IS TALKING TO VIRTUAL MORE REALISTIC?

Luc Julia and Adam Cheyer
Computer Human Interaction Center (CHIC!) - SRI International
333 Ravenswood Avenue, Menlo Park, CA 94025, USA
{luc.julia,adam.cheyer}@sri.com
http://www.chic.sri.com

ABSTRACT

Virtual worlds and animated computer avatars are becoming more realistic, more natural, and more widespread. Accordingly, we are looking at new ways of interacting with machines based on "old" methods for interacting with humans, such as talking, writing and gesturing. By applying a synergistic, multimodal approach to several application domains that incorporate avatars, augmented reality or virtual reality, we investigate whether this interface style is more suitable for realistic environments. Concrete examples and studies are used to discuss this point, raising other key questions in the design of this new generation of interfaces.

Keywords: Augmented Reality, Avatars, Virtual Worlds

1. INTRODUCTION

Several speech-enabled applications developed at SRI International explore the use of Avatars, Augmented Reality, and Virtual Reality for novel human-computer interfaces.

It is easy to assume that the realistic look and feel of virtual worlds will well accommodate an everyday life metaphor and will ease multimodal synergistic dialogs. It seems obvious that talking, gesturing, and drawing in such an environment is more natural than doing the same thing in front of desktop computers equipped with mice and keyboards. But is this assertion true? Are the virtual worlds we are building detailed enough? Intelligent enough? Comfortable enough? Or too comfortable for the kind of recognition rates we can achieve? Or too augmented, implying new paradigms? So intelligent that we are now wondering who is leading the dialog?

To begin to answer these questions, we created a number of prototype systems, which for each we discuss

- whether the use of these new modalities is appropriate
- how the integration, fusion, and disambiguation of the modalities such as speech and gesture are handled
- what are the relationships between the inputs and outputs
- limitations and expectations for each system

2. APPLICATIONS

2.1 InfoWiz

2.1.1 Description

The InfoWiz project [Figure 1] is centered around the idea of putting an interactive kiosk into the lobby of SRI. People who have a few minutes to spend will be able to learn about the institute, enjoy themselves, and walk away with a good feeling of having seen something interesting and unusual.

Figure 1: The Kiosk for InfoWiz

2.1.2 Modalities and Media

InfoWiz is a voice-only application, using a telephone handset because it is a familiar device that allows close-talking speech recognition. The animated character, in addition to talking back to the user (through text-to-speech), provides multimedia information by manipulating the contents of a Web browser, very much like in [2]. There is no fusion of modalities, but a conversational dialog manager [1] attempts to maintain a certain illusion of intelligence.

2.1.3 Comments and Issues

Compared with a touch-screen-based system, this approach is definitely more convivial. Users have the feeling they are interacting with a guide without having to look for the right path. They can jump right to a point of interest without having to go through multiple levels of menus that don't always fit their own representation of the organization of the information. However, there is a trade-off that depends on the length of the dialog. The illusion of intelligence does not last very long: after a few minutes of amazement, the user begins to realize the limitations of the system, and the InfoWiz's inability to accurately answer many questions can generate frustration on the part of the user. As the InfoWiz Kiosk is generally intended for short interactions while a visitor is waiting in the lobby, the system succeeds relatively well. However, added intelligence, learning, and natural language understanding is needed for the approach to work well in other domains.

2.2 DemoMan

2.2.1 Description

DemoMan [Figure 2] is an anthropomorphic helper who can autonomously or semiautonomously lead a demonstration of multimodal systems [3]. DemoMan can replay from a logfile complex multimodal interactions (keystrokes, mouse clicks, pen drawings, or speech utterances), automating a prototype application that accepts these as inputs. The replay can be interrupted and resumed as needed by a human, or even by another virtual assistant.

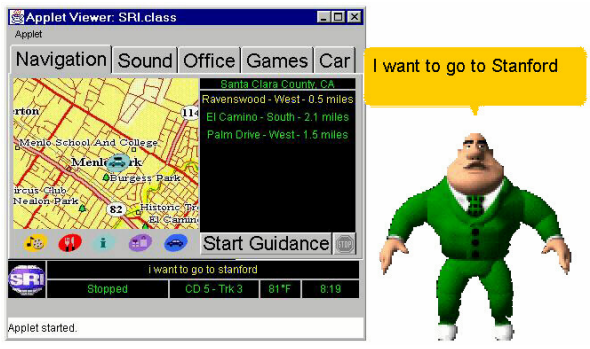

Figure 2: DemoMan, giving a CHIC! demonstration

2.2.2 Modalities and Media

One of the original features of DemoMan is that he interacts directly with an application exactly as a human user would. For example, DemoMan's voice feeds the application's speech recognizer while he gestures to recreate the very situation of synergistic multimodality. In addition, DemoMan runs his own speech recognizer to listen to the other demonstration assistant (human or virtual).

2.2.3 Comments and Issues

Watching DemoMan using an application through a fairly complex dialog, mixing spoken and deictic elements, in addition to his capability to listen to the external world while performing, gives the appearance of the completion of a cooperative task. But DemoMan is just an impersonator, reading from a logged script -- his lack of personality and proactivity reduces his dialog capabilities.

2.3 MMap

2.3.1 Description

The Multimodal Map (MMap) [Figure 3] provides an interactive display for collaborating with a set of intelligent agents able to display dynamic, multimedia information coming from the Web in response to user requests [4].

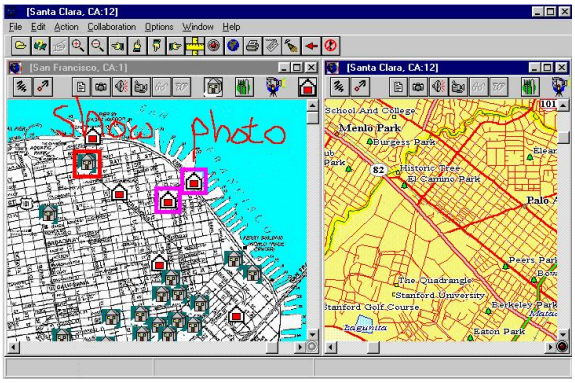

Figure 3: Multimodal, Pen and Voice, Map

2.3.2 Modalities and Media

This augmented map reacts to spoken, written, and handdrawn (multimodal synergistic) inputs produced by the user. In response to queries, the MMap generates and displays multimedia documents on its surface and using media viewers.

2.3.3 Comments and Issues

The augmentation of the pen and paper metaphor by the production of relevant documents such as pictures, videos, or sounds does not hurt the inter-

action since the user does not have to interact with them -- they are just addenda. Most of the previous experiments [5], as well as our preliminary study, [6] show that novice or expert users are taking advantage of the synergy between pen and voice inputs in a map environment.

2.4 3D/VRML Extension to MMap

2.4.1 Description

MMap's 3D/VRML extension [Figure 4] provides a way to synchronize the 2D navigation with a 3D experience [7].

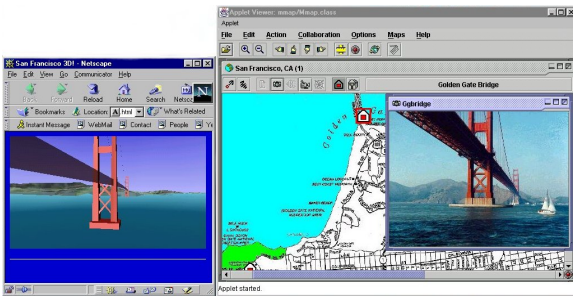

Figure 4: 3D MMap: a new navigation dimension

2.4.2 Modalities and Media

Here again, speech, 2D gestures, drawing, and handwriting are the main input modalities. A multimedia reactive map display and a 3D viewer of the matching VRML world are provided as an additional output.

2.4.3 Comments and Issues

the 3D/VRML additional degree of freedom provides new presentation and navigation features along with a larger vocabulary and set of commands to control the application, which may result in new ambiguities.

Also, since it is difficult to navigate a 3D world using a 2D metaphor (pen and paper), we feel the need to introduce some sort of 3D gesture. Augmenting the reality with too many virtual features can decrease the interaction proficiency and thus its reality.

2.5 Office MATE and Travel MATE

2.5.1 Description

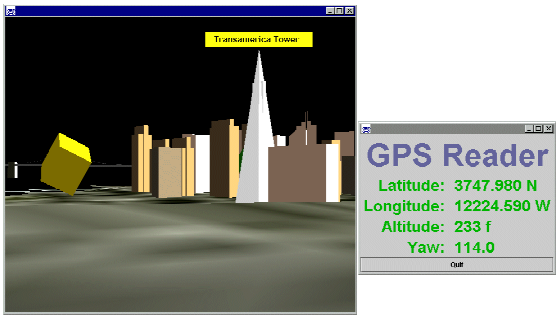

Figure 5: Travel MATE augments the reality

A series of augmented reality applications, Travel MATE [Figure 5] and Office MATE (MATE stands for "Multimodal Augmented Tutorial Environment") [8] try to correct the paradigm shift discussed with 3D MMap. In these applications, the user is immersed in a 3D, volumetric, world, avoiding any reference to 2D.

2.5.2 Modalities and Media

Navigation, which was a problem in the 3D MMap, is taken care of by the position and the orientation of GPS and compass sensors. The position of the user in the space becomes an input modality. The user can interact with the surrounding objects by talking to the world. The presentation features in the 3D world are reactive, displaying labels and talking back to the user to convey additional information.

2.5.3 Comments and Issues

Tracking the user's position in the world allows the system to be proactive, but we need a pointing device to complement the speech input.

3. CONCLUSIONS

We can definitely say that talking and gesturing within 2D or 3D virtual worlds that look real provides a more realistic communication paradigm than typing on keyboards and moving mice to get the job done. From the informal studies we already conducted, we know that user expectations when talking within a virtual world are the same as when talking within this real world. But in the systems we presented, although they take steps toward exploring natural interactions for virtual and augmented reality worlds, there are still many deficiencies. Some fail because of a lack of "intelligence", some because the domain is too broad, others, in narrower domains, through lack of focus, thus overwhelming the user and breaking envi-

ronmental rules of the real world. For each of them, we are planning to continue to collect data through user studies to improve the basic components on which they are built (analyzers, recognizers, and fusion techniques) and to try to define some limits for augmented reality. Augmenting reality does not mean inventing another reality.

4. REFERENCES

[1] Cheyer A. and Julia L. (1999), InfoWiz: An Animated Voice Interactive Information System. *Proceedings of Agents'99 (WS Communicative Agents)*, Seattle.

[2] Andre E., Rist T. and Muller J. (1998), Guiding the User Through Dynamically Generated Hypermedia Presentations with a Life-Like Character. *Proceedings of International Conference on Intelligent User Interfaces 98*.

[3] Dubreuil J. and Julia L. (1998), Collaborative Use of Multimodal Applications by Lifelike Computer Characters and Humans. *Poster session at Lifelike Computer Characters '98*.

[4] Cheyer A. and Julia L. (1995), Multimodal Maps: An Agent-based Approach. *Proceedings of CMC'95*, pp. 103-113.

[5] Oviatt, S. (1996), Multimodal Interfaces for Dynamic Interactive Maps. *Proceedings of CHI'96*, pp. 95-102.

[6] Cheyer A., Julia L. and Martin J.C. (1998), A Unified Framework for Constructing Multimodal Experiments and Applications. *Proceedings of CMC'98*, pp. 63-69.

[7] Julia L., Cheyer A., Dowding J., Bratt H., Gawron J.M., Bratt E. and Moore R. (1998), How Natural Inputs Aid Interaction in Graphical Simulations? *Proceedings of VSMM'98*, pp. 466-468

[8] Julia L., Bing J. and Cheyer A. (1999), Virtual Spaces Revive Real World Interaction. *At first EC/NSF Advanced Research Workshop "Research Frontiers in Virtual Environments and Human-Centred Computing"*, to appear.