A 3D Virtual Environment for Social Telepresence

Steve DiPaola
Technical University of British Columbia
10153 King George Hwy
Surrey, BC V3T 2W1 CANADA
+1 604 924-6162
steve@dipaola.org

David Collins
Adobe Systems Incorporated
345 Park Avenue
San Jose, CA 95110 USA
+1 408 536 4882
david.collins@adobe.com

ABSTRACT
We examine OnLive Traveler as a case study. Traveler is a client-server application allowing real-time synchronous communication between individuals over the Internet. The Traveler client interface presents the user with a shared virtual 3D world, in which participants are represented by avatars. The primary mode of communication is through multi-point, full duplex voice, managed by the server.

Our design goal was to develop a virtual community system that emulates natural social paradigms, allowing the participants to sense a tele-presence, the subjective sensation that remote users are actually co-located within a virtual space. Once this level of immersive "sense of presence" and engagement is achieved, we believe an enhanced level of socialization, learning, and communication are achievable. We examine a number of very specific design and implementation decisions that were made to achieve this goal within platform constraints. We also will detail some observed results gleaned from the virtual community user-base, which has been online for several years

Keywords
Avatars, WWW, voice, 3D, virtual environments, virtual worlds, group communications

1. INTRODUCTION
Traveler is a multi-user, voice-enabled VRML browser. It employs 3D environments and avatars with complex facial animations to provide a platform for synchronous, multi-point voice communications [1,2]. The goal in developing Traveler was to deliver a rich and compelling experience of human socialization, using a common consumer PC platform over the World Wide Web. With this goal in mind, a number of very specific design and implementation decisions were made to achieve the intended level of free-form socialization, while operating within the platform constraints. These will be examined and evaluated in detail.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission.

Western Computer Graphics Symposium, March 24-27, 2002, Vernon, BC, Canada.

Copyright 2002 Steve DiPaola and David Collins.

Figure 1. Group voice chatting in Traveler via lip-sync emotive avatars.

1.1 The Experience
In Traveler, users are immersed in a shared 3D world, with first-person perspective. Each user is able to navigate with six degrees of freedom, and each sees the other participants as fully modeled 3D characters. As a user speaks, his or her voice emanates from the corresponding avatar on each of the other clients. The avatar's lips and facial structure synchronize with the words spoken and sound of the voice is distance-attenuated and spatialized in stereo, according to its position in the 3D world relative to the local user.

The voice communication in Traveler is full-duplex and fully multi-point, i.e. the user is receiving audio while speaking and multiple streams of voice audio are delivered to each client. The overall effect of the voice delivery in conjunction with the visual environment is that of a virtual "cocktail party". Users spontaneously form and re-form conversational subgroups, using natural social conventions. Initial contacts are made using natural-world methods, such as saying "Hello, there!" towards another avatar and waiting for the other participant to turn and orient (in response to the spatialized audio cue) before proceeding.

2. GOALS AND DESIGN POINTS
Traveler was developed in response to a number of basic observations. The first was that the developing Internet was a popular medium for multi-point chat and that real-time group communication was a swiftly growing category of application. Witness the relative success of America On-line, which

concentrated on providing group chat over competitor CompuServe that focused on delivery of services. The second observation is that interleaved lines of user-typed text are a low-grade simulation of real-world social phenomena in which group communication takes place. Text-chat rooms were used to implement parties, common interest clubs, debates, discussion groups. These traditional group communication forums were simulated using what is essentially a highly artificial format that was suited to a low-bandwidth medium and that required the cognitively taxing processes of typing in real-time while simultaneously extracting multiple interleaved threads of text.

The goal in developing Traveler was to produce an intuitive communication format, which offered existing chat users a compelling experience and potential new users would find less intimidating. By including as many "organic" channels of information as the bandwidth of dial-up Internet access would allow, Traveler would allow a novice user to effectively engage in group communication by relying on intuition developed in real-world social circumstances

2.1 Voice

The basic hypothesis in implementing Traveler was that the use of human voice is the most natural way to carry on shared conversation. The implementation of an effective multi-voice audio environment was the primary design target. This bears some emphasis, since all the other aspects of the Traveler interface, including the 3D environment, were implemented in support of this goal. Implementing distributed voice over the Internet introduced an enormous amount of complexity to the implementation of Traveler, but it was considered essential for a number of reasons. As opposed to text chat, the use of voice leaves the hands free for use in navigating the 3D environment. The user is freed from the cognitively taxing task of extracting a stream of conversation from interleaved threads, while simultaneously typing a response. The visual focus of attention is on the other users and the non-vocal queues that expressed through their avatars, while the audio focus is on their voice. Finally, the human voice is tremendously rich in the layers of meaning expressed, beyond the simple stream words. Inflection and timing inject meaning into a sentence that is very hard to include in simple text. Witness the difficulty involved in introducing an ironic tone into an e-mail and the possibility for misinterpretation that one risks in the attempt.

Traveler was intended to allow for a virtual multi-way conversation, with participants contributing randomly, spontaneously and in arbitrarily shifting combinations. Since the use of voice in communication is fundamentally an interactive one, it was considered essential to allow for interjections, overlapping commentary, encouraging responses and other natural elements of verbal communication. To achieve this, it was necessary to provide a mixed stream of audio on the down channel. To create this effect in a limited bandwidth environment, Traveler provides each client with up to two audio channels on the downlink, chosen from all the available up-linked audio streams. Each client receives a different set of audio channels, based on a number of heuristics, taking into account proximity to other speakers and which of the other participants are speaking at any given moment. Since the downlink stream set is reevaluated every 60 milliseconds, the resulting voice environment appears to be perfectly fluid and arbitrarily complex.

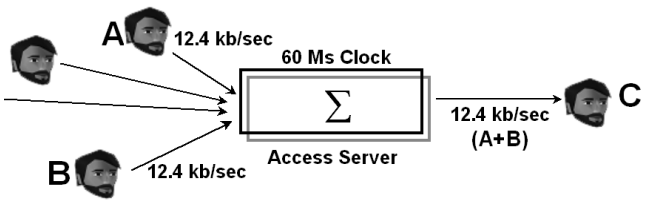

Figure 2. Multi-point, full duplex voice codec with additive bridging where the server can add incoming compressed streams from speakers A and B; f(A) + f(B) = f(A+ B).

The richness of the audio environment is further enhanced by localization of the audio data. The client software uses its knowledge of the relative positions of avatars to individually attenuate and stereo-locate the corresponding voice channels. This allows the user A to manage the influence of their voice on individuals and groups by approaching or retreating from other avatars. For example, a user listening to another speaker is vaguely aware that another group of users are speaking at some distance away. As a member of the distant group approaches, his or her audio becomes increasingly loud, combined with the voice of the original speaker. The direction and distance of the new speaker can be intuited from the attenuation and stereo queues.

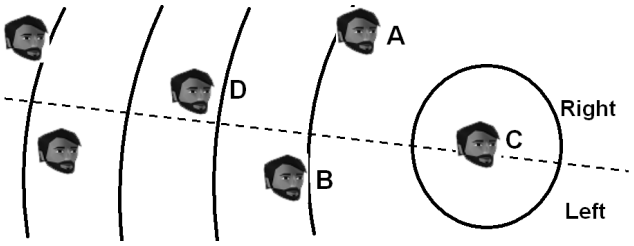

Figure 3. With Spacialized (3D) multi-point audio, avatar C hears others with distance attenuation, and stereo positioning, hearing avatar A louder and more to the right than avatar B.

By seamlessly combining these various audio techniques, Traveler provides users with a broad range of natural social behaviors in the shared environment. Mixed audio allows users to interrupt and interject as well as defer or refuse to defer to new speakers. These behaviors are all managed with standard social conventions, as opposed to artificial techniques, such as HAM-radio-style queues. Spatialized mixing allows natural and fluid formation of groups as well as smooth transition from one group to another

2.2 3D Space

The intention in integrating an immersive 3D environment into the Traveler experience was in large part to give the user intuitive tools for managing the audio experience. Frames of reference are required for navigating among conversational sub-groups. The use of distance and orientation in a social context are only meaningful in a spatial environment. In addition to providing familiar paradigms for managing voice interaction however, the shared virtual space also allows the user other quasi-organic channels of communication. Using basic navigation skills, the user can implement common gestures as non-verbal communication (e.g.

nodding, cocking the head, turning away in disgust, etc.) Dancing and exuberant motion can be used in conjunction with voice to enrich emotional expression. Landmarks can also be used to rendezvous with other individuals for planned events.

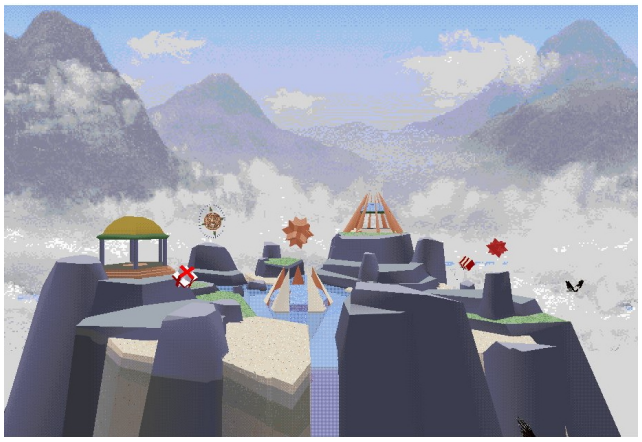

Figure 4. Condor Summit space has several social staging areas and use visuals and sound to convey an ethereal setting.

The 3D environments in Traveler also provide a symbolic and thematic background for communication. While text chat rooms are generally categorized by naming conventions, the experience of using each space is identical and the theme is maintained only by the combined consent of the group to speak on topic. Traveler adds the dimension of in-world thematic elements (e.g. space ships in a science fiction world, chairs and a podium in a business conference space, a molecular model for a scientific discussion, etc.) While the content of the speech is still determined by the participants, the world has a persistent thematic suggestion. In addition to theme, mood and setting can be suggested in a space, through the use of architectural and sculptural elements as well as background textures. Environmental audio in the form of randomly cycling layers of sound or music can also be added to a space to provide greater depth of experience.

In Traveler, facilities for interaction between the avatars and the space are limited. However, objects can be authored to produce spatialized audio in response to proximity to an avatar. Thus for example an object representing an information kiosk could give audio information to an avatar. Also objects can act as links to other Traveler spaces or as triggers to launch a web page.

2.3 Avatars and Facial Animation

In an attempt to further enhance the organic feel of the Traveler experience, the decision was made to implement avatars as smoothly morphing 3D models that animate in response to the user's voice. Usually, as in the case of anthropomorphic avatars, this animation takes the form of synchronizing the movement of the jaws and lips to the phonemes used by the speaker. This creates the profound illusion of a human face in the process of producing speech. This animation helps the user to determine which avatar in the field of view is speaking and adds to the overall illusion of being in the direct presence of living, conscious creatures. This same morphing technique is used to implement blinking, breathing, changes in emotional state and other lifelike sequences to further enhance the subtle impression of life in the avatars. Because Traveler avatars showcase the face so prominently, this organic effect is highly resonant with users, due to the extreme psychological and neuropsychological importance of the face to the human psyche. Users have reported a desire to maintain eye contact and to feel the effects of personal space during a Traveler session, indicating a high level of immersion in the social environment.

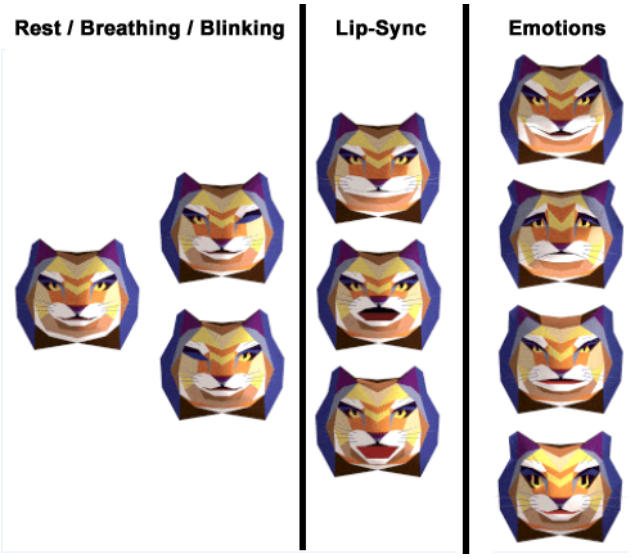

Figure 5. An avatar at a given moment is created from a morph target from each of the 3 categories: Blinking, Lip-sync, Emotions.

The implementation of the facial animation bears some amplification, since it relates to the scope of possibilities for avatar design. The strength of various vowel sounds is analyzed in the Traveler client as part of the signal processing required for speech compression. This information is encoded and bound to the audio stream for that client. When another Traveler client receives this stream, in addition to localizing and attenuating the audio based on the relative position of the corresponding avatar to the local user, it also applies the phoneme information to the 3D model of that avatar. Avatars are designed as a single 3D mesh, to which the avatar designer applies transformations, representing the extreme positions of the mesh in response to various vowel sounds. The avatar is then compiled down to a single neutral mesh and a collection of parameters describing the how varying levels of vowel-sound stimulus will shift the vertices. Thus, if a user were to pronounce "Hee-Haw", the avatar would smoothly transition between a broad flat extension of the lips and cheeks to a jaw-dropping expression, during the course of the two syllables. The avatar designer only specifies a fixed number of morph targets for the neutral mesh (in practice, twelve targets) and the actual position of the vertices will be smoothly interpreted at run-time. Thus, if a designer can produce a set of transforms that convincingly describe the phoneme states for a human face, an animal face, a fantastic creature or even an inanimate object, the avatar will be interpreted as a legitimate character in the social environment.

In addition to phoneme targets, the avatar designer also specifies the state of the neutral mesh in various emotional extremes (happy, sad, angry and surprised). The user can specify their avatar's emotional state by clicking a button on the Traveler client interface. This emotional state is then "added" in to the phoneme calculation for the state of the face, i.e. the user appears happy or sad while speaking. The influence of this emotional state decays at a regular rate over the course of several seconds, so that the transition between emotional states is relatively smooth and so that the user does not forget the current emotional state and appear incongruously happy or sad while verbally expressing some other sentiment. The emotional interface could be said to break the paradigm of managing the social experience only with body language and speech, in that a person does not explicitly choose emotional expressions in normal real-world situations. However this channel of expression was added to provide a dimension that could not be easily derived from speech, position or orientation.

3. DESIGN IMPLICATIONS

3.1 Community: 3D voice with 3D navigation

In designing the Traveler experience, we employed a consistent minimalism that served two primary purposes. The first was to provide a satisfying, responsive experience of an animated world on platforms with limited CPU power and communications bandwidth. The other was to keep the experience focused on a few essential channels of communication that maximized the user's sense of being actually co-located with other real individuals. Our approach was basically a narrative one. We used simple 3D graphical elements to merely suggest various elements of a world and its inhabitants, while at the same time insisting that these inhabitants are "real" by investing them with certain very organic characteristics (voice, fluid motion, emotions, autonomic twitches, etc.) The principle is not unlike the aesthetic employed in traditional animation. Highly stylized people and animals are convincingly portrayed as fully developed characters despite the fact that they bear little actual resemblance to a real-world person. We as viewers seem very ready to accept a character as a person, regardless of how fantastic their appearance as long as they have a recognizable face, are imbued with speech and follow certain familiar patterns of social behavior. By creating an environment that showcases these and other human characteristics, we endeavored to create an experience that was at the same time appropriate to the platform and uncompromising in its portrayal of a virtual place to meet real people.

3.2 Telepresence: Binding the pair

Our basic premise in creating Traveler as a social experience was that humans engage in community primarily with other humans. Thus, if a user is represented in a virtual world by an avatar, another user must perceive that avatar as a real person, if the world is to be useful as a social space. We chose to make extensive use of voice because it is such a rich, multi-layered channel of communication, which conveys a great deal of individual character. We chose to emphasize the face in our avatars because immediacy and intimacy implied by face-to-face communication. The voice belongs to the user, but is fully transferred to the avatar's face through the use of lip-syncing and virtual location. Thus we talk about the "binding the pair", the unification of the remote user and the corresponding avatar in the mind of the local viewer.

Some evidence that suggests a level of success in this binding emerged during early user tests of the system. It was observed that users felt the need to maintain eye contact with the virtual avatars on the screen. They seemed hesitant to turn away from the screen for fear of being perceived as "rude", despite being aware that their turning away could not be perceived by the other users. The suspension of disbelief in using the system was such that unconscious social patterns of behavior were in effect. Similarly, in participating in Traveler sessions, it is clear that certain standards of social behavior are naturally observed in the virtual world. Users describe a sense of discomfort when a novice user navigates too closely and thus violates the normal sense of "personal space". In response, the violated user will navigate backward to a "safe" distance. Users tend to unconsciously turn to orient on the current speaker as one would in the real world and generally organize themselves in social patterns with the virtual space. All of these things suggest a high level of immersion in the illusion of co-location.

3.3 Evolving Issues

As the PC platform increases in power, especially in its ability to render 3D scenes, the minimalist approach employed in implementing Traveler is less relevant from a pragmatic standpoint. However the lessons learned in realizing Traveler inform our use of the increased power available in advancing the platform. For example, adding geometric complexity to the representation of a chair in a business space would do little to improve the suggestion of common workplace affordances. However, using modern skinning techniques in implementing avatar morphing might allow even greater range of expression and thus improve the expression of a user's personality in the virtual world.

Our experience with the Traveler community suggests that we should avoid the temptation to strive for greater levels of photo-realism in presenting the illusion of co-location. While a cinematic quality is appropriate to achieve a sense of immersion for some kinds of entertainment software, a social application requires a careful balance of interface elements that are not always realistic in appearance. Adding full-bodied avatars for example might make the experience seem more realistic. However, keeping such an avatar in full view within a scene would reduce the size of the face to a very small field, diminishing the effect of voice-binding achieved through lip-syncing. Furthermore, since driving the use of the full body would have to be done through some complex mouse and keyboard interface, it would lack the spontaneous and genuine nature of the voice communication. Until the computer can "watch" a user and include the gestural component of their communication into the input stream, the use of the virtual body as a communication device would actually be quite artificial as compared to the use of the voice. Thus ironically the greater "realism" afforded by the more naturalistic avatar would add a completely non-organic element to the social construct. Finally, the use of a stylized interface avoids a problem common to all computer graphics; namely that any attempt at photo-realism is judged on how far short of that goal it falls. On

the other hand, once a user concedes that an interface is stylized, they will accept the suggestions of the stylistic elements and judge the experience on its merits.

This is not to say that the experience of Traveler should stand still in the face of an evolving PC platform. Hardware anti-aliasing and pixel shaders might make the appearance of the interface more visually appealing and therefore engaging. More importantly, greater processing power on the client might make it possible to derive a user's emotional state from the voice stream, obviating the need for the user to choose their current emotional state through the user interface.

3.3.1 Community Building

While Traveler indicates some promising directions for natural social communications over the Internet, it does little to address the wider topic of web-based community. We assert that communications and the strong illusion of co-location are essential to the development of web-based community. We have further attempted to show that voice, as the most organic form of human communication, will lead to the most natural formation of community. However, if the platform does not provide other affordances, communications in itself is not sufficient to allow the formation of community. In the real world, communities do not automatically form out of public places where strangers have occasion to talk to one another. Communities also require shared goals, interests, problems, values and economies as well as conflict over these same issues. In essence, the shared virtual world must be a place worth communicating about, or it must at least stand as a proxy for the real world with substantive issues.

There are a number of directions that can be explored along these lines, using Traveler as a point of departure. Currently the shared virtual environment in which the avatars interact is relegated to being a simple collection of narrative and contextual elements. The objects in the world function as "conversation pieces" or simple points of reference for navigation. One way in which the current community has used this limited resource to create greater cohesion is to use the spaces themselves as architectural/sculptural artworks. Some users have used world-creation as an expressive medium and have taken on the specialized role of artists within the community. This activity seeks to create value and therefore substance out of the actual structure of the shared virtual world. By adding more types of media to those that can currently be integrated into a world (video, high-quality audio, free-form animated geometry, etc.), the scope of this creative effort would be expanded, and the world might attract content-producers from more traditional areas to participate.

In Traveler currently, there is little or no interaction between the user/avatar and the world. If participants could affect the virtual worlds, then they would be empowered to engage in certain collaborative tasks and creative endeavors. These worlds would allow collaborative artwork as well as simple games that have been long proven to attract interested groups (e.g. chess, bridge, etc.). Any form of collaboration is a step toward deeper community. Again there is evidence for this within the existing Traveler community. Despite the limited facilities provided by the platform for world-based collaborative activities, the users spontaneously devised a number of group activities that were unforeseen by the developers. These include avatar races and treasure hunts. The basic ability to co-locate and use voice has been used to stage virtual dramas and to hold church services, book clubs, sing-alongs and karaoke sessions. All of these activities were sponsored and organized spontaneously by the user base, indicating the strong desire for collaboration and structure within the virtual community.

In the current Traveler community, certain users have taken on the roles of organizers, acting essentially like hosts or even "mayors" of collections of worlds. They do this by virtue of operating servers that host the worlds and in some cases by creating worlds in collections that form a thematic group. A server operator has a certain amount of moderation capability in that they can eject users from worlds, but if users had a verifiable identity with the worlds, then certain kinds of privileges could be assigned to them. Thus, a server operator could assign assistants within the community that act as local super-users to direct activities. An identity scheme would also provide for a reputation scheme to be implemented, allowing all the participants of a world to socially manage their environment. Overall, this would create a richer texture of socio-political structure, which is typical of real-world communities. It should be noted however, that by choosing voice as the medium of communication, Traveler has a de facto organic identity scheme in operation by virtue of the fact that a user's voice is highly identifiable. Certain users in the existing community (identified by their voice, since their avatar and profile are changeable) have a reputation as troublesome, eccentric, friendly or influential and are treated accordingly by the body of users as a whole. However, formalized identity tools would allow the platform to provide certain affordances for formalizing this process and thus allow the participants to take control of their social environment and turn practices into policies.

Some users have expressed the desire for the facilities required to implement e-commerce in a Traveler world. An obvious application of this would be a virtual shopping center, with avatar sales-persons facilitating the sale of real-world products. One way to quickly tie Traveler into the vast existing body of technology for e-commerce would be to improve its integration with traditional web browsers. Traveler currently has limited browser interaction capabilities, in that certain objects in the world can act as web-links, causing a browser to launch and display the data at a particular URL. If this were augmented with access to a browser-based scripting language, JavaScript being the obvious choice, Traveler could then delegate much of the e-commerce functionality while still integrating it closely into the world. While the topic of commerce is somewhat orthogonal to that of community, the ability to tie the virtual community in with the real-world economy would be one method of providing an economic dimension to the virtual world. Also, the fact that some participants want to exercise their skills as salesmen within the community indicates a desire on their part to contribute what they see as unique abilities to the overall mix, not unlike those who function as artists and hosts.

The process of exploring the various ways of expanding the existing platform to provide greater depth of community is highly interdisciplinary. It would benefit from the input of sociologists, psychologists, economists, artists and historians and could occupy a great deal of full-time research and engineering. By providing two of the essential

building blocks of community, namely free-form communication and a strong sense of presence, Traveler is a useful platform from which to begin some of the more advanced areas of investigation.

4. ACKNOWLEDGMENTS:

The authors would like to acknowledge Rod MacGregor, Henry Nash, Dave Owens, Ali Ebtekar, Stasia McGehee and James Grunke for their participation in designing and implementing OnLive Traveler. We would also like to thank the long time community members of the Traveler worlds for their support and insight.

5. REFERENCES

[1] DiPaola, Collins, A 3D Natural Emulation Design to Virtual Communities, Siggraph '99, 1999.

[2] Damer, B. Avatars!: Exploring and Building Virtual Worlds on the Internet. Peachpit Press, Berkeley. 1998.

[3] Stephenson, N., Snow Crash, New York NY: Bantam Spectra, 1992.

[4] Heim, M., Virtual Realism by Oxford U. Press, 1998

[5] Rheingold, H. The Virtual Community: Homesteading on the Electronic Frontier. Addison-Wesley, New York. 1993.

[6] Damer, B., S. Gold, J. de Bruin, D-J. de Bruin.. “Steps toward Learning in Virtual World Cyberspace: TheU Virtual University and BOWorld.” In Interactions in Virtual Worlds. : A. Nijholt, O.A. Donk, E.M.A.G. van Dijk (eds.): University Twente, Enschede, 31-43. 19