Talking Your Way Around a Conference:
A speech interface for remote equipment control

Anuj Gujar, Shahir Daya, Jeremy Cooperstock,
Koichiro Tanikoshi, William Buxton

University of Toronto
Toronto, Ontario M5S 1A4
+1-416-978-6619

Abstract

Videoconferencing enables people to attend and participate in meetings from remote locations. The key problem faced by electronic attendees is the limited sense of engagement offered by the audio-visual channel. The attendee is typically restricted to a single view of the room and has no ability to interact with presentation technology at the conference site.

As a first step to improving the situation we want to assign electronic attendees a view of the room appropriate to their particular “social roles,” which may include presenting a topic, listening to a talk, or participating in a discussion. However, attendees may change roles during a meeting, thus requiring a different position and view more suited to the new role. This involves switching video inputs and outputs to new cameras and monitors.

One possible method to enable video attendees to effect these changes independently is to provide them with the same graphical user interface (GUI) that the central site has to control the equipment. Unfortunately, using state-of-the-art systems for such control is often confusing and complex. Furthermore, this solution requires the attendees to have “extra” computer equipment (i.e. equipment not already required for videoconferencing) and learn how to operate the GUI.

Instead, using speech recognition and video overlay technologies, we are able to provide a non-technical interface to equipment in the meeting room. In doing so, we do not require any extra equipment at the attendees’ sites. Our approach provides attendees with the means of controlling their own view of the meeting, changing electronic seats, and manipulating equipment remotely, all through simple voice commands.

1 Introduction

The key problem faced by electronic attendees or “visitors” is the limited sense of engagement offered by the audio-visual channel. To improve the situation, visitors should be provided with the ability to perform tasks naturally, as though they were physically at the meeting.

As a means of increasing visitors’ sense of engagement, we implemented a “virtual window” [6], which allows them to peer around our room as if looking through a window. The virtual window is implemented via a head tracking system [2], which responds to head translations of visitors. The system functions by mapping these translations to control signals for a motorized camera located in our conference room. Making use of this technology, electronic visitors are no longer limited to a static view of our room, but instead, can move their heads to change their views as desired.

Unfortunately, controlling a conference room camera does not in itself ensure that remote attendees can participate effectively in a meeting. For example, the camera view may be obstructed. Furthermore, if an electronic visitor wishes to assume the role of presenter, the virtual window will not help the visitor change seats or control the presentation technology such as the VCR and document camera.

We would like to extend the idea of changing views by allowing remote attendees to move to new electronic seats, just as local attendees can move to new physical seats. Such a move would be useful to change social roles or simply to improve one's view. This involves switching input and output audio-video (A/V) signals among the appropriate cameras, monitors, microphones and speakers. The move should be possible without requiring the assistance of attendees who are physically present in the conference room.

We could add to this functionality the ability for visitors to control presentation equipment in the conference room. Combined with the seat changing ability, this would greatly enhance the electronic attendee's sense of engagement. One way to accomplish this is through the use of a graphical user interface (GUI) that communicates with the A/V devices in our conference room. However, this approach requires "additional" computer equipment (i.e. equipment not already required for traditional videoconferencing) at the remote site, in addition to extensive training. A further problem with the GUI is that its use is highly distracting, a problem identified in an earlier configuration of our videoconference environment [3].

Instead, using speech recognition technology to control the A/V devices, we can provide an interface to the remote attendee that eliminates the need for additional equipment and computer communications at the remote site. It is important to recognize that the microphone and audio channel are already in place to permit voice communication for the videoconference.

To indicate what options are available to the attendee, we can make use of a video-overlaid menu, which appears on the attendee's monitor. Again, the monitor and video channel are already in place, so no extra equipment is required. Together, the speech interface and video overlay technique replace the GUI and computer display of a conventional interface.

We designed and implemented the Hyper Doorway [4], which provides all of the above functionality. The remainder of this paper describes the system in more detail.

2 System Overview

2.1 Architecture

Electronic attendees communicate with the conference room through nodes, which consist of a microphone, speaker, video camera and video monitor. An "electronic seat," consisting of the same hardware, is provided in the conference room for each attendee. This setup serves as a video surrogate, so that local attendees can communicate with the visitors [7].

As shown in Figure 1, the interconnections between all of the A/V equipment in the conference room are controlled by a Desk Area Network (DAN). The DAN consists of an Akai A/V switch

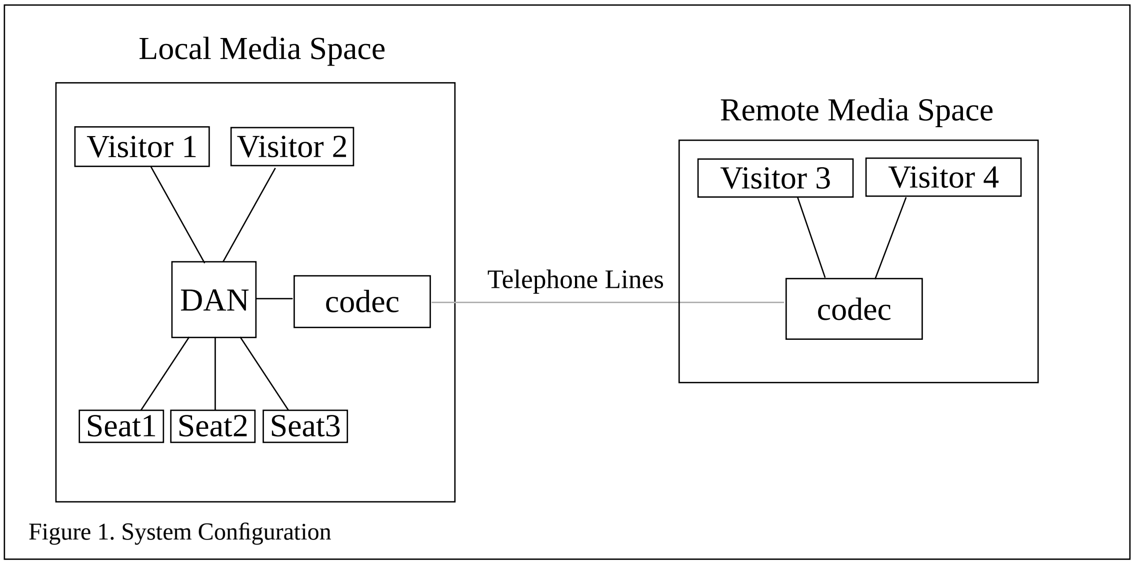

Figure 1. System Configuration

and software running on a Sun UNIX platform.

To permit videoconferencing from outside of our media space [5], an A/V coder/decoder (codec) is required. Remote sites with a codec can connect to our conference room through telephone lines.

We use a centrally located IBM PC[1] 486 running Microsoft[2] Windows[3] 3.0 as the link between the remote attendee and the DAN, as depicted in Figure 2. A RocGen VGA[4] card, installed in the PC, is used to generate a composite image containing the computer-generated video overlay. The overlay provides a list of DAN services to the attendee, who can use speech to select one of the options. The Voice Assist[5] software, running on a Sound Blaster 16[5] audio card in the PC, then performs speech recognition to extract the attendee's request and relay it to the DAN. Communication between the PC and the DAN is handled by the Hyper Doorway system, which was implemented using the Windows Sockets Application Programming Interface.

2.2 Voice Recognition

Voice recognition systems appear in many forms. The system we chose is a speaker-dependent, unlimited-vocabulary, discrete-utterance system. Although a speaker-independent system would not require user-specific training, we feel that accurate and reliable interaction is essential in producing an interface that is readily accepted by users. As a result, we decided that for the purposes of prototyping, we would use the speaker-dependant system, which, while requiring training, is more reliable.

There are still two reliability problems with our speech recognition system. First, we have observed that due to inconsistent background noise, speaker utterances, and room acoustics, the number of recognition errors can be significant. Second, user specific training for large vocabularies is very impractical. To address both of these problems, for the purpose of prototyping, we chose to limit the vocabulary to the digits "0" through "9".

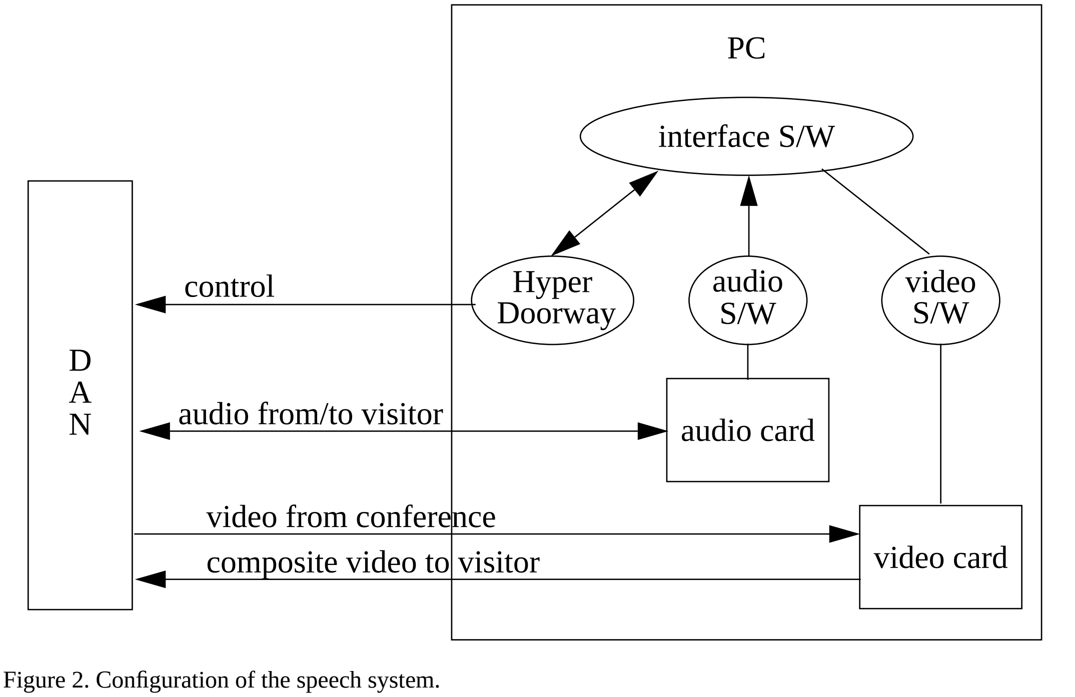

Figure 2. Configuration of the speech system.

[1] IBM PC is a registered trademark of International Business Machines Corporation.
[2] Microsoft is a registered trademark of the Microsoft Corporation.
[3] Windows is a trademark of the Microsoft Corporation.
[4] Manufactured by Roctec Electronics Incorporated.
[5] Voice Assist and Sound Blaster 16 are trademarks of Creative Technology Limited.

2.3 Video Overlay

Using speech, we have solved the problem of providing input without additional equipment, but this introduces two new problems. First, electronic attendees have no way of knowing what services are available from the DAN, and second, they do not know how to invoke these services.

One simple way to correct these problems would have been to replace the video image on the visitor's monitor with a graphical menu displaying available services and providing instructions for their invocation. However, visitors might feel disengaged when the conference room view is replaced by the menu. Our design attempts to avoid introducing such discontinuities.

Instead, we provide the menu using video overlay technology. This enables us to combine a video image with computer graphics into a composite image that can be displayed on any NTSC video monitor.

The technology we are using provides four major services:

- conversion of VGA signal to NTSC
- genlock, for synchronization of different video signals
- graphics display on top of video
- fade in/out effects on VGA and/or video signals

We are presently exploiting the first three of these services, and will soon be adding the fourth, to provide an interface that is both effective and usable. Video overlay minimizes disturbances that would be caused by a more simplistic computer-generated menu that hides the visitor's view of the conference room.

3 Interface Design

The major advantage of a speech interface is that it eliminates the need for a keyboard or mouse as input devices to the system. However, because our system operates on speech that is provided to the conference room, we are able to go one step further. We need only one computer, installed at this central site, rather than computers at the location of each remote attendee.

Our interface progressed through several iterations. Each iteration explored a new method of presentation to the visitor, addressing the downfalls of the previous stage. The following sections describe the progression of the interface.

3.1 Text-Based

The first prototype used a text-based menu. To invoke the menu the user would say, "computer." The menu then appeared, displaying the seat changing services offered by the DAN, as illustrated in Figure 3. Each service appeared by name, alongside the associated voice command, a number between zero and nine. The menu would disappear either after a selection was made or a preset time-out period expired. From our own experience, we found 7 seconds to be long enough for users to make a selection, but not so long as to be distracting.

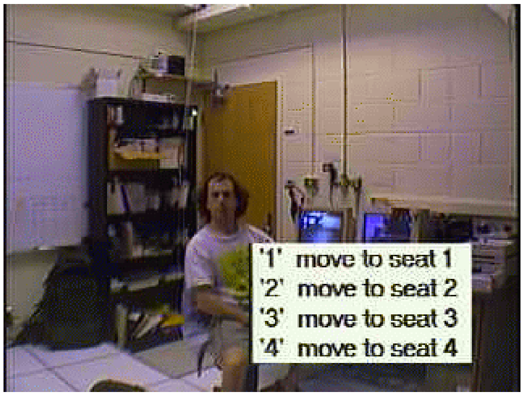

Figure 3. First iteration menu overlaid the video image. The voice commands required to select the individual services are displayed in quotes.

This prototype served as a proof-of-concept to show that visitors could change seats successfully using a speech-based interface. The novelty here was that visitors could control equipment in our conference room without any extra computer equipment at their site.

However, the text-based prototype had one major drawback. Visitors, having little experience with our conference room, lacked information regarding the location of devices and electronic seats. Without this, it was impossible to know what location was best suited to one's social role. This problem would be difficult to address with a text-only interface. Therefore, our next interface exploited the two-dimensional spatial information conveyed through a graphical representation of our conference room.

3.2 Floor Plan

To provide orientation to the visitor, our second interface displayed a floor plan sketch of the conference room, overlaid on the video signal, as shown in Figure 4. The overlay was invoked and cleared as in the previous interface.

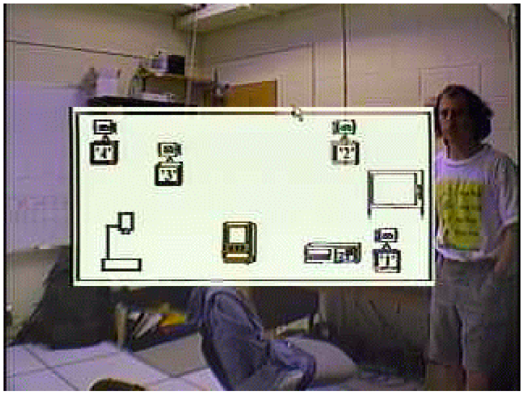

Figure 4. Floor plan overlaid on the video image.

The floor plan identified several devices in the room, as well as the relative locations and corresponding commands, in quotes, of the electronic seats to which a visitor could move. As a result, visitors had enough information to choose their desired location.

In one scenario, if the visitor wished to present, using the electronic white board, he would likely want to move to seat 2. From this position, he can face the local attendees, who will now be looking in his direction because of his proximity to the white board. If, instead, the visitor appeared at seat 4, he would only see the backs of the local attendees while they faced the white board. Local attendees would also suffer, having to switch their attention between the white board and the presenter.

A drawback of this interface was that the visitor could not predict the view from a particular seat until a move had been performed. The resulting view might not be satisfactory, either due to camera orientation or visual obstruction. In such a situation, the visitor would likely want to change seats yet again. Such trial-and-error seat changing was found to be time consuming and frustrating for the user; thus, reducing the sense of engagement.

3.3 Floor Plan with Seat Views

To reduce the trial and error aspect of the floor plan interface, we added camera views, as in Figure 5. The seat view interface provides these views in the form of snapshot images taken from each available seat. Snapshots are captured periodically by a frame grabber, converted to Windows device-independent bitmaps and then made available to the interface.

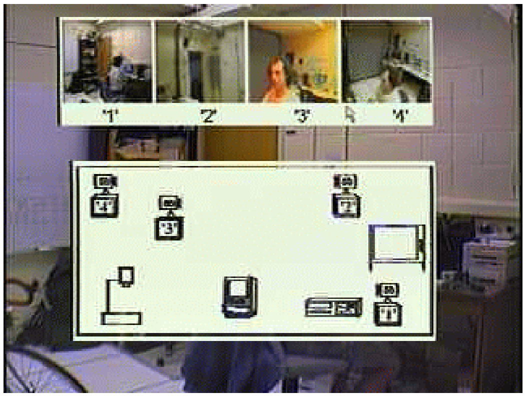

Figure 5. Floor plan and seat views overlaid on the video image.

The seat view interface provides far more information than the previous iterations. As a result, visitors can see what they will view before they actually switch seats.

4 Ongoing Work

The final iteration works very well, but from Figure 5, it is clear that much of the original video image is being obstructed by the graphics overlay. We are currently working on reducing the impact of the overlay by decreasing its size and making the overlay translucent instead of opaque. We must, however, keep in mind the problem of the low image quality that results from video that is degraded when transmitted through codecs. Our early experiments with translucent overlays provided text that appeared reasonable on local video monitors, but relatively unreadable at remote sites.

Another issue currently being addressed is how to distinguish between commands directed toward the speech system and speech intended for the meeting participants. Our current system requires the visitor to explicitly say “Go to Sleep” to deactivate the speech recognizer and “Wake Up” to reactivate it. The problem is that it is disruptive to a meeting to hear these commands. Consequently, we will be investigating alternative methods, such as the use of a mute button or gesture recognition, to act as agents allowing the activation and deactivation of the speech recognizer.

Our system serves as a proof of concept and clearly shows that we can use speech for remote equipment control. However, due to limitations in technology we were forced to limit the selection of possible commands in order to increase the reliability of the system. This restriction diminishes the potential advantages of speech in interfaces. Hence, we will be investigating alternative methods by which we can assure the reliability of the system without sacrificing the potentially large set of commands. We will also investigate the use of speech recognizers that are speaker independent and accept long strings of words instead of short discrete utterances.

This research has demonstrated the possibility of seat changing through a speech interface. However, our media space offers many more services, such as control over the VCR. We are presently extending the system described here to allow interaction with additional services using a speech interface. One such extension involves the “video server attendant,” which would allow any visitor with a codec to directly connect to one of our nodes, without the need for any additional computer equipment. Currently, this is only possible for remote sites running our iiif software [1]. The attendant will also provide access to a video answering service and demos-on-demand.

5 Conclusions

We have successfully implemented a remote control-system, with no extra equipment at remote sites, that allows visitors to change seats using voice and video overlay technologies. We were also able to increase the usability of the system by exploiting spatial information offered by visual representation of the environment. Although extensive user testing has not been performed, preliminary feedback indicates that an increased sense of engagement was achieved by providing users with more control over their view of the environment. Further user studies need to be run, and we are in the process of expanding the Hyper Doorway system in several new directions.

Acknowledgments

We wish to thank Tracy Narine for his invaluable assistance with the Hyper Doorway equipment. We thank Radek Nowicki and Don-Jun Wang who assisted in development of the Hyper Doorway system. We also thank Alex Mitchell and Victor Ng-Thow-Hing, of the Dynamic Graphics Project, and Mike Ruicci of the University of Toronto CS Lab, for their many hours of help with the system. Finally, we thank the various users who provided feedback for our work.

About the Authors

Anuj Gujar is currently working on his M.Sc. at the University of Toronto in the field of Human-Computer Interaction. He completed his B.Sc. in Computer Science with a major in software systems at the University of New Brunswick in 1994. From January 1991 to May 1991 he worked at IBM’s advanced technology center and developed a client-support multimedia application allowing clients to access detailed information on their support staff. From September 1991 to December 1991 Anuj worked at the IBM Canada Limited’s TD tower, in Toronto, as a marketing assistant. During the summer months of 1992 he received a summer NSERC research assistantship at the University of New Brunswick, where he researched the use of Monte Carlo methods for solving large sets of linear equations. From January 1993 to August 1993 he worked at Bell-Northern Research’s Captive Office in Ottawa, Ontario. Currently funded by NSERC, Anuj’s research focuses on investigating the use of speech as an alternative and more intuitive input medium for human-computer interaction.

Shahir Daya completed his B.A.Sc. in Computer Engineering at the University of Toronto in 1995. He worked in IBM Canada’s Application Development and Maintenance department for 20 months as an internship student and will be joining IBM in their Systems Integration department this May. Shahir’s interests include all aspects of Software Engineering.

Jeremy Cooperstock is currently working towards the Ph.D. in Electrical and Computer Engineering at the University of Toronto. He received the B.A.Sc. in Computer Engineering from the University of British Columbia, Vancouver, in 1990 and the M.Sc. in Computer Science from the University of Toronto in 1992. From 1987 to 1988, he worked at IBM Research in Haifa, Israel, and in 1989, at the IBM T.J. Watson

Research Center in Yorktown Heights, New York. His research interests include reactive environments, learning in robotic and autonomous systems, communication in distributed systems, and competitive analysis of trading strategies.

Koichiro Tanikoshi is a researcher working at Hitachi Research Laboratory, Hitachi Ltd. He holds an M.S. in Information Science from the Tokyo Institute of Technology. He was a visiting researcher with the Input Research Group at the University of Toronto from April 1994 to April 1995. Koichiro is now interested in user interface customizations, ubiquitous computing and situated information.

Bill Buxton began his career as a musician. He became involved in designing electronic musical instruments, which drew him into the fields of computer graphics and user interface design. He is Principal Scientist - User Interface Research, at Alias Research Inc., a Toronto- based company specializing in computer graphics systems. He is also an Associate Professor in the Department of Computer Science at the University of Toronto where he is the Scientific Director of the Ontario Telepresence Project and the Input Research Group. Finally, Buxton has had a long term relationship as a consulting researcher at Xerox's Palo Alto Research Centre (PARC).

Buxton has published extensively and, with Ron Baecker, Johnathen Grudin and Saul Greenberg, is co-author/editor of the text Readings in Human-Computer Interaction: A Multi-Disciplinary Approach, published by Morgan-Kaufmann. He is also nearing completion on two new books for Cambridge University Press. One is Haptic Input to Computer Systems, a discussion of input techniques and technologies. The other, with Sara Bly and Bill Gaver, is The Use of Nonspeech Audio Displays in Human Computer Interaction.

References

[1] Buxton, W. and Moran, T. EuroPARC's Integrated Interactive Intermedia Facility (iiif): Early Experience, In S. Gibbs & A.A. Verinj-Stuart (Eds.). Multi-user interfaces and applications, Proceedings of the IFIP WG 8.4 Conference on Multi-user Interfaces and Applications, Heraklion, Crete. Amsterdam: Elsevier Science Publishers B.V. (North-Holland), pages 11-34, 1990.

[2] Cooperstock, J., Tanikoshi, K., and Buxton, W. Turning Your Video Monitor into a Virtual Window, Proc. of IEEE PACRIM, Pacific Rim Conference on Communications, Computers, Visualization and Signal Processing, Victoria, May 1995.

[3] Cooperstock, J., Tanikoshi, K., Beirne, G., Narine, T., and Buxton, W. Evolution of a Reactive Environment. Proceedings of CHI'95, Denver, Colorado, May 1995.

[4] Daya, S., Hyper Doorway - Controlling a remote system over a video link, Computer Engineering Undergraduate Thesis, University of Toronto, April 1995.

[5] Gaver, W. The affordances of media spaces for collaboration. Proceedings of CSCW'92.

[6] Gaver, W., Smets, G., and Overbeeke, C. (1995). A Virtual Window on Media Space. Proceedings of CHI'95, Denver Colorado.

[7] Riesenbach, R. The Ontario Telepresence Project, CHI'94 Conference Companion, pages 173-174.