Autonomous Communicative Behaviors in Avatars

by

Hannes Högni Vilhjálmsson
B.Sc., Computer Science
University of Iceland, Reykjavík (1994)

SUBMITTED TO THE PROGRAM IN MEDIA ARTS AND SCIENCES,
SCHOOL OF ARCHITECTURE AND PLANNING,
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE
in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1997

© Massachusetts Institute of Technology 1997
All Rights Reserved

Signature of Author _____
Program in Media Arts and Sciences
May 9, 1997

Certified by _____
Justine Cassell
AT&T Career Development Assistant Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by _____
Stephen A. Benton
Chairman, Department Committee on Graduate Students
Program in Media Arts and Sciences

Autonomous Communicative Behaviors in Avatars

by

Hannes Högni Vilhjálmsson

Submitted to the Program in Media Arts and Sciences,

School of Architecture and Planning,

on May 9, 1997, in partial fulfillment of the requirements for the degree of

Master of Science

in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Abstract

Most networked virtual communities, such as MUDs (Multi-User Domains), where people meet in a virtual place to socialize and build worlds, have until recently mostly been text-based. However, such environments are now increasingly going graphical, displaying models of colorful locales and the people that inhabit them. When users connect to such a system, they choose a character that will become their graphical representation in the world, termed an *avatar*. Once inside, the users can explore the environment by moving their avatar around. More importantly, the avatars of all other users, currently logged onto the system, can be seen and approached.

Although these systems have now become graphically rich, communication is still mostly based on text messages or digitized speech streams sent between users. That is, the graphics are there simply to provide fancy scenery and indicate the presence of a user at a particular location, while the act of communication is still carried out through a single word-based channel. Face-to-face conversation in reality, however, does make extensive use of the visual channel for interaction management where many subtle and even involuntary cues are read from stance, gaze and gesture. This work argues that the modeling and animation of such fundamental behavior is crucial for the credibility of the virtual interaction and proposes a method to automate the animation of important communicative behavior, deriving from work in context analysis and discourse theory. BodyChat is a prototype of a system that allows users to communicate via text while their avatars automatically animate attention, salutations, turn taking, back-channel feedback and facial expression, as well as simple body functions such as the blinking of the eyes.

Thesis Supervisor: Justine Cassell

Title: AT&T Career Development Assistant Professor of Media Arts and Sciences

Autonomous Communicative Behaviors in Avatars

by

Hannes Högni Vilhjálmsson

The following people served as readers for this thesis:

Reader _____

Bruce M. Blumberg

Assistant Professor of Media Arts and Sciences

MIT Program in Media Arts and Sciences

Reader _____

Carl Malamud

Visiting Scientist

MIT Media Laboratory

Acknowledgements

I would like to thank my advisor, Justine Cassell, for giving me freedom to pursue my interests while making sure that I founded my research on a strong theoretical background. I am grateful for her support and supervision. My thanks to Hiroshi Ishii and Henry Lieberman for the class on Collaboration between People, Computers and Things, where my thesis work started to take shape. I am indebted to my readers Bruce Blumberg and Carl Malamud for guidance on drafts of both my proposal and this document.

Special thanks to the members of the Gesture and Narrative Language Group for their encouragement and great company. In particular I would like to thank my officemate and friend Marina Umaschi for bringing the human factor into the office life and my fellow countryman Kris Thórisson for the humor and long nostalgic conversations over sandwiches at the local coffee franchise.

This work would not have been possible without the sponsorship of a number of companies. I would especially like to mention Intel for donating the hardware and Microsoft and TGS for donating the software.

Thanks to my housemate Brygg Ullmer for making me feel at home.

Thanks to my family for life and love.

Thanks to Deepa for being a good supportive friend, and for showing me that avatars can only go so far…

Table of Contents

1. INTRODUCTION 8

1.1. SCENARIO 8

1.1.1. THE VISION OF SCIENCE-FICTION 8

1.1.2. FICTION TURNED REAL, BUT NOT QUITE 8

1.1.3. MY CONTRIBUTION 9

1.2. APPLICATION DOMAIN 9

1.2.1. VIRTUAL BODIES 9

1.2.2. CHATTING 10

1.2.3. TELECOMMUTING 10

1.2.4. GAMING 11

1.3. OVERVIEW OF THESIS 11

2. CURRENT SYSTEMS AND THEIR SHORTCOMINGS 12

2.1. TYPES OF SYSTEMS 12

2.2. AN EXISTING SYSTEM: ACTIVE WORLDS 12

2.3. SHORTCOMINGS 15

2.3.1. TWO MODES OF OPERATION 15

2.3.2. EXPLICIT SELECTION OF BEHAVIOR 15

2.3.3. EMOTIONAL DISPLAYS 16

2.3.4. USER TRACKING 16

3. PREVIOUS RESEARCH 18

3.1. SOCIAL SCIENCE STUDIES OF EMBODIED COMMUNICATION 18

3.1.1. MULTIMODAL CONVERSATION 18

3.1.2. AN ANALYZED CONVERSATION 19

3.1.3. GAZE AND THE INITIATION OF A CONVERSATION 20

3.1.4. THE FUNCTIONS OF THE FACE DURING A CONVERSATION 21

3.2. COMMUNICATING VIRTUAL BODIES **22**

3.3. ELECTRONIC COMMUNITIES **23**

3.4. MULTI-USER PLATFORMS **24**

4. THE SYSTEM: BODYCHAT **26**

4.1. SYSTEM DESCRIPTION **26**

4.1.1. OVERVIEW 26

4.1.2. A NOVEL APPROACH 27

4.2. SYSTEM ARCHITECTURE **28**

4.2.1. AVATAR CREATION AND DISTRIBUTION 28

4.2.2. AVATAR CONTROL 29

4.3. USER INTENTION **30**

4.4. AVATAR BEHAVIOR **32**

4.4.1. PRESENCE AND MOVEMENT 32

4.4.2. SIGNS OF LIFE 32

4.4.3. COMMUNICATION 32

4.5. SAMPLE INTERACTION **35**

4.5.1. OVERVIEW 35

4.5.2. NO INTEREST 35

4.5.3. PARTNER FOUND 36

4.5.4. A CONVERSATION 38

4.6. IMPLEMENTATION **39**

4.6.1. PROGRAMMING PLATFORM 39

4.6.2. CONSTRAINTS 39

4.6.3. MAJOR CLASSES 39

4.7. PORTABILITY **39**

5. CONCLUSION **41**

5.1. SUMMARY **41**

5.2. EVALUATION **42**

5.3. FUTURE DIRECTIONS **43**

5.3.1. EXPANSION IN TWO AREAS 43

5.3.2. AVATAR BEHAVIOR 43

5.3.3. USER INPUT 43

APPENDIX A: USER INTERFACE **45**

APPENDIX B: REACTION IMPLEMENTATION **46**

APPENDIX C: WORD ACCOMPANIMENT **47**

REFERENCES **48**

1. Introduction

1.1. Scenario

1.1.1. The vision of science-fiction

In the novel *Neuromancer,* Science-fiction writer William Gibson let his imagination run wild, envisioning the global computer network being an immersive space, much like a parallel dimension, into which people could jack via neural implants (Gibson 1984). This was a shared graphical space, not constrained by the laws of a physical reality, allowing people to interact with remote programs, objects and other people as if they were locally present. This novel stirred many minds and is frequently referred to as the origin of the term *Cyberspace*.

Another influential science-fiction novel is *Snowcrash*, written by Neal Stephenson, where a near-future scenario describes the Metaverse, a computer generated universe in which people can go about their digital business clad in 3D graphical bodies, termed *avatars* (Stephenson 1992). (The word *avatar* comes from Sanskrit and means incarnation).

1.1.2. Fiction turned real, but not quite

In 1985, Lucasfilm created Habitat, an online service in which each user was represented as an avatar that could be moved around a common graphical space using the keys on a keyboard (see Figure 1). Users could manipulate the environment as if they were playing a computer game, or they could interact with other users through text messages displayed along with the figures.

Figure 1: In Habitat users are shown as graphical figures.

Now, as the Internet embraces sophisticated graphics, dozens of similar Internet-based systems have emerged. Some are 2D in nature, like Habitat, others plunge into the third dimension, partly fueled by the VRML standardization of 3D graphics interchange on the Internet. While still not using Gibson's neural implants or Stephenson's goggles, these

environments provide windows into a shared visually rich universe in which you can see remote users float around. However, when you step up to an avatar to start a conversation, the spell is broken because current avatars don't exploit embodiment in the discourse. At best they move their lips while a user is speaking, but things like shifting the gaze or gesture with the hands are absent or totally irrelevant to the conversation.

1.1.3. My contribution

I use a model derived from work in discourse theory, dealing with multiple modes of communication, to animate communicative visual behavior in avatars. I have built a working prototype of a multi-user system, BodyChat, in which users are represented by cartoon like 3D animated figures. Interaction between users is allowed through a standard text chat interface. The new contribution is that visual communicative signals carried by gaze and facial expression are automatically animated as well as body functions such as breathing and blinking. The animation is based on parameters that reflect the intention of the user in control as well as the text messages that are passed between users. For instance, when you approach an avatar, you will see from its gaze behavior whether you are invited to start a conversation, and while you speak your avatar will take care of animating its face and to some extent the body. In particular it animates functions such as salutations, turn-taking behavior and back channel feedback.

1.2. Application Domain

1.2.1. Virtual Bodies

This work introduces an approach to animating virtual bodies that represent communicating people. The following sections present three different types of applications where the avatar technology presented here could be employed to enhance the experience. The existence and popularity of these applications serves as a motivation for the current work.

1.2.2. Chatting

Pavel Curtis, one of the creators of LambdaMOO (Curtis 1992), advocates that the Internet "killer app of the 90's" is *people*. His point is that whatever business we go about on the global network, we shouldn't have to be alone, unless we want to. You should be able to see and communicate with people strolling the isles of a supermarket, hanging out in the café or waiting in lines, be it in an old fashioned mall or an on-line shopping center. A new era in technology is upon us: the age of social computing (Braham and Comerford 1997).

Systems that allowed a user to see who was on-line and then enabling them to exchange typed messages in real-time date back to the first time-sharing computers of the 1960s (Rheingold 1994) . Later systems, such as the Internet Relay Chat (IRC), have been widely popular as a way to convene informal discussions among geographically distant people, "but the continuing popularity of IRC appears to be primarily a function of its appeal as a psychological and intellectual playground" (Rheingold 1994, 179). The IRC and more recently, various Distributed Virtual Environments (DVEs), seem to serve a purpose as public meeting places analogous to their real world counterparts, but not confined to physical distances.

1.2.3. Telecommuting

In today's global village where multi-national companies keep growing and research institutions in different countries join forces to address major issues, the demand for efficient channels of communication across long distances has never been greater. The field of Computer Supported Collaborative Work (CSCW) is exploring ways to create systems and techniques that help distributed workgroups to jointly perform a task and share experience.

One aspect of such a system is real-time communication between the members of the group in the form of a virtual meeting. There it is important to incorporate some mechanisms to assist in managing the flow of turns to avoid a chaotic situation. For dealing with this and other issues of mediating presence, representing participants visually is a powerful approach. Consider a large meeting where half of the participants are physically present in the room but the other half is participating through a speakerphone. The remote people are soon

dominated by the others, and often reduced to mere overhearers (according to personal communication with various sponsors).

1.2.4. Gaming

Computer gaming has until recently been mostly a solitary experience, but with the sudden surge in household Internet connectivity the global network is fast becoming a sprawling playground for all kinds of game activity. Text based games and role-playing environments have been on-line for awhile, such as the popular MUD (Multi-User Dungeon) that has been around for almost two decades. But now a wide selection of simulations, war games, action games, classic games as well as different versions of role-playing games offer a graphically rich environment in which you can interact with other game players across continents. Although many of those games pose players head-to-head in combat, others encourage group co-operation and interaction. These games already provide captivating virtual worlds to inhabit and they often represent users as avatars, adapted for the environment and the particular game experience.

1.3. Overview of Thesis

The previous chapter has served as an introduction to the domain of this work, and motivated the subject by presenting some applications. The remainder of the thesis is divided into four chapters that in a general sense present, in order, the problem, the theoretical tools for working on the problem, how this work applies the tools, and conclusions. Chapter 2 starts by describing in detail an already existing system and then goes on to discuss the shortcomings of current systems with regard to avatars. Chapter 3 is a review of relevant work from various research areas related to and supporting this work, establishing a solid foundation. Chapter 4 discusses the working prototype, how it starts to address the stated problems, its system architecture and how it is theoretically rooted. Finally Chapter 5 gives a summary, evaluation and suggests future directions.

2. Current systems and their shortcomings

2.1. Types of systems

The term avatar has been used when referring to many different ways of representing users graphically. As described in section 1.2, the range of applications is broad and the requirements for the user's virtual presence differ. This work implements the type of avatars that inhabit what has been technically referred to as Distributed Virtual Environments (DVEs). The ideas presented here are still applicable to other kinds of systems and should be viewed with that in mind. The next section describes an existing graphical chat system that is a good example of a DVE. The particular system was chosen because it is popular and sports a few advanced features. The system is described here in detail primarily to give readers of this work some insight into the current state of the art.

2.2. An existing system: Active Worlds

The Active Worlds Browser (AWB) from Worlds Incorporated is a client program running under Windows that connects the user to an Active World server maintained by Worlds Inc. or one of its collaborators. The client renders a view into the Active World as seen by the avatar or optionally a floating camera (see Figure 2). Other users are seen as articulated 3D models that they have chosen from a menu of available bodies. The user can freely move through the 3D scene using either a mouse or the cursor keys. To communicate, the user types a sentence into an edit field, transmitting it into the world by hitting Carriage Return. A scrollable text window directly below the rendered view displays all transmitted sentences along with the name of the responsible user. The sentence also appears floating above the head of the user's avatar. Only sentences from the closest 12 users are displayed.

Figure 2: The Active Worlds Browser is an example of a Distributed Virtual Environment (DVE).

Before using AWB one must choose a nickname and acquire a unique user ID number from Worlds Inc.'s "immigration" service. The nickname and ID are written into the AWB configuration file ensuring consistent identity between sessions. After executing the browser, the user can select from a list of avatar models to represent them in the world. The user is free to switch to another model at any time. The models are human figures of both sexes and various ethnicities. Each body has a set of distinctive idle motion sequences that are executed at random for an interesting visual effect. Some avatars seem to be checking their watches once in awhile, others rock their hips or look pensive.

Once moving through the world, the user is allowed to switch between a view of the surroundings through the eyes of the avatar and an overhead view following the avatar around. This allows the user to look at other users face-to-face or to observe themselves along with the other users. When the user wants to initiate a contact with another person, three steps can be taken. First the user can navigate up to another avatar, making sure to enter the other person's field of view. Then the user can select from a limited set of animation sequences for the avatar to play, 'Waving' being the most appropriate for this situation. Lastly, the user starts a conversation by transmitting a sentence into the space, preferably addressing the person to contact. In fact, only the last step is necessary; the user's greeting sentence will be 'heard' by the 12 closest avatars, regardless of their location or orientation. During the conversation, the user keeps typing messages for transmission, switching between animations

from a set of ‘Happy’, ‘Angry’ and ‘Wave’ as appropriate. Between the selected animation sequences, the idle motions are randomly executed.

Upon entry into an Active World using the AWB, one notices how lively and in fact life-like the world seems to be. A crowd of people gathered on the city square is crawling as avatars move about and stretch their bodies. However, one soon realizes that the animation displayed is not reflecting the actual events and conversations taking place, as transcribed by the scrolling text window beneath the world view.

Although the avatars allow the user to visually create formations by controlling position and orientation in relation to other avatars, this does not affect the user’s ability to communicate as long as the desired audience is among the 12 closest persons. One reason for this redundancy is that the bodies in this system are not conveying any conversational signals. The automated motion sequences are not linked to the state of the conversation or the contents of the messages, but are initiated at random, making them irrelevant. The manually executed motion sequences allow a few explicit (and somewhat exaggerated) emotional displays, but since they are chosen by the user via buttons on a control panel, they tend not to be used while the user is engaged in a conversation, typing away on the keyboard.

2.3. Shortcomings

> Paul walks up to Susan who stands there staring blankly out into space. *"Hello Susan, how are you?"* Susan looks at her watch as she replies *"Paul! Great to see you! I'm fine, how have you been?"* Paul returns the stare and without twitching a limb he exclaims *"Real Life sucks, I don't think I'm going back there :) "*. Susan looks at her watch. Paul continues *"I mean, out there you can't just walk up to a random person and start a conversation"*. Susan looks at her watch. Karen says *"Hi"*. While Paul rotates a full circle looking for Karen, Susan replies *"I know what you mean"*. Karen says *"So what do you guys think about this place?"*. Karen is over by the fountain, waving. Susan looks blankly at Paul as she says *"I think it is great to actually see the people you are talking to!"*. Paul is stiff. Karen is waving. Susan looks at her watch.

2.3.1. Two modes of operation

In most current systems (such as the popular Active Worlds and The Palace) the user has to switch between controlling the avatar and chatting with other users. While the user is creating the message for her interlocutor, her avatar stands motionless or keeps repeating a selected animation sequence. This fails to reflect the relationship between the body and the communication that is taking place, potentially giving misleading or even conflicting visual cues to other users. Some systems, such as the voice based OnLive world, offer simple lip synching, which greatly enhances the experience, but actions such as gaze and gesture have not been incorporated.

2.3.2. Explicit selection of behavior

The creators of multi-user environments realize that avatars need to be animated in order to bring them to life, but their approach does not take into account the number of different communicative functions of the body during an encounter. They provide menus where users can select from a set of animation sequences or switch between different emotional representations. The biggest problem with this approach is that every change in the avatar's state is explicitly controlled by the user, whereas many of the visual cues important to the conversation, are spontaneous and even involuntary, making it impossible for the user to explicitly select them from a menu. Furthermore, the users are often busy producing the content of their conversation, so that simultaneous behavior control becomes a burden.

2.3.3. Emotional displays

When people looked at the stiff early versions of avatars and considered ways to make them more life-like, generally they came to the conclusion that they were lacking *emotions*. Users should be allowed to express emotions in order to liven up the interaction. Naturally we associate the display of emotions to being human and the way we relate to our environment and other people. As repeatedly emphasized throughout a book on Disney animation, written by professional animators (Thomas and Johnson 1981), rich and appropriate emotional display is essential for the illusion of life.

However, lively emotional expression in interaction is in vain if mechanisms for establishing and maintaining mutual focus and attention are not in place (Thorisson and Cassell 1996). A sole person standing on a street corner, staring fixedly at a nearby wall and sporting a broad smile will be lucky if people other than suspicious officers dare to approach. We tend to take communicative behaviors such as gaze and head movements for granted, as their spontaneous nature and non-voluntary fluid execution makes them easy to overlook when recalling a previous encounter (Cassell, forthcoming). This is a serious oversight when creating avatars or humanoid agents since emotion displays do not account for the majority of displays that occur in a human to human interaction (Chovil 1992).

2.3.4. User tracking

Many believe that employing trackers to map certain key parts of the user's body or face onto the graphical representation will solve the problem of having to explicitly control the avatar's every move. As the user moves, the avatar imitates the motion. This approach, when used in a non-immersive setting, shares a classical problem with video conferencing: The user's body resides in a space that is radically different from that of the avatar. This flaw becomes particularly apparent when multiple users try to interact, because the gaze pattern and orientation information gathered from a user looking at a monitor doesn't map appropriately onto an avatar standing in a group of other avatars. Thus whereas tracking may be appropriate for Virtual Reality applications where head mounted displays are employed, it does not lend itself well to Desktop Virtual Environments.

3. Previous Research

3.1. Social science studies of embodied communication

3.1.1. Multimodal conversation

A face-to-face conversation is an activity in which we participate in a relatively effortless manner, and where synchronization between participants seems to occur naturally. This is facilitated by the number of channels we have at our disposal to convey information to our partners. These channels include the words spoken, intonation of the speech, hand gestures, facial expression, body posture, orientation and eye gaze. For example, when giving feedback one can avoid overlapping a partner by giving it over a secondary channel, such as by facial expression, while receiving information over the speech channel (Argyle and Cook 1976). The channels can also work together, supplementing or complementing each other by emphasizing salient points (Chovil 1992, Prevost 1996), directing the listener's attention (Goodwin 1986) or providing additional information or elaboration (McNeill 1992, Cassell forthcoming). When multiple channels are employed in a conversation, we refer to it as being multimodal.

We can think about the process as being similar to collaborative weaving. Each person contributes a bundle of different colored threads, the color representing a modality of communication, such as speech or gesture. Over the course of the conversation, the group of people weaves a continuous and seamless textile where each band consists of multiple strings from different participants and different modalities. When looking at the finished tapestry, an emerging pattern may be observed, suggesting an ordered affair. However, unlike the skilled textile worker, the people involved in the conversation will not be able to recall the specifics of laying out the strings, since most of it happened spontaneously.

Of course the pattern observed will be unique for each encounter, given a unique situation and cast of characters, but the relationship between the different colors and bands, is to some extent governed by general principles (Kendon 1990). Researchers from different disciplines, such as linguistics and sociology, have conducted the search for these principles of multimodal communication, each from a different point of view.

Even though methods differ and approaches to explanation vary, it is made clear that our body, be it through gesturing or our facial expression, displays structured signals that are an integral part of communication with other people. These are behaviors that should be exploited in the design of autonomous and semi-autonomous characters that are intended to be a part of or assist in a natural dialog.

The current work focuses on gaze and communicative facial expression mainly because these are fundamental in establishing and maintaining a live link between participants in a conversation. The displaying of gesture and body posture is also very important, but the required elaborate articulation of a human body is beyond the scope of this thesis and will be pursued later.

To illustrate what is meant by communicative behavior, the following section describes a scenario where two unacquainted people meet and have a conversation. The behaviors employed are referenced to background studies with relevant page numbers included. The two subsequent sections then elaborate on some of these behaviors and serve as a theoretical foundation for the automated behaviors in BodyChat.

3.1.2. An analyzed conversation

Paul is standing by himself, looking out for interesting people. Susan (unaquainted to Paul) walks by, mutual glances are exchanged, Paul nods smiling, Susan looks at Paul and smiles [distance salutation] (Kendon 1990, 173. Cary 1978, 269) Susan touches the hem of her shirt [grooming] as she dips her head, ceases to smile and approaches Paul (Kendon 1990, 186, 177). She looks back up at Paul when she is within 10' [for initiating a close salutation], meeting his gaze, smiling again (Kendon 1990, 188; Argyle 1976, 113). Paul tilts his head to the side slightly and says "Paul", as he offers Susan his hand, which she shakes lightly while facing him and replying "Susan" [close salutation] (Kendon 1990, 188, 193). Then she steps a little to the side to face Paul at an angle (Kendon 1990, 193; Argyle 1976, 101). A conversation starts.

During the conversation both Paul and Susan display appropriate gaze behavior, such as looking away when starting a long utterance (Kendon 1990, 63; Argyle 1976, 115; Chovil 1992, 177; Torres 1997), marking various syntactic events in their speech with appropriate

facial expressions, such as raising their eyebrows while reciting a question or nodding and raising eyebrows on an emphasized word (Argyle 1973; Chovil 1992, 177; Cassell 1994), giving feedback while listening in the form of nods, low "mhm"s and eyebrow action (Chovil 1992, 187; Schegloff 1968; Cassell 1994) and finally giving the floor or selecting the next speaker using gaze (Kendon 1990, 85; Chovil 1992, 177; Argyle 1973; Argyle 1976, 118).

3.1.3. Gaze and the initiation of a conversation

The eyes are a powerful channel for intimate connection between people. Not only does the "look" suggest a being with consciousness and intentions of its own, as Sartre (Sartre 1956) describes it, but it also works as a device for people to commonly establish their "openness" to one another's communication (Kendon 1990, Argyle 1976, Goffman 1963).

Merely meeting a person's gaze is an important first step but will not initiate a conversation. In fact what E. Goffman refers to as "civil inattention" is a fundamental social behavior of unacquainted people that happen to come into each other's proximity without any intentions to converse:

> One gives to another enough visual notice to demonstrate that one appreciates that the other is present (and that one admits openly to having seen him), while at the next moment withdrawing one's attention from him so as to express that he does not constitute a target of special curiosity or design.
>
> (Goffman 1963, 84)

If your initial glance and orientation towards the other person was not met by interest in a greeting, your behavior can pass as a part of the "civil inattention" ritual and thus you are saved the embarrassment of explicitly requesting a conversation from an unwilling person (Goffman 1963, Cary 1978, Kendon 1990).

The showing of mutual awareness asserts that the other person's subsequent actions take your approach into account. A second glance or a sustained gaze and a smile, act as indicators of the other person's intentions to greet you. A distance salutation is performed, an approach follows and finally a close salutation occurs once a comfortable conversational distance is established. A few studies have focused on the verbal aspect of opening a

conversation (Schegloff 1968, Schegloff and Sacks 1973), while others have specifically looked at gaze (Kendon 1990, Cary 1975), and Adam Kendon (Kendon 1990) has done a thorough study of the role of the body in a salutation sequence.

3.1.4. The functions of the face during a conversation

Michael Argyle (Argyle and Cook 1976) argues that gaze serves 3 main functions during a face-to-face conversation:

1. Information Seeking
2. Sending signals that accompany the speech
3. Controlling the flow of the conversation

Perhaps the most obvious function of gaze is Information Seeking, since the primary function of the eyes is to gather sensory input. In order to read visual signals from our environment, we have to direct our attention and thus our gaze towards the source. In a face-to-face conversation we rely on various kinds of gestural information given by our partner and therefore we have to glance at them, at least from time to time. Listeners spend more than half of the time looking at the speaker, supplementing the auditory information. Speakers on the other hand spend much less time looking at the listener, partially because they need to attend to planning and don't want to load their senses while doing so (Argyle and Cook 1976). The speaker will at least look at the listener when feedback is expected, such as at the end of utterances, after speech repairs or a word search and during questions (Argyle and Cook 1976, Kendon 1990).

Facial movement, including the gaze, eyebrow action and mouth movement, accompanies the speech and is synchronized at the verbal level. These signals sent during the course of speaking have been classified into syntactic displays and semantic displays (Chovil 1992). The syntactic displays include the raising of eyebrows and a slight head nod on a stressed or an accented word, raised eyebrows during an offer or a suggestion and blinking on a pause. The semantic displays convey something about what is being said. They either emphasize a word by showing an appropriate expression or a reference to an emotion (lowering eyebrows and wrinkle nose when saying "not") or stand in place of a word by acting out what is being

meant (showing surprise by dropping the jaw after saying “when I opened the door, I was like”). Facial movements such as nodding and brow raising are also used as listener feedback sometimes accompanying a low verbal chant like “mhm” or a “yeah”.

Finally the face serves an important function in organizing how the conversation flows between participants. This is of course related to the speaker’s use of gaze to gather information on feedback, since it also signals the listener in question to elicit what is expected. It has been observed that the person whom the speaker last looked at before ending is more likely than other members of the group to speak next (Kendon 1990, Argyle 1976); thus, looking can serve “to coordinate group action by controlling the succession of speeches” (Weisbrod 1965).

3.2. Communicating virtual bodies

The real-time animation of 3D humanoid figures in a lifelike manner is a large research issue. The Improv system (Perlin and Goldberg 1996) demonstrates a visually appealing humanoid animation and provides tools for scripting complex behaviors, ideal for agents as well as avatars. However, coming up with the appropriate communicative behaviors and synchronizing them with an actual conversation between users has not been addressed yet in Improv. Real-time external control of animated autonomous actors has called for methods to direct animated behavior on a number of different levels (Blumberg and Galyean 1995).

Creating fully autonomous agents capable of natural multi-modal interaction deals with integrating speech, gesture and facial expression. By applying knowledge from discourse analysis and studies of social cognition, systems like The Animated Conversation (Cassell et al. 1994b) and Gandalf (Thorisson 1996) have been developed. The Animated Conversation renders a graphical representation of two autonomous agents having a conversation. The system’s dialog planner generates the conversation and its accompanying communicative signals, based on the agent’s initial goals and knowledge. Gandalf is an autonomous agent that can have a conversation with a user and employs a range of communicative behaviors that help to manage the conversational flow. Both these systems are good examples of discourse theory applied to computational environments, but neither is concerned with user embodiment and issues of avatar control.

Studies of human communicative behavior have seldom been considered in the design of believable avatars. Significant work includes Judith Donath's Collaboration-at-a-Glance (Donath 1995), where on-screen participant's gaze direction changes to display their attention, and Microsoft's Comic Chat (Kurlander et al. 1996), where illustrative comic-style images are automatically generated from the interaction. In Collaboration-at-a-Glance the users lack a body and the system only implements a few functions of the head. In Comic Chat, the conversation is broken into discrete still frames, excluding possibilities for things like real-time backchannel feedback and subtle gaze.

3.3. Electronic communities

To understand the importance of addressing the issue of communicative behavior in avatars, it is enlightening to examine the literature on electronic communities. The phenomenon of electronic communities where people gather to socialize without bringing their own physical bodies, has fascinated researchers in sociology, anthropology, ethnography and psychology. In particular the text-based MUDs (Multi-User Domains) have been the subject of a variety of studies, due to their popularity and their strong sense of community construction (Curtis 1992). MUDs have been used to build research communities (Bruckman and Resnick 1995) and learning environments (Bruckman 1997) as well as game worlds and chat rooms. A certain conversational style has emerged in these systems, where a body is simulated in the text messages passed between users (Cherney 1995), emphasizing how the body is intimately involved in the discourse even in the absence of graphics. While some argue that purely text-based MUDs allow for a richer experience than graphical environments by engaging the user's imagination, graphic MUD-like systems are gaining popularity partly because of their familiar video game like interface. Graphical electronic communities introduce the whole new field of avatar psychology, the study of how people present themselves graphically to others (Suler 1996). A recent dissertation at the Media Lab explores in depth various aspects of on-line societies and compares different conversational interfaces (Donath 1997).

3.4. Multi-user platforms

Implementing multi-user environments is a complex research topic first seriously tackled by the military in the large scale SIMNET system developed by DARPA in the mid 80's. The goal was to create a virtual battlefield where multiple manned vehicle simulators could be present in the same environment. A scaled down version, dubbed NPSNET, was developed at the Naval Postgraduate School in Monterey, California, and has spawned many interesting research projects in the field of distributed simulation (Falby et al. 1993; Macedonia et al. 1994; O'Byrne 1995; Waldorp 1995). Other large multi-user environment projects, not necessarily affiliated with the military, include DIVE at SICS, Sweden (Carlsson and Hagsand 1993), SPLINE at MERL (Anderson et al. 1996), MASSIVE at CRG Nottingham University, UK (Greenhalgh and Benford 1995) and the GreenSpace project at the HITLab (Mandeville et al. 1995). These projects have mostly contributed to the development of a reliable infrastructure, but are now increasingly touching on issues concerned with human interaction within the systems. Because of the technical focus, none of them however, have applied discourse theory to the problem.

Commercially, many companies provide low-end client software to connect Internet users to graphical multi-user environments. The first Internet based graphical chat system that incorporated 3D graphics was WorldChat from Worlds Inc. The first system to allow voice communication and implement lip-synching is OnLive! Traveler from OnLive! Technologies and the first to include a selection of motion-captured animation for avatars was OZ Virtual from OZ Interactive. So far most solutions have been proprietary, but are starting to converge with the developing Virtual Reality Modeling Language (VRML), a standard language for describing interactive 3-D objects and worlds delivered across the Internet. Standardizing VRML extensions dealing with avatars and multi-user issues are currently being worked on.

4. The System: BodyChat

4.1. System description

> Paul is standing by himself on the sidewalk, looking about. Susan walks by on the other side of the street, mutual glances are exchanged as they sight each other, and Paul tosses his head smiling and calls *"Susan!"* Susan lowers her head smiling while replying *"Paul!"* emphasized by raised eyebrows. Susan straightens a fold in her jacket as she glances to the side, and approaches Paul across the street. She looks back at Paul when she steps up to him, meeting his gaze, smiling broadly again. Paul tilts his head slightly, opening a palm towards Susan and says *"Susan, how are you?"* Susan's face lights up as she exclaims *"Paul! Great to see you!"*

4.1.1. Overview

BodyChat is a system prototype that demonstrates the automation of communicative behaviors in avatars. Currently BodyChat only implements appropriate behavior for conversations involving no more than two users at a time (see section 5.3). However, this is an actual Distributed Virtual Environment that allows multiple users to share the space, potentially creating a number of conversations running in parallel.

The system consists of a Client program and a Server program. Each Client is responsible for rendering a single user's view into the DVE (see Appendix A). When a Client is run, the user is asked for the host name of a Server. All users connected to the same Server will be able to see each other's avatars as a 3D model representing the upper body of a cartoon-like humanoid character. Users can navigate their avatar around using the cursor keys, give command parameters to their avatar with the mouse and interact textually with other users through a two-way chat window. A sentence entered into the chat window will also be displayed word by word above the user's avatar, allowing the avatar to synchronize facial expression with the words spoken. A camera angle can be chosen to be from a first person perspective (from the eyes of the avatar), from a point just behind the avatar's shoulder or from a distance, encapsulating all participating users.

4.1.2. A novel approach

The novel approach to avatar design presented here, treats the avatar somewhat as an autonomous agent acting on its own accord in a world inhabited by other similar avatars. However the autonomy is limited to animating a range of communicative expressions of the face, leaving the user in direct control of movement and speech content. The avatar continuously tries to show appropriate behavior based on the current situation and modified by the user's intentions, described as a set of parameters toggled by the user. One can think of this as control at a higher level than in current avatar based systems. This approach starts to addresses the following problems:

- **Control complexity:** The user manipulates a few high-level parameters, representing the user's current intention, instead of micromanaging every aspect of animating a human figure.

- **Spontaneous reaction:** The avatar can show spontaneous and involuntary reactions towards other avatars, something that a user would not otherwise initiate explicitly.

- **Discrete user input:** By having the avatar update itself, carry out appropriate behaviors and synchronize itself to the environment, the gap between meaningful samples of user input or lag times is bridged to produce seamless animation.

- **Mapping from user space into Cyberspace:** The user and the user's avatar reside in two drastically different environments. Direct mapping of actions, such as projecting a live image of the user on the avatar's face, will not produce appropriate avatar actions (consider a user in front of a monitor and an avatar in a group of 5 other avatars giving the floor). Control at an intentional level may however allow the avatar to give the cues that are appropriate for the virtual situation.

4.2. System architecture

4.2.1. Avatar creation and distribution

The Server acts as a simple router between the Clients. When a message is sent to the Server, it gets routed to all other connected Clients. The Server gives each Client a unique ID number when they connect. If a Client receives a message from another Client it has not heard from before, it will assume this is a new user, and will spawn a new avatar representing that user. It then sends an update message of its own, to elicit the same creation procedure in the new Client (see Figure 3). The term *Shadow Avatar* is used here to refer to the avatars in a Client that represent other users, in contrast with the one avatar that represents the user of that particular Client (a users can elect to see their own representation by selecting the appropriate camera angle).

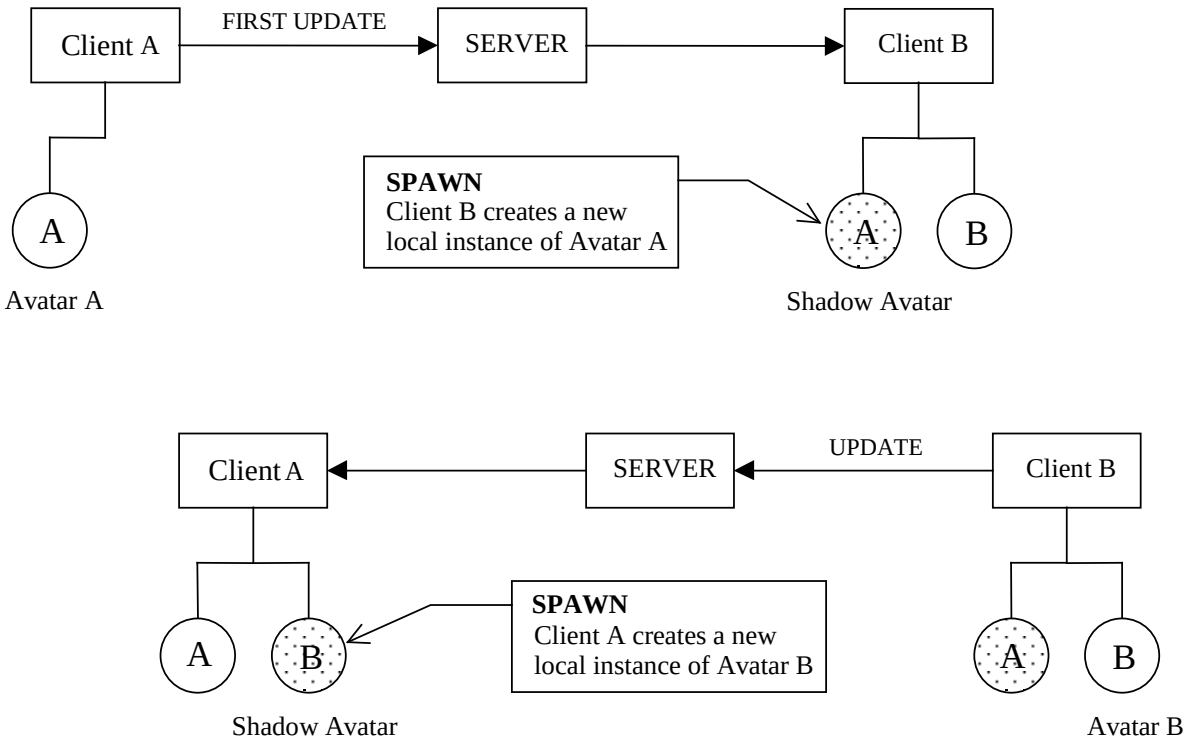


Figure 3: When Client A connects to the Server, other Clients spawn local instances of A's avatar. Client A in turn spawns local instances of all other avatars present.

4.2.2. Avatar control

As stated earlier, the avatar can be thought of as a partially autonomous entity. This entity will live parallel lives in different Clients. The avatar's automated facial expression and gaze will depend on (a) the user's current intentions, as indicated by parameters set by the user, (b) the current state and location of other avatars, (c) its own previous state and (d) some random tuning to create diversity. All user direction of an avatar is shared with all Clients, including the setting of control parameters. This ensures that all instances of an avatar are behaving similarly, although network lag and the produced randomness factor may vary the details.

A user's Client distributes three types of update messages, plus a closing message. These messages act as a remote control for the corresponding avatar instances at other Clients. The messages are listed in Table 1.

Message	**Description**
MOTION	Changes in position and orientation caused by the user's manipulation of cursor keys
CHAT	Strings typed by the user as the content of a conversation
SETTINGS	Control Parameters that describe the user's intentions

Table 1: Avatar update messages broadcast from a user's Client to all connected Clients

Sending a few discrete control settings and then fleshing out the behavior locally in each client instead of having a master instance of the avatar broadcast its behavior to all of its other instances for them to replicate, saves a lot of network traffic and lends itself well to scaling (see Figure 4). With regard to lag times on a network, it is also important to note that each instance is responsible for bringing together the different modalities and producing output that is synchronized within each Client.

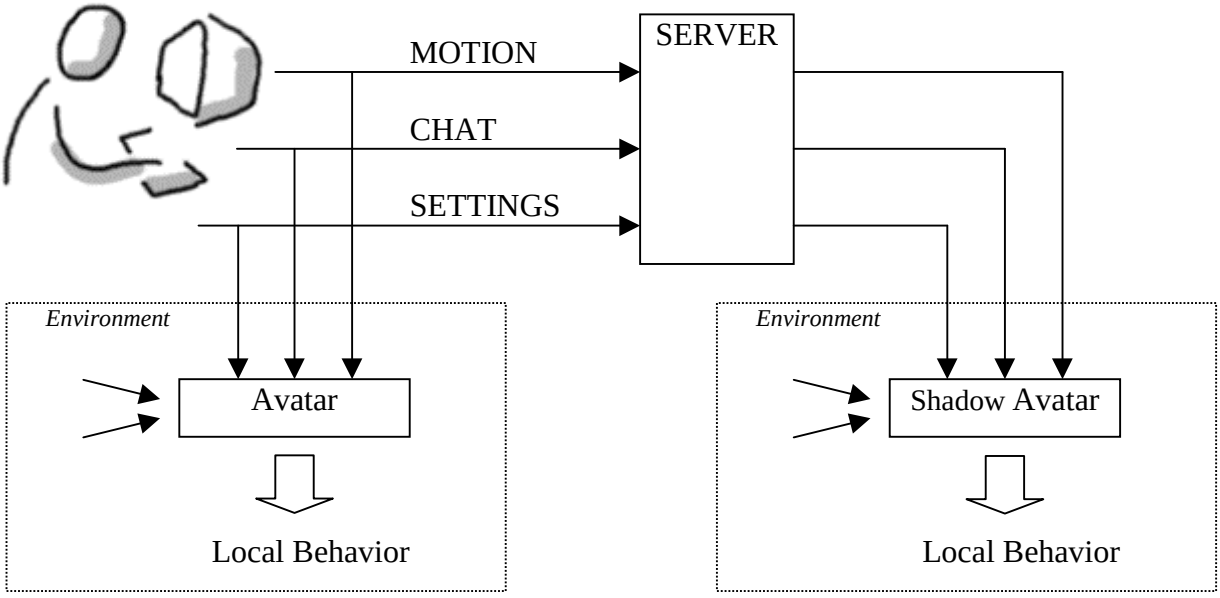


Figure 4: Only a few control settings are sent to the avatar and its remote instances. Each instance then fleshes out the behavior locally.

4.3. User intention

The avatar's communicative behavior reflects its user's current intentions. The user's intentions are described as a set of control parameters that are sent from the user's Client to all connected Clients, where they are used to produce the appropriate behavior in the user's Shadow avatars. BodyChat implements three control parameters as described in Table 2.

Parameter	Type	Description
Potential Conversational Partner	Avatar ID	A person the user wants to chat with
Availability	Boolean	Shows if the user is available for chatting
Breaking Away	Boolean	Shows if the user wants to stop chatting

Table 2: Control Parameters that reflect the user's intention

The ***Potential Conversational Partner*** indicates whom the user is interested in having a conversation with. The user chooses a Potential Conversational Partner by clicking on another avatar visible in the view window. This animates a visual cue to the chosen Avatar that in turn reacts according to that user's *Availability*.

Availability indicates whether the user wants to welcome other people that show interest in having a conversation. This has effect on the initial exchange of glances and whether salutations are performed that confirm the newcomer as a conversational partner. Changing Availability has no effects on a conversation that is already taking place. The user switches Availability ON or OFF through a toggle switch on the control panel (see Appendix A).

During a conversation, a user can indicate willingness to ***Break Away***. The user informs the system of his or her intention to Break Away by placing a special symbol (a forward slash) into a chat string. This is elicits the appropriate diverted gaze, giving the partner a visual cue along with the words spoken. For example, when ready to leave Paul types "/well, I have to go back to work". The partner will then see Paul's avatar glance around while displaying the words (without the slash). If the partner replies with a Break Away sentence, the conversation is broken with a mutual farewell. If the partner replies with a normal sentence, the Break Away is cancelled and the conversation continues. Only when both partners produce subsequent Break Away sentences, is the conversation broken (Kendon 1990, Schegloff and Sacks 1973).

4.4. Avatar behavior

4.4.1. Presence and movement

One of the primary functions of avatars is to indicate a particular user's presence in the virtual environment and pinpoint his or her location. In BodyChat a new avatar is dynamically created in the environment when the user logs on and removed when a user logs off. For moving around, the system directly translates each press of the forward/backward arrows on the keyboard to a forward/backward shift of the avatar by a fixed increment. Press of the left/right keys is translated to a left/right rotation of the avatar body by a fixed increment. When using either a first person perspective camera or a shoulder view, the viewpoint is moved along with the avatar. The shadow avatars precisely replicate the movement of the primary avatar.

4.4.2. Signs of life

Breathing and eye blinks are automated by the avatar throughout the session, without the user's intervention. Breathing is shown as the raising of the shoulders and chest. Blinking fully covers the eyes for a brief moment. Some randomness is introduced to prevent mechanical synchrony. The shadow avatars execute this behavior independently from the primary avatar.

4.4.3. Communication

When discussing the communicative signals, it is essential to make clear the distinction between the *Conversational Phenomena* on one hand and the *Communicative Behaviors* on the other. Conversational Phenomena describe an internal state of the user (or avatar), referring to various conversational events. For example, a *Salutation* is a Conversational Phenomenon. Each Phenomenon then has associated with it a set of *Communicative Behaviors*, revealing the state to other people. For example, the Salutation phenomenon is associated with the *Looking*, *Head Tossing*, *Waving* and *Smiling* Behaviors.

The avatars in BodyChat react to an event by selecting the appropriate Conversational Phenomenon that describes the new state, initiating the execution of associated Communicative Behaviors. Essentially the avatar's behavior control consists of four tiers, where the flow of execution is from top to bottom (see Figure 5).

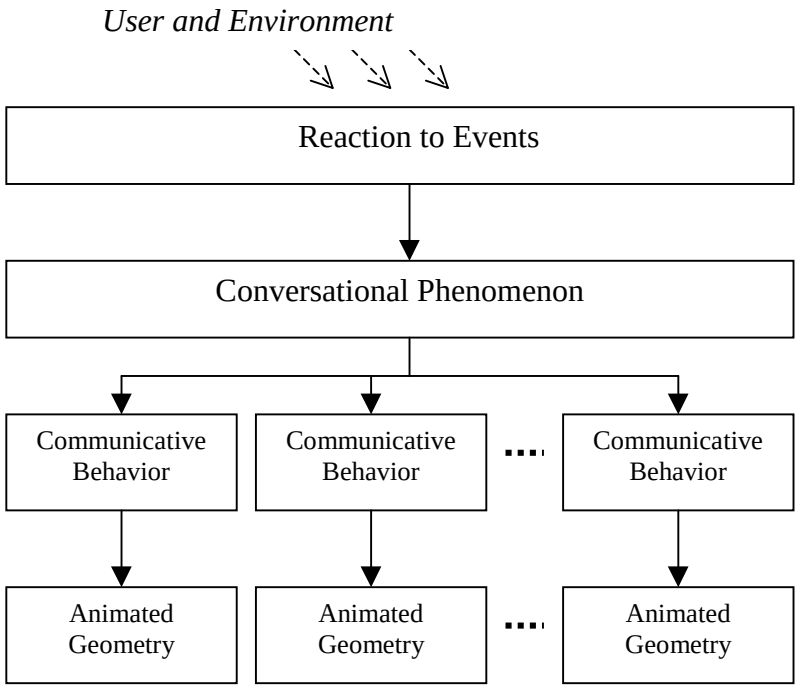


Figure 5: The avatar's behavior control consists of four tiers, where the flow of execution is from top to bottom.

The ***Reaction to Events*** tier defines the entry point for behavioral control. This tier is implemented as a set of functions that get called by the Client when messages arrive over the network or by the avatar's "vision" as the environment gets updated. These functions are listed in Table 3.

This tier is the heart of the avatar automation, since this is where it is decided how to react in a given situation. The reaction involves picking a Conversational Phenomenon that describes the new state of the avatar. This pick has to be appropriate for the situation and reflect, as closely as possible, the user's current intentions. The selection rules are presented in Appendix B.

Function	Event
ReactToOwnMovement	User moves the avatar
ReactToMovement	The conversational partner moves
ReactToApproach	An avatar comes within reaction range
ReactToCloseApproach	An avatar comes within conversational range
ReactToOwnInitiative	User shows interest in having a conversation
ReactToInitiative	An avatar shows interest in having a conversation
ReactToBreakAway	The conversational partner wants to end a conversation
ReactToSpeech	An avatar spoke
Say (utterance start)	User transmits a new utterance
Say (each word)	When each word is displayed by the user's avatar
Say (utterance end)	When all words of the utterance have been displayed

Table 3: The Behavior Control functions that implement the Reaction to Events

The ***Conversational Phenomena*** tier implements the mapping from a state selected by the Event Reaction, to a set of visual behaviors. This mapping is based on the literature presented in section 3.1 and is described in Table 4.

Conversational Phenomena	Communicative Behavior
Approach and Initiation	
Reacting	SHORTGLANCE
ShowWillingnessToChat	SUSTAINEDGLANCE, SMILE
DistanceSalutation	LOOKING, HEADTOSS/NOD, RAISEEYEBROWS, WAVE, SMILE
CloseSalutation	LOOKING, HEADNOD, EMBRACE OR OPENPALMS, SMILE
While chatting	
Planning	GLANCEAWAY, LOWEREYEBROWS
Emphasize	LOOKING, HEADNOD, RAISEEYEBROWS
RequestFeedback	LOOKING, RAISEEYEBROWS
GiveFeedback	LOOKING, HEADNOD
AccompanyWord	Various (see Appendix C)
GiveFloor	LOOKING, RAISEEYEBROWS
BreakAway	GLANCEAROUND
When Leaving	
Farewell	LOOKING, HEADNOD, WAVE

Table 4: The mapping from Conversational Phenomena to visible Behaviors

Finally, each ***Communicative Behavior*** starts an animation engine that manipulates the corresponding avatar geometry in order change the visual appearance. In the current version of BodyChat, merging is not performed when different behaviors attempt to control the same degree of freedom. The behavior that comes in last takes control of that degree.

4.5. Sample interaction

4.5.1. Overview

This section describes a typical session in BodyChat, illustrated with images showing the various expressions of the avatars. The images are all presented as sequences of snapshots that reflect change over time. First is a failed attempt to initiate a conversation, followed by a successful attempt, a short exchange and a farewell.

4.5.2. No interest

User A is walking around, seeking out someone interested in chatting. After awhile A spots a lone figure that is apparently not occupied. A clicks on the other avatar, expressing *Willingness To Chat* (see 4.3). The other Avatar reacts with a brief glance without a change in expression. This lack of sustained attention signals to A that the other user is not *Available* (see 4.3). The automated sequence of glances is shown on figure 6.

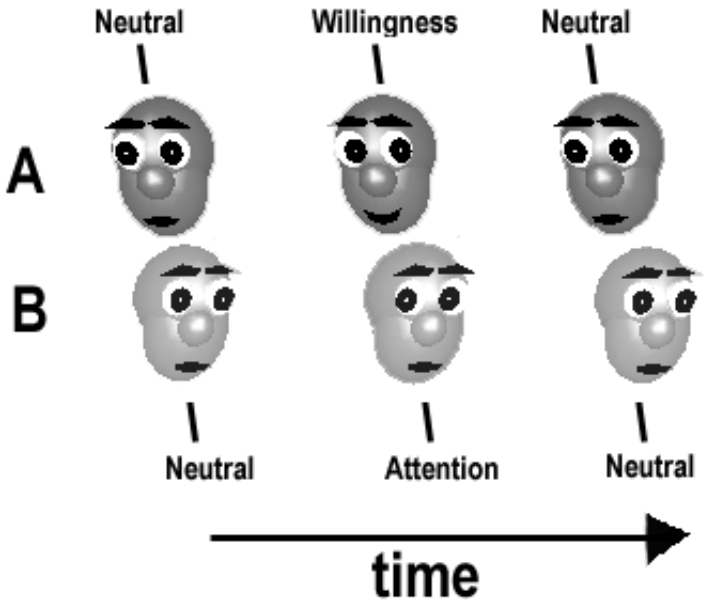


Figure 6: The sequence of glances when user A clicks on avatar B to express willingness to chat while user B is not available.

4.5.3. Partner found

User A continues to walk about looking for a person to chat with. Soon A notices another lone figure and decides to repeat the attempt. This time around the expression received is an inviting one, indicating that the other user is Available. The automated sequence of glances can be seen in figure 7.

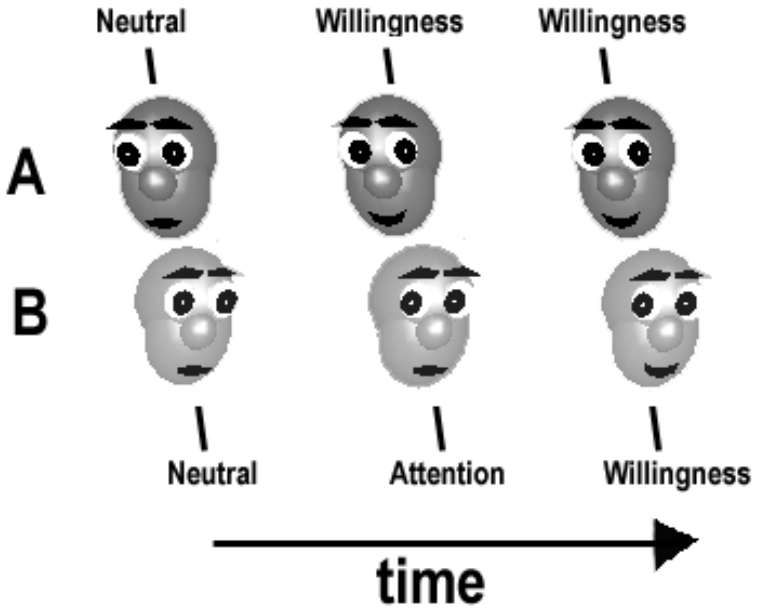


Figure 7: The sequence of glances when user A clicks on avatar B to express willingness to chat and user B is available

Immediately after this expression of mutual openness, both avatars automatically exchange *Distance Salutations* to confirm that the system now considers A and B to be conversational partners. *Close Salutations* are automatically exchanged as A comes within B's conversational range. Figure 8 shows the sequence of salutations.

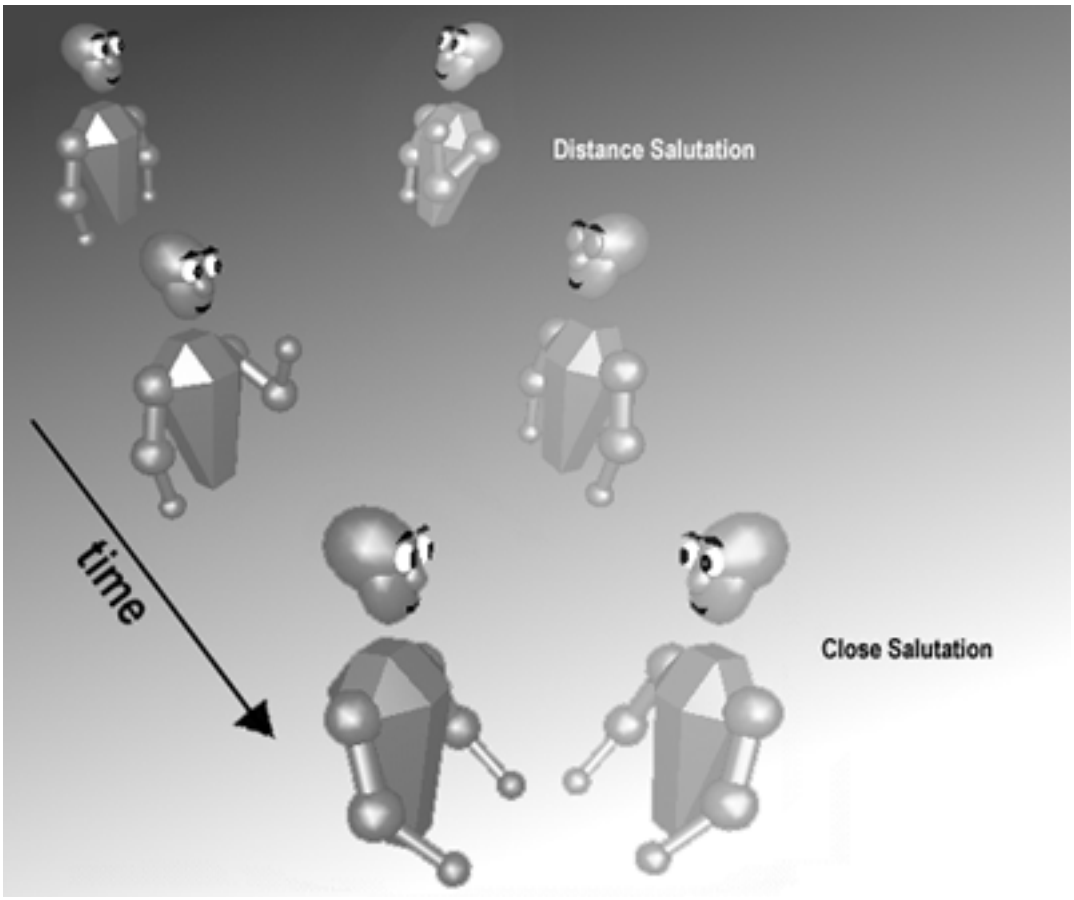


Figure 8: Avatars A and B exchange Distance Salutations when the system registers them as conversational partners. When they get within a conversational range, Close Salutations are exchanged.

4.5.4. A conversation

So far the exchange between A and B has been non-verbal. When they start chatting, each sentence is broken down into words that get displayed one by one above the head of their avatar. As each word is displayed, the avatar tries to accompany it with an appropriate expression (See Appendix C). An example of an animated utterance can be seen in figure 9.

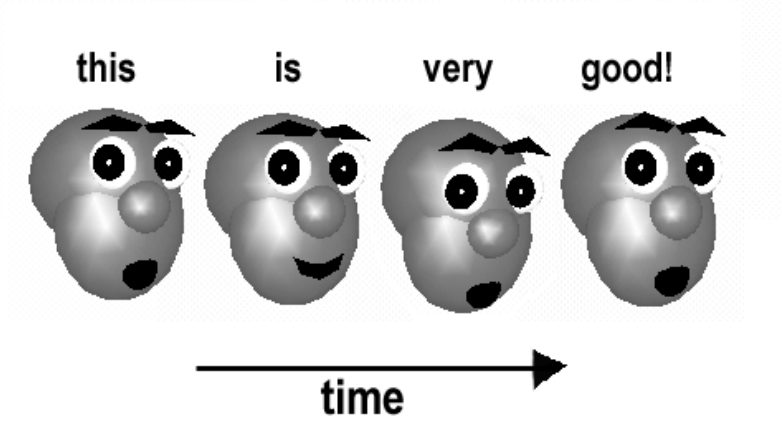


Figure 9: Some words are accompanied with a special facial expression. Here "very" is being emphasized with a nod. The exclamation mark elicits raised eyebrows at the end of the utterance.

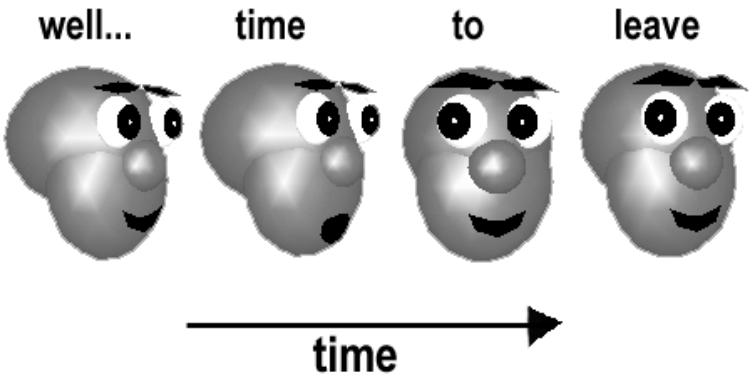


Figure 10: When the user marks a sentence as a Break Away utterance, the avatar displays diverted gaze while reciting the words to give subtle cues to the conversational partner.

Finally, after A and B have been chatting for awhile, A produces a *Break Away* utterance by placing a forward slash at the beginning of a sentence (see 4.3). This makes A's avatar divert its gaze while reciting the words as shown in figure 10. User B notices this behavior and decides to respond similarly, to end the conversation. The avatars of A and B automatically wave farewell and break their eye contact.

4.6. Implementation

4.6.1. Programming platform

BodyChat was written in C++ on an Intel Pentium Pro running Microsoft Windows NT 4.0. Coding and compilation was performed in the Microsoft Visual Studio integrated development environment using Open Inventor graphics libraries from TGS.

4.6.2. Constraints

Keeping graphics performance adequate imposed limits on model complexity. Texture maps were avoided since they slowed down performance considerably.

4.6.3. Major classes

Interface classes were built on the MFC Application Framework, conforming to the document-view approach. The document class contains the state of the client and takes care of communicating with the server. The document also holds a pointer to an Open Inventor scene graph representing the virtual environment and maintains a list of all avatars currently active. Three views on the document are vertically laid out in a splitter window. The largest is the World View that contains an Open Inventor scene viewer for displaying the document's scene graph and a control panel for the user to select the avatar's control parameters. The smaller views are for displaying incoming messages from other users and composing an outgoing message. An avatar is defined and implemented as a separate class.

4.7. Portability

Originally the idea was to build the behavior demonstration on top of an existing Distributed Virtual Environment, in stead of implementing a system from scratch. A lot of effort went into researching available options and finding a suitable platform. It seemed viable to implement the avatar geometry in VRML 2.0 and the behaviors in Java and then use

a VRML/Java compatible browser to view the result. However, that approach was abandoned for a couple of reasons. First, current implementations of the interface between VRML and Java are still not robust enough to warrant reliable execution of complex scene graph manipulation. This may stem from the fact that the VRML 2.0 standard emerged a less than a year ago and browsers have not implemented a full compliance yet. Secondly, most browsers that already have multi-user support implement avatars as a hard-coded proprietary feature of the user interface, rather than a part of an open architecture suitable for expansion. Since this thesis work was not concerned about standardization or reverse engineering of current systems, it was decided to opt for flexibility by using C++ and Open Inventor.

Although the VRML/Java approach was abandoned for current demonstration purposes, it should by no means be discarded as an option, especially when browsers become more robust. In fact, BodyChat introduces an architecture that lends itself well to the separation of animated geometry (i.e. VRML 2.0) and behavior control (i.e. Java). The VRML model would then implement the set of basic communicative behaviors, such as SMILE, NOD, AND RAISEEYEBROWS and the Java module would take care of communicating with the user, environment and other clients to choose an appropriate state for the avatar.

5. Conclusion

5.1. Summary

This thesis has introduced a novel approach to the design and implementation of avatars, drawing from literature in context analysis and discourse theory. The thesis opened by revisiting the notion of cyberspace as a virtual gathering place for geographically separated people. As motivation, it went on to specifically mention chatting, telecommuting, and gaming as some of the major applications for avatar technology. By presenting examples of current systems, it was argued that today's avatars merely serve as presence indicators, rather than actually contributing to the experience of having a face-to-face conversation. In order to understand the important communicative functions of the body, the thesis covered previous research in social sciences on multi-modal communication. Finally the thesis described BodyChat, a system that employs those findings in the automation of communicative behaviors in avatars.

This thesis is more than a presentation of a solution to an engineering problem. It touches on a very important problem concerning embodiment in virtual spaces, notably *how do we map a person onto that person's virtual representation*. In particular, by discussing the various communicative functions of the body, this work brings up the issue of displaying spontaneous and involuntary visual cues that are essential for initiating and sustaining a face-to-face conversation. Since the person sitting at the desktop neither shows the appropriate visual cues for the virtual setting nor consciously thinks about them, we need a way to generate them. This work suggests looking at the avatar as a personal conversational agent, monitoring the user's intentions and applying knowledge about social behavior to come up with appropriate non-verbal cues.

5.2. Evaluation

BodyChat is a prototype that is intended to demonstrate a particular approach to avatar design and implementation. It is not meant as a product ready for distribution and general use, and therefore lacks many of the functions featured in comparable products. However, when comparing the communicative behaviors of avatars in different systems, it is clear that BodyChat starts to fill a vacuum. It presents a new approach that takes avatars from being a mere visual gimmick to being an integral part of a conversation. Although no organized user testing has been performed, reaction to BodyChat has been positive and encouraging, reinforcing the belief that the modeling of autonomous communicative behavior is worthwhile.

Regarding the approach in general, a few limitations should be considered. The first thing to keep in mind is that although communicative non-verbal behavior adheres to some general principles, it is far from being fully understood. Any computational models are therefore going to be relatively simplistic and constrain available behavior to a limited set of displays void of many real world nuances. This raises concerns about the system's capability to accurately reflect the user's intentions under unforeseen circumstances or resolve issues of ambiguity. If the avatar makes a choice that conflicts with what the user had in mind, reliability is severely undermined and the user is left in an uncomfortable skeptical state. The balance between autonomy and direct user control is a really tricky issue.

Another consideration is that it is hard to personalize the autonomous behavior and give it a flavor that reflects the distinct character and mood of the user. A solution may be provided by the use of a template of personality traits filled out for each user that then affects the manner of behavior execution. However the dynamic nature and context dependency of these traits pose a major challenge. Again the question is how much autonomy should be incorporated into the avatar and to what extent the direct control of the user carries the character.

5.3. Future Directions

5.3.1. Expansion in two areas

The issue of avatar control is far from trivial and presents many interesting problems. As described above, the current work introduces an approach rather than a solution. This invites further research, both to see how well the approach can be applied to more complex situations and how it can be expanded through integration with other methods and devices. The following two sections elaborate on two different aspects of expansion. The first deals with the capabilities of the avatar and the second with the monitoring of the user's intentions.

5.3.2. Avatar behavior

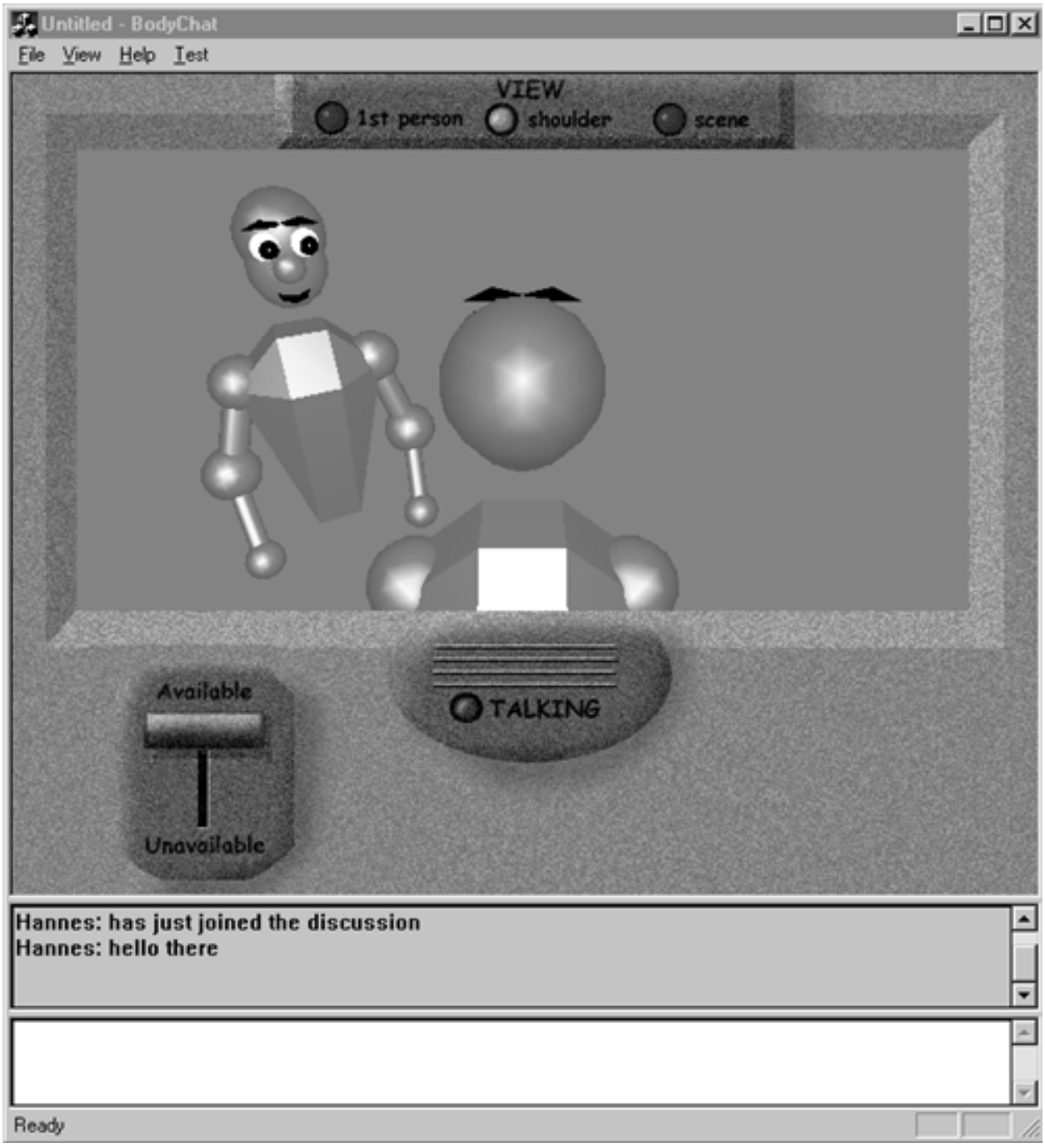
This thesis only starts to build a repertoire of communicative behaviors, beginning with the most essential cues for initiating a conversation. It is important to keep adding to the modeling of conversational phenomena, both drawing from more literature and, perhaps more interestingly, through real world empirical studies conducted with this domain in mind. Behaviors that involve more than two people should be examined and attention should be given to orientation and the spatial formation of group members. The humanoid models in BodyChat are simple and not capable of carrying out detailed, co-articulated movements. In particular, the modeling of the arms and hands needs more work, in conjunction with the expansion of gestural behavior.

5.3.3. User input

An issue that did not get a dedicated discussion in this work, but is nevertheless important to address, is the way by which the user indicates intention to the system. BodyChat makes the user point, click and type to give clear signals about intention, but other input methods may allow for more subtle ways. For example, if the system employed real-time speech communication between users, parameters, such as intonational markers, could be extracted from the speech stream. Although using cameras to directly map the living image of a user onto an avatar is not a good approach, as discussed in section 2.3.4, cameras could still gather

important cues about the user's state. This gathered information would then be used to help constructing the representation of the user's intentions. Other ways of collecting input, such as novel tangible interfaces and methods in affective computing, can also be considered.

Appendix A: User Interface



Appendix B: Reaction Implementation

(Words in *Italics* represent Conversational Phenomena, see section 4.4.3)

ReactToOwnMovement and **ReactToMovement**
→ LookAtPartner

ReactToApproach
→ *Reacting*

ReactToCloseApproach
If Already Saluted at Distance → *CloseSalutation*

ReactToOwnInitiative
→ *ShowWillingness*

ReactToInitiative
If SELF.AVAILABLE
If in CONVERSATIONAL RANGE → *CloseSalutation*
Else If in REACTIONAL RANGE → *DistanceSalutation*
Else
If in REACTIONAL RANGE → *Reacting*

ReactToBreakAway
If SELF.BREAKAWAY → *Farewell*

ReactToSpeech
If it is the current partner that is speaking → *LookAtPartner*
Else → *Reacting*

Say (utterance start)
If long utterance → *Planning*

Say (each word)
→ *AccompanyWord*

Say (utterance end)
If SELF.BREAKAWAY and PARTNER.BREAKAWAY → *Farewell*
Else → *GiveFloor*

Appendix C: Word Accompaniment

In BodyChat utterances are broken into words that are displayed one by one above the avatar's head. The method **`AccompanyWord(Cstring word)`** is called for each word, allowing the avatar to take action based on the words spoken. The current implementation spots a few keywords and punctuation markers and selects an appropriate conversational phenomenon for accompaniment. The mapping presented here is a simple demonstration, but it is easily extendable to elicit a wider range of behaviors.

```
AccompanyWord(Cstring word) {
   if(word.GetLength()>0) {
         word.MakeLower();
         if(word[0]=='*') Emphasize(); // Allows user to emphasize any word
         if(word.Find("you")  > -1) Beat();   // A slight hand wave
         if(word.Find("this") > -1) Beat();
               if(word.Find("very") > -1) Emphasize();
         if(word.Find("yes")  > -1) Emphasize();
         if(word.Find("aha")  > -1) Emphasize();
               if(word.Find(',') > -1) RequestFeedback();
         if(word.Find('.') > -1) RequestFeedback();
         if(word.Left(1)=='?') RequestFeedback();
         if(word.Left(1)=='!') RequestFeedback();
         // …
         // Add more actions here
         // …
   }
}
```

References

(Anderson et al. 1996) Anderson, D.B., Barrus, J.W., Brogan, D., Casey M., McKeown, S., Sterns, I., Waters, R., Yerazunis, W. (1996). Diamond Park and Spline: A Social Virtual Reality System with 3D Animation, Spoken Interaction, and Runtime Modifiability. Technical Report at MERL, Cambridge.

(Argyle and Cook 1976) Argyle, M., Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge University Press.

(Argyle et al. 1973) Argyle, M., Ingham, R., Alkema, F., McCallin, M. (1973). The Different Functions of Gaze. *Semiotica.*

(Blumberg 1995) Blumberg, B. M., Galyean, T. A. (1995). Multi-Level Direction of Autonomous Creatures for Real-Time Virtual Environments. *Proceedings of SIGGRAPH '95.*

(Braham and Comerford 1997) Braham, R., Comerford, R. (1997). Special Report: Sharing Virtual Worlds. *IEEE Spectrum*, March, 18-19.

(Bruckman 1997) Bruckman, A. (1997). MOOSE Crossing: Construction, Community, and Learning in a Networked Virtual World for Kids. Ph.D. Thesis. MIT Media Lab.

(Bruckman and Resnick 1995) Bruckman, A., Resnick M. (1995). The MediaMOO project: constructionism and professional community. In *Convergence* 1(1).

(Carlsson and Hagsand 1993) Carlsson, C., Hagsand, O. (1993). DIVE: A Platform for Multi-User Virtual Environments. *Computers and Graphics,* 17 (6), 663-669.

(Cary 1978) Cary, M. S. (1978). The Role of Gaze in the Initiation of Conversation. *Social Psychology*, 41(3).

(Cassell et al. 1994a) Cassell, J., Stone, M., Douville, B., Prevost, S., Achorn, B., Steedman, M., Badler, N., Pelachaud, C. (1994). Modeling the Interaction between Speech and Gesture. *Proceedings of the Cognitive Science Society Annual Conference*

(Cassell et al. 1994b) Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M. (1994). Animated Conversation: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents. *Proceedings of SIGGRAPH '94.*

(Cassell forthcoming) Cassell, J. (forthcoming). *A Framework For Gesture Generation And Interpretation*. In R. Cipolla and A. Pentland (eds.), *Computer Vision in Human-Machine Interaction*. Cambridge University Press.

(Chovil 1992) Chovil, N. (1992). Discourse-Oriented Facial Displays in Conversation. *Research on Language and Social Interaction*, 25, 163-194.

(Cherny 1995) Cherny, L. (1995). "Objectifying'' the Body in the Discourse of an Object-Oriented MUD. *CyberSpaces: Pedagogy and Performance on the Electronic Frontier.* 13(1,2)

(Curtis 1992) Curtis, P. (1992). Mudding: social phenomena in text-based virtual realities. *Proceedings of the 1992 Conference on Directions and Implications of Advanced Computing*. Berkeley, May 1992.

(Donath 1995) Donath, J. (1995). The Illustrated Conversation. *Multimedia Tools and Applications*, 1, 79-88.

(Donath 1997) Donath, Judith. 1997. Inhabiting the virtual city: The design of social environments for electronic communities. Ph.D. Thesis. MIT Media Lab.

(Falby et al. 1993) Falby, J. S., Zyda, M. J., Pratt, D. R., Mackey, R. L. (1993). NPSNET, Hierarchical Data Structures for Real-Time Three, Dimensional Visual Simulation. *Computers & Graphics*, 17(1), 65-69.

(Gibson 1984) Gibson, W. (1984). *Neuromancer*. New York: Ace Books.

(Goffman 1967) Goffman, E. (1967). *Interaction Ritual: Essays in Face-to-Face Behavior*. Aldine Publishing Company. Chicago.

(Goodwin 1986) Goodwin, C. (1986). Gestures as a Resource for the Organization of Mutual Orientation. *Semiotica*, 62(1/2).

(Greenhalg and Benford 1995) Benford, S., Bowers, J., Fahlen, L.E., Greenhalgh, C., Snowdon, D. (1995). User Embodiment in Collaborative Virtual Environments. *In Proceedings of CHI'95*, 242-249.

(Kendon 1992) Kendon, A. (1992). The negotiation of context in face-to-face interaction. In A. Duranti and C. Goodwin (eds*.), Rethinking context: language as interactive phenomenon*. Cambridge University Press. New York.

(Kendon 1990) Kendon, A. (1990). Conducting Interaction: Patterns of behavior in focused encounters. Cambridge University Press. New York.

(Kurlander et al. 1996) Kurlander, D., Skelly, T., Salesin, D. (1996). Comic Chat. *Proceedings of SIGGRAPH '96.*

(Macedonia et al. 1994) Macedonia, M. R., Zyda, M. J., Pratt, D. R., Barham, P. T., Zeswitz, S. (1994). NPSNET: A Network Software Architecture for Large, Scale Virtual Environments. *Presence, Teleoperators and Virtual Environments*, 3(4), 265-287.

(Mandeville et al. 1995) Mandeville, J., Furness, T., Kawahata, M., Campbell, D.,Danset, P., Dahl, A., Dauner, J., Davidson, J., Kandie, K. and Schwartz, P. (1995). GreenSpace: Creating a Distributed Virtual Environment for Global Applications. *In Proceedings of IEEE Networked Virtual RealityWorkshop.*

(McNeill 1992) McNeill, D. (1992*). Hand and Mind: What Gestures Reveal about Thought*. University of Chicago.

(O'Byrne 1995) O'Byrne, J. (1995). Human Interaction within a Virtual Environment for Shipboard Training. MS Thesis. Naval Postgraduate School, Monterey, California.

(Perlin and Goldberg 1996) Perlin, K., Goldberg, A. (1996). Improv: A System for Scripting Interactive Actors in Virtual Worlds. *SIGGRAPH 1996 Course Notes #25*.

(Prevost 1996) Prevost, S. (1996). Modeling Contrast in the Generation and Synthesis of Spoken Language. *In Proceedings of ICSLP '96*.

(Rheingold 1994) Rheingold, H. (1994). *The Virtual Community*. Secker & Warburg. London.

(Sartre 1956) Sartre, J. P. (1956). *Being and Nothingness*. Translated by Hazel E. Barnes. Philosophical Library. New York.

(Schegloff 1968) Schegloff, E. (1968). Sequencing in Conversational Openings. *American Anthropologist*, 70, 1075-1095.

(Schegloff and Sacks 1973) Schegloff, E., Sacks, H. (1973). Opening up closings. *Semiotica*, 8, 289-327.

(Stephenson 1992) Stephenson, N. (1992). *Snowcrash*. New York: Bantam Books.

(Suler 1996) Suler, J. (1996). The Psychology of Avatars and Graphical Space in Visual MOOs. *WWW, URL=" http://www1.rider.edu/~suler/psycyber/psyav.html"*.

(Thomas and Johnson 1981) Thomas, F., Johnson, O. (1981). *The Illusion of Life: Disney Animation*. Hyperion Press. New York.

(Thorisson 1996) Thorisson, K. R. (1996). Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills. Ph.D. Thesis. MIT Media Lab.

(Thorisson and Cassell 1996) Thórisson, K.R., J. Cassell (1996) Why Put an Agent in a Human Body: The Importance of Communicative Feedback in Human-Humanoid Dialogue. (abstract*) In Proceedings of Lifelike Computer Characters '96*, Snowbird, Utah, 44-45.

(Torres 1997) Torres, O. (1997). Modeling Gaze Behavior as a Function of Discourse Structure. Submitted to the First International Workshop on Human-Computer Conversation to be held in July.

(Weisbrod 1965) Weisbrod, A. T. (1965). Looking behavior in a discussion group. Term Paper submitted for Psychology 546 under the direction of Professor Longabaugh, Cornell University, Ithaca, New York.

(Waldorp 1995) Waldrop, M.S. (1995). Real-Time Articulation of the Upper Body for Simulated Humans in Virtual Environments. MS Thesis. Naval Postgraduate School, Monterey, California.

Additional references and web links can be found at
http://avatars.www.media.mit.edu/avatars