ISSUES IN MEASURING THE BENEFITS OF MULTIMODAL INTERFACES

James Flanagan and Ivan Marsic

CAIP Center, Rutgers University
Piscataway, NJ 08855–1390
http://www.caip.rutgers.edu/multimedia/

ABSTRACT

Multimedia interfaces are rapidly evolving to facilitate human/machine communication. Most of the technologies on which they are based are, as yet, imperfect. But, the interfaces do begin to allow information exchange in ways familiar and comfortable to the human—principally through natural actions in the sensory dimensions of sight, sound and touch. Further, as digital networking becomes ubiquitous, the opportunity grows for collaborative work through conferenced computing. In this context the machine takes on the role of mediator in human/machine/human communication—the ideal being to extend the intellectual abilities of humans through access to distributed information resources and collective decision making. The challenge is how to design machine mediation so that it extends, not impedes, human abilities. This report describes evolving work to incorporate multimodal interfaces into a networked system for collaborative distributed computing. It also addresses strategies for quantifying the synergies that may be gained.

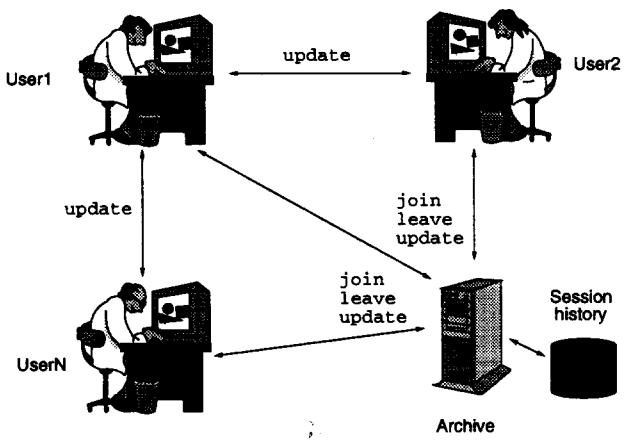

Figure 1: *Information flow in a distributed real–time collaborative system.*

1. INTRODUCTION

Human communication typically relies on multiple senses employed simultaneously. Information from parallel channels is fused for the task at hand. But, traditionally, we measure the performance of machine aids singly (usually in a one–human on one–machine scenario—such as assessing the word accuracy of an automatic speech recognizer). As multimodal human/machine interfaces become an integral part of computing systems, we must understand how to design systems to support the user's rich set of interaction skills—rather than requiring humans to adapt to arbitrary constraints of technology–driven designs. We expect that proper employment of multimedia technologies (cooperative and simultaneous) can deliver value greater than the sum of the parts. To arrive at prudent designs, we require means to measure the synergy gained from particular combinations. This measurement implies an application scenario, recognizing that synergy is generally task–dependent.

Toward this objective we are researching a multiuser data network constituting a *d*istributed *s*ystem for collaborative *i*nformation *p*rocessing and *l*earning (DISCIPLE). This network, illustrated in Figure 1, is presently implemented on internal Ethernet, and is evolving to an internal Asynchronous Transfer Mode facility. It permits collaborative access and manipulation of objects in shared workspace at geographically separate locations. Presently mouse and keyboard impose severe limits on the ability to interact. Therefore we are experimenting in DISCIPLE by combining several sensory interfaces we have developed in related research. These techniques include: (**sight**) eye–tracking, foveating sensing, and region–of–interest identification and extraction; (**sound**) automatic speech and speaker recognition, speech synthesis, and autodirective microphone arrays; (**touch**) gesture detection and position sensing, force–feedback transduction, and collaborative manipulation of tactile data. A scenario based on design of a research laboratory by three participants is projected for measuring the synergies in the combined, simultaneous use of these interface technologies (such as point and speak, look and speak, and grasp, look, and speak).

1.1. Multimodality in Multiuser Environments

Multimodal interfaces have special value for multiuser synchronous collaboration, where users manipulate objects in real–time. Real–time collaboration often involves multimedia objects whose geometric relationships must be explicated. Multimodal interaction is particularly suitable for this. It is powerful in domains having significant spatio–

COMPONENTS OF RESEARCH DESCRIBED HERE ARE SUPPORTED BY DARPA CONTRACTS NO. DABT63–93–C–0037, DAST63–93–0064 AND N66001–96–C–8510, AND BY NSF CONTRACTS IRI–9314946 AND MIP–9314625.

temporal components, such as manipulation of whiteboard objects. Multimodal interfaces permit straightforward expression of spatial relations and locations, gesturing, and the monitoring and coordination of activity.

A major goal of collaboration is to transmit one's ideas and concepts to other participants and to understand the intentions of all collaborators—and to reach joint decisions and actions.

Current systems for synchronous computer–supported collaborative work typically consist of a video/audio teleconferencing system and an electronic whiteboard. In this arrangement communication and data manipulation are *disjoint* processes. But communicated information should be extracted, summarized, and *integrated* with commands for manipulating the objects of collaboration. The extracted communication information should be incorporated into the task of manipulation and visualized at other locations, to provide collaborators the greatest benefit. What is desired is integration of "communication with" and "communication about" multimedia objects. This objective highlights the role of software agents that can intelligently fuse potentially unreliable, multiple sensory inputs to arrive at reliable decisions and actions.

2. RELATED WORK

Little prior work addresses measuring the benefits of multimodal interfaces because these technologies have not been available until recently. But integration of multiple modalities in human/machine interfaces has for a long time been viewed as a means for getting natural communication with machines [8]. Most recent studies deal with integration of only two modalities within a system: speech and sight [2, 5, 11], speech and gesture [3, 19], or text and images [22, 15]. Another example is the CAIP video/audio conferencing system which integrates sight and sound [9]. Efforts such as the ESPRIT Project "*Multimodal Integration for Advanced Multimedia Interfaces*" (MIAMI)[1], sponsored by the Commission of the European Communities and CMU Interactive Systems Laboratories [21, 23] also aim towards building integrated interfaces. Previous uses of the tactile or gesture modality have usually not included a force–feedback capability. But, this capability is necessary to effectively grasp, move, and place virtual objects [4]. A popular technique for studying multimodal interfaces is the Wizard of Oz technique [6, 7]. The basic idea is to model the interface, which is not yet or only partly available, by a human (the hidden "wizard") and to hide this fact from the user. At least one effort has aimed at automatic assessment of multimodal interfaces [6].

Separately, work has addressed the usefulness of computer–supported cooperative work [16, 17, 18, 24], and the utility of digital video in multiuser collaborative applications [12]. Our own studies in determining acceptable audio and video quality necessary to accomplish collaborative tasks found that proper allocation of resources and transmission bandwidth strongly depends on the nature of the collaborative work [9].

[1]Description available at
`http://www.nici.kun.nl/~miami/reports/reports.html`

To our best knowledge there has not been work on evaluating multimodal interfaces in collaborative multiuser environments.

3. GOALS AND METHODOLOGY

Building upon the multiuser collaborative system DISCIPLE, our effort is to incorporate sight, sound and touch modalities into each user terminal. The system is illustrated in Figure 2. An important component at the client is a software agent that is able to fuse simultaneous and complementary (and possibly redundant) sensory inputs to make reliable decisions and take appropriate actions.

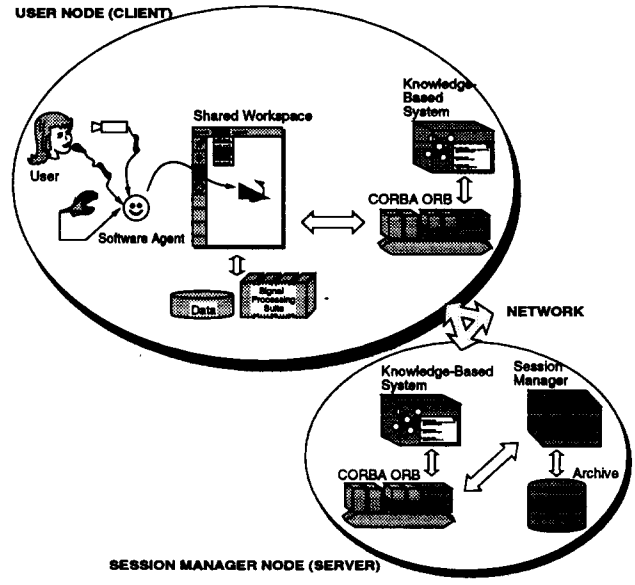

Figure 2: *Software architecture of the* DISCIPLE *system. A server supports several client nodes. Each client node incorporates sight, sound and touch modalities, whose outputs are fused by a software agent into a workspace command.*

If effectively used, these modalities eliminate the need for a menu bar and a tools palette on the whiteboard. Current software packages for complex tasks typically contain a maze of menu options that are user–interface operations. Very few of the operations do useful work. In our system, the whiteboard normally consists only of the shared workspace, as shown in Figure 3. The user position is supported by a resident signal processing suite and by local storage. A knowledge–based system moderates the work of an object communication infrastructure (CORBA–compliant Object Request Broker [20]).

3.1. Interface Technologies

We are incorporating several sensory technologies, established in related research, into the user interface:

sight: visual gesture from eye tracking and head position; region–of–interest recognition steered by foveated sensing [25], together with iconic reporting symbols and operators for managing their semantics [1];

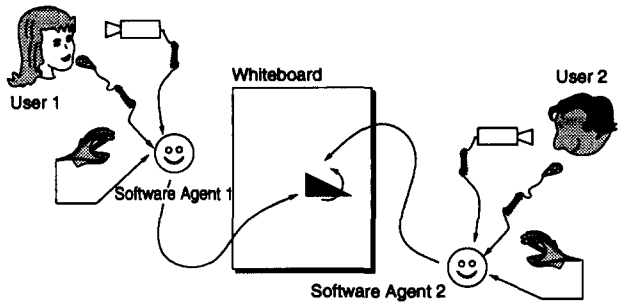

Figure 3: *Users give commands to software agents, which evaluate and fuse these commands and carry them out.*

sound: hands–free large–vocabulary speech recognition from microphone array and hidden Markov model (HMM) classifier [9]; speaker identification for user security; text–to–speech synthesis for answerback;

touch: position–sensing, force feedback data glove [4], for manual gesture, grasp and placement of data objects.

The system that we are implementing includes not only multimodal input, but also multimodal output in the form of screen presentation, speech answerback, and tactile force–feedback.

3.2. Benchmark Tasks

Central to assessments of multimodal synergies is the quantification of:

- *ease–of–use* (which implies making an application available to a broad range of users who may be non–technical. This resonates with the current theme of how to make computers as ubiquitous as TV's or microwaves [13]);
- *productivity increase* (to achieve shortening of the time needed to accomplish a task);
- *expressiveness* (to represent complex concepts and ideas).

While some significant research experience exists with the first goal, the last two goals remain less well understood. Our work focuses on the last two goals for multimodal interfaces. Here, different tasks present different levels of difficulty, e.g., direct manipulation with a well–defined goal vs. negotiation to reach a consensus.

An additional important parameter is the network bandwidth needed to support dynamic allocation of resources [14].

One issue in evaluation is measuring the benefits of multimedia communication in collaborative systems where experiments are set–up *ad hoc*. Normally, we would expect that richer interactions would lead to more productive and/or higher quality results. However, some previous studies have observed that multimedia dimensions do not improve the quality of interaction among remote participants [10]. Isaacs and Tang [12] conjectured that a key reason for this was that the studies were conducted among strangers who were asked to accomplish an artificial task. That is, the participants did not have working relationships with each other, and were not dealing with issues that motivated them. These authors also observed that digital video improves collaboration results for groups working on real tasks [12].

For an initial evaluation and preliminary exercise of a software agent for fusing sensory input, we are implementing only the sound and touch modalities (with speech recognition and synthesis and force–feedback tactile data glove). First trials for fusing the sensory data are projected in a single user environment, and then expanded to the network collaboration.

The collaborative task is taken as the layout of a laboratory/office room which three peers (presently separated geographically) are to occupy jointly for conducting circuit board debugging and validation. Audio and video communication are provided through the interface. They must reach consensus on layout and facilities, choosing components from a standard stock with a budget cap. Time to accomplish the solution and quality of the solution (as judged by experts) constitute the metrics. Parameters include single and double interface modalities, and network bandwidth. Comparisons are made to solutions obtained only with mouse and keyboard, with and without audio and video. Network capacity, and its fluctuations, impact the utility of multimodal interaction. Service demands can vary widely, as illustrated by the preliminary data in Figure 4. A separate software agent considers this network traffic and adjusts requests for resources accordingly.

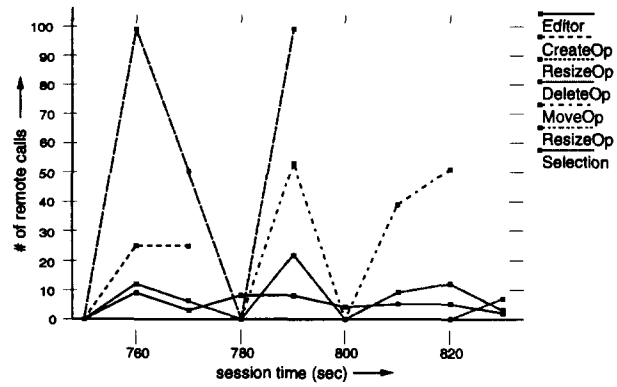

Figure 4: *Network statistics on service demands for software objects used in a collaborative session.*

In our case it is important to separate multiuser from single–user aspects of the experiments. The benefits of multimodality for the user's peers come from the multimodal output. The benefits of multimodal input are for the user alone, whereas for the user's peers the benefits come indirectly in the sense that they can more rapidly understand the user's actions. Delays and convoluted explanations, caused by limitations in the human/machine interface, are frustrating to collaborating users.

4. CONTINUING WORK

The preceding discussion emphasizes the need for metrics to assess the effectiveness and utility of multimodal inter-

faces, especially in the multiuser collaborative environment. As yet, there is little established consensus about methodologies and on the design of task scenarios for measurement. We expect the work described and projected will delineate central design principles and preliminary measures for quantifying the synergies that can be derived from simultaneous use of multiple sensory modalities.

Acknowledgments This research is part of a joint project conducted with Professors C. Kulikowski, P. Meer, J. Wilder, and G. Burdea in the CAIP Center of Rutgers University.

5. REFERENCES

[1] S. AbadMota, C. A. Kulikowski, R. Mezrich, L. Gong, S. Stevenson, A. Tria, and K. Klein. Iconic Reporting: A New Way of Communicating Radiological Findings. In *Proceedings of the Symposium on Computer Applications in Health Care*, page 915, November 1995.

[2] C. Benoit, M. Lallouache, T. Mohamadi, and C. Abry. A Set of French Visemes for Visual Speech Synthesis. In G. Bailly and C. Benoit, editors, *Talking Machines: Theories, Models, and Designs*, pages 485–504. Elsevier Science Publ., Amsterdam, 1992.

[3] R. Bolt. Put–That–There: Voice and Gesture at the Graphic Interface. *Computer Graphics*, 14(3):262–270, August 1980.

[4] G. Burdea, P. Richard, and P. Coiffet. Multimodal Virtual Reality: Input–Output Devices, System Integration, and Human Factors. *Int'l J. Human–Computer Interaction*, 8(1):5–24, January 1996.

[5] M. Cohen and D. Massaro. Synthesis of Visible Speech. *Behavior Research Methods, Instruments and Computers*, 22(2):260–263, 1990.

[6] J. Coutaz, D. Salber, and S. Balbo. Towards Automatic Evaluation of Multimodal User Interfaces. *Knowledge–Based Systems*, 6(4):267–274, Dec. 1993.

[7] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. Wizard of Oz Studies—Why and How. *Knowledge–Based Systems*, 6(4):258–266, December 1993.

[8] J. L. Flanagan. Technologies for Multimedia Communications. *Proceedings of the IEEE*, 84(4):590–603, April 1994.

[9] J. L. Flanagan. Multimodality. In R. Cole and J. Mariani, editors, *Survey of the State of the Art of Human Language Technology*, chapter 9, pages 277–300. National Science Foundation and Directorate General XIII of the European Commission, Cambridge University Press, (in press). Available at `http://www.cse.ogi.edu/CSLU/HLTsurvey`.

[10] S. Gale. Human Aspects of Interactive Multimedia Communication. *Interacting With Computers*, 2:175–189, 1990.

[11] C. Henton and P. Litwinovitz. Saying and Seeing it With Feeling: Techniques for Synthesizing Visible, Emotional Speech. In *Proc. ESCA '94*, pp.73–76, 1994.

[12] E. A. Isaacs and J. C. Tang. What Video Can and Cannot do for Collaboration: A Case Study. *ACM Multimedia Systems*, 2(2):63–73, August 1994.

[13] S. J. Johnston. Out in Front: A Candid Conversation with Bill Gates. *Information Week*, pages 14–17, November 11 1996.

[14] Q. Lin and K. Srinivas. Infrastructure Support for Multimedia Communication: A Survey. *International Journal of Intelligent and Cooperative Information Systems*, 3(2):189–202, June 1994.

[15] P. McKevitt, editor. *Integration of Natural Language and Vision Processing: Computational Models and Systems.* Kluwer Academic Publ., Dordrecht, The Netherlands, 1995.

[16] R. Ocker, S. Hiltz, and J. Fjermestad. The Effects of Distributed Group Support and Process Structuring on Software Requirements Development Teams: Results on Creativity and Quality. *J. Management Information Systems*, 12(3):127, Winter 1995.

[17] G. M. Olson, J. S. Olson, M. R. Carter, and M. Storrosten. Small Group Design Meetings: An Analysis of Collaboration. *Human–Computer Interaction*, 7(4):347–374, 1992.

[18] J. S. Olson, G. M. Olson, M. Storrosten, and M. R. Carter. Groupwork Close Up: A Comparison of the Group Design Process With or Without a Simple Group Editor. *ACM Transactions on Information Systems*, 11(4):321–348, October 1993.

[19] R. Sharma, T. S. Huang, and V. I. Pavlović. Multimodal Framework for Interacting With Virtual Environments. In C. E. Ntuen and E. H. Park, editors, *Human Interaction With Complex Systems: Conceptual Principles and Design Practice*, pages 53–71. Kluwer Academic Publ., Dordrecht, The Netherlands, 1996.

[20] The Object Management Group. The Common Object Request Broker: Architecture and Specification. Technical Report 96-03-04, Object Management Group, Inc., Framingham, MA, July 1995. Revision 2.0.

[21] M. Vo, R. Houghton, J. Yang, U. Bub, U. Meier, A. Waibel, and P. Duchnowski. Multimodal Learning Interfaces. In *Proc. ARPA Spoken Language Technology Workshop*, Austin, TX, January 1995.

[22] W. Wahlster. One Word Says More Than a Thousand Pictures: On the Automatic Verbalization of the Results of Image Sequence Analysis Systems. *Computers and Artificial Intelligence*, 8:479–492, 1989.

[23] A. Waibel, M. Vo, P. Duchnowski, and S. Manke. Multimodal Interfaces. *Artificial Intelligence Review*, 10(3–4), 1995.

[24] S. Whittaker, E. Geelhoed, and E. Robinson. Shared Workspaces: How Do They Work and When are They Useful? *International Journal of Man–Machine Studies*, 39(5):813–842, November 1993.

[25] J. Wilder, L. Xiao, and W. Kosonocky. A Foveating Sensor for High Speed Data Capture and Pattern Recognition Invariant to Translation, Rotation and Scale. In *Proc. Smart Focal Plane Workshop*, pages 32_1–32_15, Fort Belvoir, VA, July 1993.