Face-to-Face Communication System in Cyberspace by Voice Driven Avatar

Shigeo MORISHIMA, Tatsuo YOTSUKURA and Eishi FUJII

Morishima-lab., Seikei University

Abstract

Recently computer can make cyberspace to walk through by an interactive virtual reality technique. An avatar in cyberspace can bring us a virtual face-to-face communication environment. In this paper, we realize an avatar which has a real face in cyberspace and construct a multi-user communication system by voice transmission through network. Voice from microphone is analyzed and transmitted, then mouth shape and facial expression of avatar are synchronously estimated and synthesized on real time.

1. Introduction

Recently, virtual reality technology is focused to produce cyberspace in which user can chat and give cooperative work through network. The final goal is to make the virtual space close to the real communication environment between network users. In this paper, multi-users virtual face-to-face communication environment in cyberspace is presented. There is an avatar projecting the feature of each user in cyberspace which has a real texture mapped face to generate facial expression and can give action and move by user's control. User can get a view in cyberspace through the avatar's eyes, so he can communicate with other people by gaze crossing. And also user's transmitted voice can control the lip shape and facial expression of avatar on real-time by our media conversion algorithm[1][2].

2. Modeling of Avatar

To generate a realistic avatar, a generic face model is manually adjusted to user's frontal face image to produce a personal face model and all of the control rules for facial expressions are defined as a movement of grid points in a generic face model (Figure 1). Synthesized face is coming out by texture mapping of user's frontal image onto the modified personal face model. The body of avatar is simply modeled as a balloon. User's emotion condition can be transmitted to other users as a feature and motion of balloon as well as facial expression. Figure 2 shows an avatar balloon model on which user's face is located.

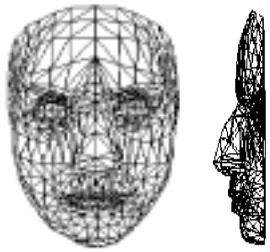


Figure 1: *Face model*

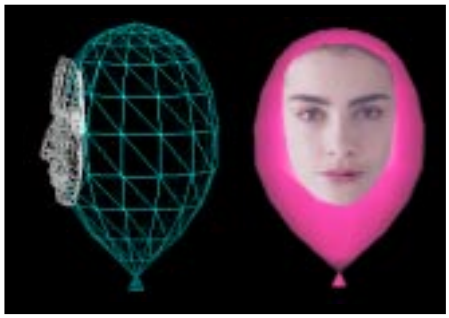


Figure 2: *Avatar model*

3. System Feature

Each process of a 3-users' communication system in cyberspace works as follows (Figure 3).

1) Voice Capturing

At client system, on-line captured voice of each user is A/D converted by 16KHz and 16bits, and is transmitted to server system frame-by-frame through network.

2) Voice Analysis and Parameter Conversion

At server system, voice from each client is phonetically analyzed and converted to mouth shape and expression parameters. LPC Cepstrum parameters are converted into mouth shape parameters by neural network trained by vowel features. Figure 4 shows neural network structure for parameter conversition and Figure 5 shows locations of control points around mouth. Figure 6 shows the example mouth shape parameter vowel "a". Emotion condition is decided by LPC Cepstrum , Voice Power and Pitch frequency using another neural network into Anger, Disgust, Happiness, Fear, Sadness and Surprise (Figure 7). Each basic emotion has a specific facial expression parameters described by FACS (Facial Action Coding System)[3].

Figure 3: *A photograph communication system in cyberspace*

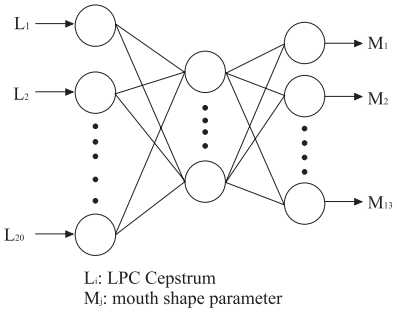


Figure 4: *Neural network for parameter conversion*

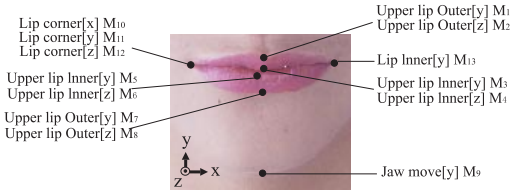


Figure 5: *Mouth shape parameters*

Figure 6: *Mouth shape parameter for vowel "a"*

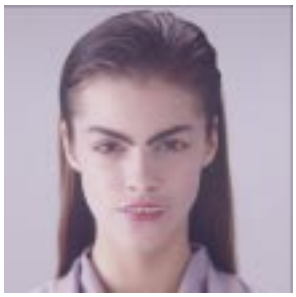


Figure 7a: *Basic emoion "Anger"*

Figure 7b: *Basic emoion "Disgust"*

Figure 7c: *Basic emoion "Happiness"*

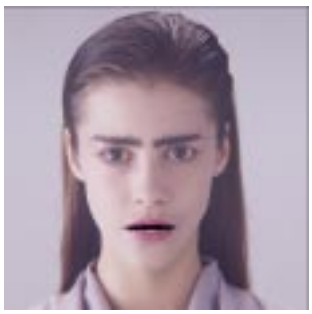


Figure 7d: *Basic emoion "Fear"*

Figure 7e: *Basic emoion "Sadness"*

Figure 7f: *Basic emoion "Surprise"*

3) Location Control

Each user can walk through and fly through cyberspace by mouse control and current locations of all users are always observed by server system. Avatar image is generated in the client space by the location information from the server system.

4) Emotion Key-in

Emotion condition can always be decided by voice, but sometimes user give his avatar a specific emotion condition by pushing function key. This process-works with first priority. For example, push anger and then red face and bigger balloon are coming out (Figure 8). When happiness, bouncing balloon is coming out, and so on.

5) Information Management at Server

Location information of each avatar, mouth shape parameters and emotion parameters are transmitted every 1/30 seconds to client system. Distance between every 2 users are calculated by the avatar location information, and voice from every user except himself is mixed and amplified with gain according to the distance. So the voice from the nearest avatar is very loud and one from far away is very small.

6) Agent and Cyberspace Generation at Client

Based on facial expression parameters and mouth shape parameters, avatar face is synthesized frame by frame. And avatar body is located on cyberspace according to the location information. There are two modes for displaying, view from avatar's own eyes (Figure 9) and view from sky (Figure 10) which can be chosen by menu in window.

7) Voice Output

Playback volume of an avatar's voice depends on the distance to that avatar. To add multiple speakers system make 3D audio output possible. To realize lip synchronization, 64ms delay is given to voice playback.

Figure 8: *Balloon "Anger"*

Figure 9: *Eye contact*

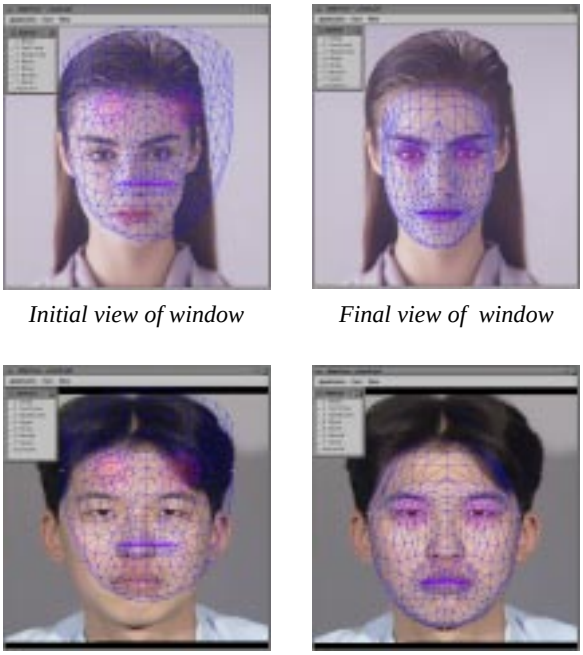
Figure 10: *View from sky*

4. Speaker Adaptation

When new speaker comes in, his face model and voice model have to be registered before operation. In case of voice, new learning for neural network has to be performed ideally. However, it takes a very long time to get convergence of backpropagation. To simplify the face model construction and voice learning, the GUI tool for speaker adaptation is prepared.

1) Face Model Fitting

To register the face of new user, a generic 3D face model is modified to fit on the input face image. Only 2D frontal image is needed. Figure 11 shows the initial and final view of fitting tool window. Some of the control points on face model are shifted manually. It takes a few minutes to complete user's face model because of the easy mouse operation by GUI tool. Expression control rules are defined onto the generic model, so every user's face can be equally modified to generate basic expression using FACS based expression control mechanism.



Initial view of window *Final view of window*

Initial view of window *Final view of window*

Figure 11: *Fitting tool window*

2) Voice Adaptation

75 persons' voice data including 5 vowels are pre-captured and database for weights of neural network and voice parameters are constructed. So speaker adaptation is performed by choosing the optimum weight from database. Training of neural network for every 75 persons' data is already finished before. When new non-registered speaker comes in, he has to speak 5 vowels into microphone before operation. LPC Cepstrum is calculated for every 5 vowels and this is given into the neural network. And then mouth shape is calculated by selected weight and error between true mouth shape and generated mouth shape is evaluated. This process is applied to all of the database one by one and the optimum weight is selected when the minimum error is detected.

5. Conclusion

Natural communication environment between multiple users in cyberspace by transmission of natural voice and real-time synthesis of avatar's facial expression is presented. Synthesis speed of cyberspace and avatars is about 10.5 frame per second by SGI Onyx2 (R10k, 180MHz). Current system is working on 3 users and intra-network environment. To increase the number of users, it's necessary to reduce the traffic in network by compressing voice signal and reduce the cost of server processing. Our final goal is to realize the system on Internet environment.

Reference

[1]Sigeo Morishima, H.Harashima: "A Media Conversion from Speech to Facial Image for Intelligent Man-machine Interface", IEEE journal of Selected Areas in Communication Special Issue on Human Interface in Telecommunication Vol.9, No.4, pp.595-600, 1991

[2]Sigeo Morishima: "Virtual Face-to-Face Communication Driven by Voice Through Network",Workshop on Perceptual User Interfaces, pp85-86, 1997

[3]Ekman and W.V.Friesen: "Facial Acion Coding System", Consultiong Psychologist Press, 1977