when compared with face-to-face interaction, communication using the ISDN system was found to have longer conversational turns, fewer interruptions, overlaps and backchannels and increased formality when switching speakers. Communication over the system with broadcast quality audio and video was more similar to face-to-face meetings, although it did not replicate face-to-face interaction. Contrary to our expectations, formal techniques were still used to achieve speaker switching. We suggest that these may be necessary because of the absence of certain speaker switching cues. The results imply that the advent of high speed multimedia networking will

**Conversations over Video-conferences:
An Evaluation of the spoken aspects of Video
Mediated Communication¹**

**Brid O'Conaill, Steve Whittaker
Hewlett Packard Research Laboratories, UK.**

&

**Sylvia Wilbur
Queen Mary and Westfield College,
London.**

1. Requests for reprints should be sent to the first author at Hewlett Packard Laboratories, Filton Rd., Stoke Gifford, Bristol, BS12 6QZ, UK.