NETWORKED COLLABORATIVE ENVIRONMENT WITH ANIMATED 3D AVATARS [1]

Wing Ho Leung and Tsuhan Chen
Dept. of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA
Email: {wingho,tsuhan}@ece.cmu.edu

Abstract - In this demonstration, we present a prototype networked collaborative environment. This prototype supports audiovisual communication among several terminals. The voice is transmitted over the network and used to animate 3D avatars that represent the users in a virtual environment. Intelligent interfaces create a virtual environment in which users can interact with the sense of immersion.

INTRODUCTION

Our prior research has shown that the quality of audiovisual communication, such as in video telephony and videoconferencing, can be improved by combining speech and image processing techniques in an unconventional way. In particular, we have shown that we can improve the lip synchronization by utilizing speech recognition followed by facial image analysis and synthesis [1]. Based on these results, we build a networked collaborative environment to enable people to communicate and cooperate with each other from anywhere. By building intelligent interfaces into the terminals, we create an environment of the sense of immersion, so that the users can ignore the presence of the physical network and concentrate on the task at hand.

TECHNOLOGIES

In a traditional videoconferencing setup, each participant sees other participants in separate windows, and the voices are all mixed together. There is no sense of immersion, and it is often difficult for the participants to tell who is talking, and talking to whom. A more appealing environment can be obtained by dissecting individual images of the participants from their background and reassemble them around a virtual conference table. Furthermore, 3D avatars can be rendered to represent users in the virtual environment to replace live video. The positions of the 3D avatars in the virtual environment determine what each user sees on the display, and hears from the speakers. Also, the face orientation of these 3D avatars will reflect who is talking to whom in this collaborative environment. Directional sound and position-responsive sound technologies can be used to render a realistic and immersive sound environment. In addition, these 3D avatars can be animated faces that are lip-synched by the human voice. Other useful techniques include text-to-speech synthesis, speech recognition, sign language analysis/synthesis, and language translation. Lip-reading and lip-sync

[1] IEEE Multimedia Signal Processing Workshop, Los Angeles, Dec 1998

animation are also essential for people who are hearing impaired or with other language problems. Finally, a video display can respond to the position of the user. The stereoscopic display technology can provide a real 3D visual experience.

DEMO

We demonstrate a prototype system that supports three or more terminals, as shown in Figure 1, with the following capabilities:

1. One of the terminals generates a virtual conference table and virtual people around it. We call it the main terminal and others auxiliary terminals. Each terminal is equipped with a microphone and a speaker. Since the system supports three (or more) terminals, the main terminal creates two (or more) virtual people around the table.

2. The speaker at an auxiliary terminal may specify whether he/she wants to talk to a specific person or to the public by pressing a button. If the speaker chooses to speak to the user at another auxiliary terminal instead of the user at the main terminal, then the virtual face of the speaker will turn to the virtual face of the corresponding person. On the other hand, if the speaker chooses to speak to the public, then the virtual face of the speaker will face each audience in turn to maintain eye contacts with them.

3. The mouth of the speaker's virtual face moves in sync with the voice so that listeners know who is speaking.

4. Different proportions of sound will be steered among the left and right sound channels so that the listener can feel which person is talking by listening to the direction of the sound.

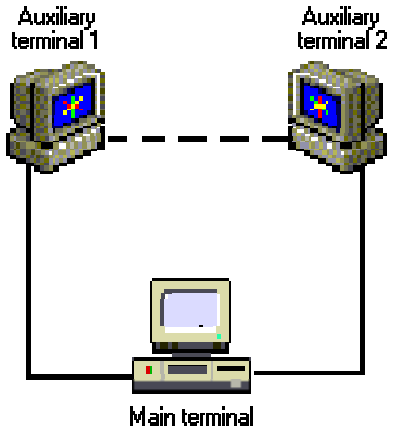

Figure 1 Demo setup.

Figure 2 gives an example of the screen of the main terminal. In this system two auxiliary terminals are connected to the main terminal. Here the left person is talking to the right person, therefore the user at the main terminal sees that the speaker is looking at the listener and the listener is looking at the speaker.

Figure 2 Virtual conference table.

CONCLUSION AND FUTURE WORK

As technology becomes more and more advanced, a networked collaborative environment is essential to enable people to have a more effective communication. The demo provides an example of how intelligent interfaces can be implemented to provide a virtual collaborative environment. With this technology, people from different parts of the world can work together efficiently and are not limited by their actual physical locations.

For the future work, we will use eye tracking to determine which person the speaker wants to talk to. Besides, face tracking will be performed to determine the position of the user. This information is used to define the projection of the 3D objects on the screen so that the user will have a more realistic 3D visual environment. For user who does not have a microphone, text-to-speech synthesis will be used and an animated face can be generated.

REFERENCE

[1] T. Chen and R. Rao, "Audio-visual integration in multimodal communication," *Proceedings of IEEE, Special Issue on Multimedia Signal Processing*, pp. 837-852, May 1998.