Informal, desktop, audio-video communication

J M Thorne‡

‡ BT

Abstract: Audio-Video systems have been developed to support many aspects and modes of human communication, but there has been little support for the informal, ongoing nature of communication that occurs often in real life. Most existing systems implement a call metaphor. This presents a barrier to initiating conversation that has a consequent effect on the formality of the resulting conversation. By contrast, with informal communication the channel is never explicitly opened or closed. This paper examines the range of previous systems and seeks to build on these to develop plans for supporting informal communication, in a desktop environment.

1 Introduction.

This paper is concerned with informal desktop audio-video communications. Interactions between workers in a physical office can be viewed "as one long intermittent communication comprising multiple brief related fragments over an open channel" [1]. There appears to have been relatively little attempt to allow geographically separated participants to engage in this form of audio-visual communication, that occurs naturally between workers who are co-located.

The 1994 study by Frolich et al [1] illuminates this mode of communications further: He discovered that some 30% of all work time was consumed by informal face to face communications and about half of this time was spent in conversation with others in the same office. Exchanges are mostly brief with 50% lasting less than 38 seconds and over 80% involving just two people. Only a small minority either began with a greeting or ended in a farewell and most assume a large degree of shared prior context.

Most existing systems implement a call model where effort is required to begin and terminate an interaction, which in turn affects the content and use of that interaction. This paper begins to examine how technology might be provided to enable such an open channel where there is no explicit beginning or end, bearing in mind both the social aspects of how such a system is presented and configured, and the scalability challenge of providing large numbers of uncoordinated video connections. We begin by presenting some background to other relevant systems and studies of human behaviour and follow with some plans on how we might proceed.

2. Background

Issacs and Tang [2] observe that 'raw technology' is not itself useful; it must fit within people's lives. Tools must take advantage of users' existing communications skills without requiring conscious effort to accomplish what is normally an unconscious activity [3]. Thus it is essential that we understand the context and purpose of the communication we are trying to support. This is borne out by the range of existing video conferencing systems from the large, expensive room based systems to the free packages that form part of today's instant messaging offerings. There is no 'one size fits all' system.

Room based systems such as Access Grid [4] (widely adopted in the academic world) allow complete control of lighting, cameras, speakers and microphones to enable very compelling and involving experiences. However, they require significant effort to setup and run (booking the rooms at both ends, travelling to them, changing cameras at appropriate moments). They attempt to replicate the physical meeting room experience, enabling a simple cost benefit analysis e.g. is the cost of tolerating "imperfect" video communication smaller than the cost of flying across the Atlantic for a meeting?

Other systems address different communications scenarios and needs and have different cost compromises. Forum [2] addressed a speaker presenting to a large distributed audience, multicasting video out to the listeners, but the inability to see that audience made the task of presenting less rewarding. VSee [5] supported the smaller grouping of a teacher to a classroom, leveraging the general asymmetry of the situation by providing only low bandwidth video from students back to the teacher but allowing individual links to become full rate when a student raised their hand. Montage [2] tried to

replicate the experience of peering into offices to see if someone was available by providing the ability to make brief reciprocal video glances that fade in on the users' desktops. Porthole [6] systems try to give background awareness of many users by assembling a matrix of very low rate video images, updated perhaps every 5 minutes. The Office of the Future project [7] seeks to connect two offices as if a hole had been cut in the wall between them, by using multiple cameras and projectors to capture and transmit depth and reflection information. Such a system may provide convincing immersion and sense of connection but unfortunately remains impractical in open plan offices where there are no large surfaces to project onto and each user may be engaged in different private conversations.

Xerox PARC's Media Spaces [8] grew out of the premise that work is fundamentally social, that the activities of the participants consist of more than just the task they are engaged in. In the early 1980's they evolved a system that connected offices and common areas together using analogue audio-video connections and allowed the configuration to be altered through a computer controlled switch. The system was on 24 hours a day such that connections always went somewhere, though individual monitors, cameras and microphones could be switched off. It provided the opportunity to connect two groups into the same meeting, to allow chance encounters with others passing through remote common areas, background awareness of distant activity and one to one connections. The flexibility allowed users to develop their own patterns of use. This comes close to supporting the informal communications we are interested in, but its dedicated audio-video links and the central switch make the system expensive and un-scaleable. The analogue nature of the links also makes integration with other flows of data difficult. Usage for direct communication between pairs was limited, perhaps because doing so would cut you off from awareness of others.

3. Discussion

The closest that existing technology comes to allowing informal always on communication is through the now ubiquitous, text based instant messaging systems, where conversations can be started, resumed, or allowed to fall silent almost effortlessly. 'Push to Talk' [9] may provide a similar experience using voice on mobile phones. Neither requires the high bandwidth of video, but typing is slow and inexpressive while the half duplex audio present in current Push to Talk systems may make coordination of speech awkward.

So what does video offer that improvements to text and audio systems wouldn't? Early studies of video conferencing found little or no improvement over audio only systems. These tended to focus on short task based experiments and the products of those tasks such as completion times or decisions reached. Issacs and Tang [3] suggest that video finds its use in the process of interaction not the product, and that its value might only become obvious in the long term between participants who know each other well. From their studies they saw that participants used the visual channel to express understanding, forecast responses, enhance verbal information, express purely non verbal information, manage pauses and express attitude through posture and facial expression. They suggest that video should be of use in highly interactive situations such as creating rapport, negotiating and conflict resolution. However, high quality, low latency audio is important. Without audio a video conference would be almost useless, without video, that's just a telephone call.

Desktop video systems don't, however, succeed in faithfully emulating face to face conversations; they lack immersion, stereo vision and the ability to make mutual eye contact. Eye contact is typically not possible because a user cannot look simultaneously at the camera and the image of the remote participant. In real face to face communication gaze has been shown to be of extensive use in coordinating utterances. Different behaviour is shown by the listener and the speaker, with listeners looking at the speaker twice as much as the speaker looks at the listener [10]. Speakers look away as they begin speaking and then back again as they end, enabling them to see the reaction of the listener. Mutual gaze occurs about 30% of the time. Gaze and eye contact allow greater informality in the communication with interruptions that would not be tolerated in an audio only situation, being perceived as ok.

Garau et al [11] used avatars to represent real pairs of participants in an immersive shared virtual world. They specifically isolated gaze behaviour as a factor for analysis. Simulating realistic eye gaze behaviour using the statistics for listeners and speakers was found to improve the perceived quality of

the communication. Consequently there has been significant effort expended on attempts to provide eye contact. “Video tunnels” can align the optical paths of the camera and screen using mirrors but are bulky, and many combinations of computer vision and graphics techniques have been proposed. Both Zitnick et al [12] and Taylor and Rowe [13] suggest mapping frames of video onto 3D face models that can be re-rendered from the correct viewpoint. Criminisi et al [14] and Xu et al [15] address the use of pairs of cameras to provide view interpolation without an explicit scene model.

Immersion and stereo vision are lost because desktop monitors only occupy a small, 2D part of the visual input. The peripheral vision of the user roots them firmly in the local space and not in a shared space with the remote user. In real life people adjust their relative positions in space to seek a certain degree of proximity with those they are conversing with [10]. This ability to interpret and control interpersonal distance is reduced in desktop systems. CAVE systems and virtual reality headsets can help restore immersion and stereo vision. Garau et al [11] note that in VR users automatically adopt appropriate interpersonal distances between themselves and the virtual representation of the other user. Anecdotal evidence also suggests users of VR display other forms of socially responsible behaviour such as apologising for walking through each other. One attempt at restoring the awareness of peripheral action on the desktop is given by Buxton et al [16], they suggest providing multiple linked camera views, both close-up and wide view, augmented to show the links between them.

If mutual sharing cannot be achieved, then trust and privacy become important issues. Issacs and Tang discovered that when presented with the idea of using an audio-video system, many users are concerned about their privacy and also about invading the privacy of others. Bradner and Mark [17] showed that being observed via video conferencing or application sharing had a negative impact on people’s ability to solve tasks. Subjects reported unease at exposing current thoughts and rough working. However, Issacs and Tang note that few users actually took advantage of privacy controls when they were offered. Lewis and Cosier [18] suggest that to ease the user’s fear that their privacy is being invaded, the metaphor to aim for in videotelephony is one of mutual giving of pictures, not taking. He suggests that this might be achieved by giving the user more control over how they are perceived including easy to use manipulations of lighting and shot compositon.

3. Direction

How do we proceed with enabling informal always on connections? We have seen that video is useful, and eye contact particularly so. Audio, privacy, space and awareness are important too. This work aims to construct a system that allows informal one-to-one desktop audio-video communication over an always open channel, such that there is no more effort required to begin or end a fragment of communication than when talking to someone else in your own office. Allowing the richness of synchronous eye contact that is absent from text and audio only systems. This must be done in a smart way, since always-on, uncoordinated, high rate video communications will not scale well to many users. We will need to sufficiently understand user behaviour so that we can adapt the bandwidth provided to how active the channel is but without denying users the awareness cues needed during the lulls in conversation. Presentation and user interface metaphors are important too such that users know how to simply switch between conversations, are not cut off from the real office they inhabit, can adjust the perceived distance between themselves and others, and can control the privacy of their communications and behaviour between fragments.

There are many further questions that need to be answered, such as: how many conversations do users simultaneously engage in? How does removing the barrier to initialising a call affect the formality, intensity and length of the resulting communication? How do we make it acceptable to leave a connection on but idle? What cues are necessary during a lull to maintain awareness and allow easy re-engagement? How long are the gaps between fragments of conversation, what percentage occupancy is there in an audio-video channel? How do we integrate this continuous awareness within a desktop environment where screen real estate is limited and immersion impossible?

So how do we know when we’ve succeeded? We need to achieve scalability and user acceptance. Simulations and user studies can be used to evaluate our progress. User studies should take place over long time scales in users’ normal environment, with users who are not involved in the development of the system [2] such that a successful system is one adapted to their needs, not one to which the users

have successfully adapted themselves. There should be a focus on deriving objective measures such as usage statistics to support hypotheses on what informal communication is like.

4. Conclusions.

We have presented a brief overview of the range of systems that have been developed to support communication over audio and video. This has been used to illuminate the key considerations that must be implemented in any system. And we have outlined plans for how we might apply these principles to supporting informal, desktop, always on communication.

Acknowledgments.

This work has been supported by BT, the EPSRC and UCL.

References.

[1] Frohlich, D., Whittaker, S., Daly-Jones, O. (1994) Informal Workplace Communication: What Is It Like And How Might We Support It? CHI'94.

[2] Isaacs, E., Tang, J. (1997) Studying video-based collaboration in context: From small workgroups to large organizations, in K. Finn, A. Sellen, & S. Wilbur, Video-Mediated Communication, Mahwah, NJ: Lawrence Ehrlbaum.

[3] Isaacs, E., Tang, J. (1993) What Video Can and Can't Do for Collaboration: A Case Study. Proc. ACM Multimedia 1993.

[4] Access Grid - http://www.accessgrid.org/

[5] Chen, M. (2002) Achieving Effective Floor Control with a Low-Bandwidth Gesture-Sensitive Videoconferencing System. ACM Multimedia, 2002.

[6] Lee, A., Girgensohn, A., Schlueter, K. (1997) NYNEX Portholes: Initial User Reactions and Redesign Implications. ACM GROUP 97.

[7] Raskar, R., Welch, G., Cutts, M., Lake, A., Stesin, L., Fuchs, H. (1998) The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays. SIGGRAPH 98.

[8] Bly, S., Harrison, S., Irwin, S., Media spaces: bringing people together in a video, audio, and computing environment. ACM 1993.

[9] Push To Talk http://www.nextel.com/services/directconnect/ptt_overview.shtml

[10] Argyle M. Bodily Communication 2nd Ed. 1988, ISBN0-415-051142

[11] Garau, M., Slater,M., Vinayagamoorhty, V., Brogni, A., Steed, A., Sasse, M.A. (2003) The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment. SIG-CHI 2003.

[12] Zitnick, Gemmell, Toyama (1999) Manipulation of Video Eye Gaze and Head Orientation for Video Teleconferencing. Microsoft Technical Report June 1999.

[13] Taylor, Rowe (200) Gaze Communication using Semantically Consistent Spaces. CHI 2000.

[14] Criminisi, A., Shotton, J., Blake, A., Torr, P.H.S. (2003) Gaze Manipulation for One-to-one Teleconferencing, Ninth IEEE International Conference on Computer Vision Volume 1, 2003.

[15] Xu, L., Lei, B., Hendriks, E. (2002) Computer Vision for a 3-D Visualisation and telepresence Collaborative Working Environment. BT Technology Journal Vol 20 No 1, January 2002.

[16] Yamaashi, K., Cooperstock, J., Narine, T. & Buxton, W. (1996) Beating the limitations of camera-monitor mediated telepresence with extra eyes. CHI '96.

[17] Bradner, E., Mark, G. (2001) Social Presence with Video and Application Sharing GROUP'01.

[18] Lewis, A.V., Cosier, G. (1997) Whither video ? — pictorial culture and telepresence. BT Technology Journal Vol 15, Issue 4, 1997.