2 Literature review

This chapter describes research related to animating avatars and other kinds of animated characters. Included here are descriptions of both research and commercial virtual environment software, systems for specifying animation through scripting or other higher level language, computer-mediated communication software, and systems for controlling music and abstract animation. We have chosen these fields either because they offer other examples of the type of system we are developing, or because they demonstrate control techniques that may be applicable to this problem.

2.1 Avatars in virtual worlds

The popular conception of an avatar[1] comes from science fiction novels about adventures in virtual worlds. One of the most popular of these was Neal Stephenson's *Snow Crash,* published in 1992. The concept became a reality when the World Wide Web enabled fast transmission of graphical content over the Internet.

The origins of nonverbal communication (NVC) via avatars lie in these online social virtual worlds.The earliest social virtual worlds were developed by the commercial sector.

1. The original meaning of the word comes Hindu mythology: "descent, as of a deity to earth in bodily form."[5]

Essentially they are chat rooms with graphical content added on top. Unlike simple text based chat rooms, where the virtual room is a window in which visitors can see each other's chat text, graphical chat worlds allow users to see the room itself and the people (via their avatars) inside of it. In addition, users can interact with the environment and other people. The simplest form of interaction is navigating the avatar around the room. An example of avatars in such a virtual world is shown in Figure 2-1

Figure 2-1. Typical avatar world in *blaxxun*.

2.1.1 Controlling NVC

A natural next step was to provide a means for visual NVC. NVC did exist prior to the addition of graphics. User's of MUD's, a purely text form of virtual world, developed a number of mechanisms: emoticons such as the ubiquitous `:-)`; well understood acronyms such as `LOL` meaning "laugh out loud"; and parenthetical text meant to be interpreted as descriptions of actions and gestures. No additional user interface widgets were required. Producing nonverbal communication based on avatar motion meant designing interaction techniques beyond typing.

By its nature, the existence of a visualizable and navigable virtual world affords interpersonal distance as a from of communication, a form of NVC often referred to as *proxemics*. Since all virtual world browsers provide a means for navigating the world, control of proxemics is automatically taken care of. Controls for other forms of NVC require widgets that are spe-

cialized for controlling particular modes of nonverbal communication. Most virtual world browsers use standard widgets such as menus and buttons to control discrete forms of nonverbal expression. For instance, selection of an item on a menu changes the facial expression or body posture of the avatar.

2.1.2 Palettes of discrete expressions

One of the first and most popular graphical virtual worlds was *The Palace*. The Palace is a two dimensional, image based virtual world. Each virtual place or room is like a theatre backdrop against which two dimensional avatars can be placed. Figure 2-2 shows a typical room. Visitors communicate through typing text, and their chat text appears in balloons near the visitor's avatar.

Figure 2-2. The Palace, one of the first graphical chat worlds

The Palace provides a simple interface for facilitating nonverbal communication. The user controls the image used for their avatar. If a user has a set of avatar images which repre-

sent the same character with different facial expressions or postures, then changing the avatar image can represent changing the avatar's expression. In The Palace, thumbnails for these different images appear on a panel as shown in Figure 2-3. This panel is in essence a palette of

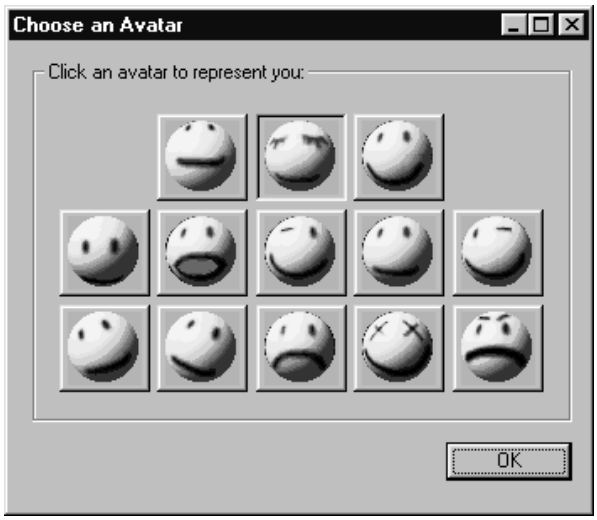

Figure 2-3. Thumbnail panel from The Palace showing the default avatar.

expressions. Selecting a thumbnail (or its equivalent hotkey) changes the avatar image that is displayed. Expression can also be changed by typing emoticons into the chat text box. For instance, typing a `:)` will invoke the *smile* expression.

More technologically sophisticated virtual world systems, such as *Active Worlds* and OZ, employ three dimensional computer graphics. The avatars in these worlds are composed of articulated segments and can perform short animated sequences. Many animations, such as dancing the *macarena* or shooting bolts of lightning, are intended as a form of entertainment. Others are gestural, such as waving and bowing. Some animations are predefined and provided with the software for the virtual world. User's personalize their avatars by designing their own animation sequences. For instance, one AlphaWorld avatar, designed as a martial artist, is programmed to perform karate forms [28]. The mechanism for invoking an animation remains

essentially the same: the user selects an animation from a palette or presses the associated hot keys.

The *Traveler* software developed by the OnLive! is a full three-dimensional avatar world. Traveler software allows spoken verbal communications in addition to text chat. In fact, speaking is the normal mode of verbal communication, and text is often reserved for private messages. The effect of being in a three-dimensional space is enhanced by the simulated two-dimensional audio. As an avatar gets closer, the user's voice becomes louder. The sound attenuates with distance. The avatar's facial animation is automatically synched to the user's voice.

Figure 2-4. Traveler avatar world where avatars are heads.

In Traveler, all avatars are sans bodies (as show in Figure 2-4), so all gestures in Traveler are head and facial gestures. Users can select expressions from a drop down menu. Also, using the keyboard, users can rotate their avatar's head up and down and from side to side. These motions allow them to easily perform the familiar gestures of nodding and shaking their head. When a user is not active for a certain amount of time, their avatar's eyes close and the face "relaxes" into what looks like dozing. This is one of several automated expressive behaviors in Traveler.

The preceding are just a few examples of the many virtual worlds available on the web. (For detailed descriptions of these and other worlds, see [28].) These worlds vary greatly in degree of visual sophistication. Some support audio thought most do not. All of the worlds provide some kind of palette or menu of expressions. The ability to change expression lacks fine control since the user can only select an expression but not modify it in any continuous way. Still, users have made the most of these limited capabilities to express themselves as best they can. In The Palace, users sometimes communicate by rapidly cycling through a series of images. In Traveller, users sometimes spin their head around in space to express exuberance. The limitations of the medium do not seem to stifle the desire of users to communicate non-verbally.

2.2 Desktop virtual environments and nonverbal communication

When researchers became interested in the problem of nonverbal communication in avatar worlds, they developed novel types of interaction widgets in order to provide finer degrees of expression. Like the preceding examples, the interfaces provide a set of gestures from which a user can select. In addition, the widgets allow the user to modulate the expression or gesture in some manner.

2.2.1 Selection and modulation

A nonverbal communication application was developed for *VLNET* by Guye-Vuillème [45]. The VLNET project comprises several research projects to develop the technologies required for various tasks in very large networked 3D virtual environments with realistic looking avatars. The nonverbal communication application addresses the need for users to express themselves using facial expression and body language within the virtual environment. The interface consists of three panels: a panel for selecting affect display, a panel for

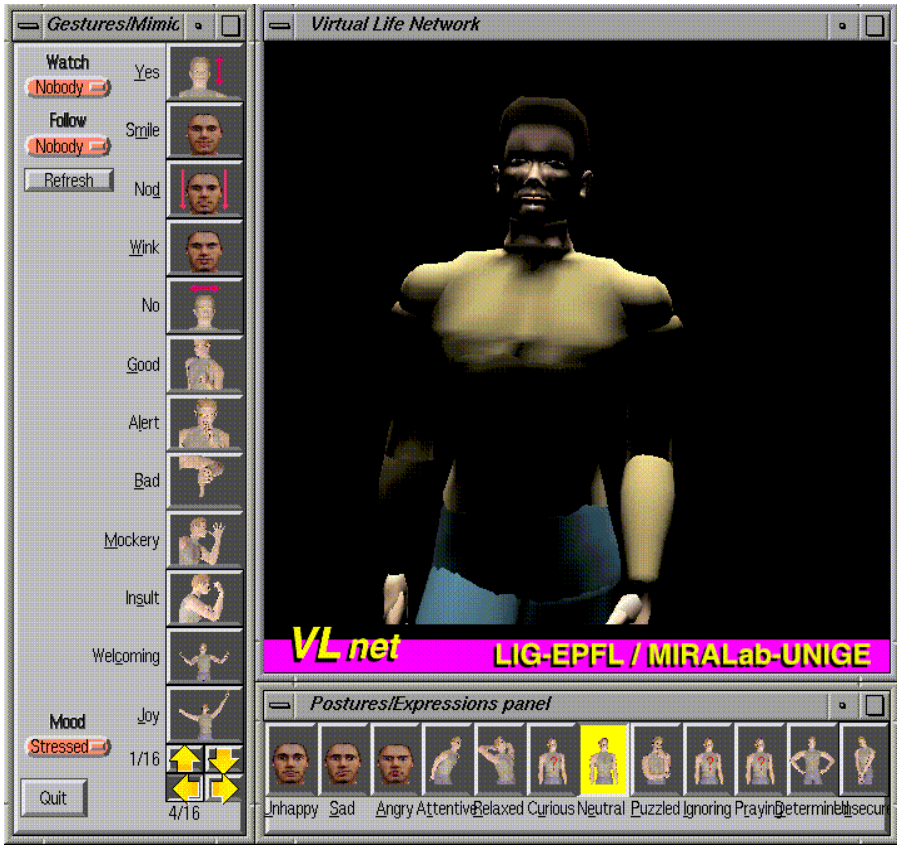

Figure 2-5. Gesture/Mimics panel from VLNET NVC application.

selecting body and facial gesture, and a panel for selecting actions. This organization separates postures and expressions which usually signify a user's state of mind or emotion, from actions and gestures which have a more symbolic meaning. (See Figure 2-5.) Users are also able to modify the animation speed of avatar actions using a slider.

The expressions that are available are *affect displays*—facial expressions and body postures that indicate emotional state—and *emblematic* gestures. The researchers are particularly concerned with the control of affect displays because of their focus on the use of NVC to create a "more friendly" virtual environment. Emblematic gestures were included because this type of gesture is produced consciously by people in face to face conversation. This intentionality makes it easier to map standard desktop UI's to their control. In addition, emblematic ges-

tures have a specific form, so they are easy for a designer to create in advance without input from the user. They deferred the issue of controlling gestures that accompany speech because of the technical problem of synchronizing the gestures with speech, and also because they feel that controls for these gestures would require too much of the user's attention.

They conducted user studies in which they asked if users could replicate their real life relationships with another person. They had three pairs of interactants and each pair had a different level of mutual familiarity. They were given two hours of interaction time during which they were not given any specific instructions as to how to interact. Following the study, they were given a survey about their reactions towards the interaction.

Slater et al developed an NVC interface for avatar expressions in order to study the feasibility of virtual reality for rehearsing what would eventually be a live dramatic performance [99]. Verbal communication took place using speech. Unwittingly they had designed a system in which the users were able to gesture with their avatar's head. Users control their viewpoints, and subsequently the position of the avatar's head, by moving the mouse. Minute movements of the mouse allowed the actors to generate head nods, wags and glances towards and away from their interlocutors. The arms can be moved up and down either separately or independently using sliders. Simple full body actions such as standing and sitting were predefined.

This project also developed a novel interface for facial expression. The actors created expressions by drawing eyebrows and a stylized mouth onto a generic smiley face template. Mouths and eyebrows are drawn with the mouse using $\vee$ and $\wedge$ shaped strokes, respectively. The widget is shown in Figure 2-6. The intensity of the expression is modified by adjusting the slopes of the angles on the input gesture.

Figure 2-6. Controls for acting in virtual reality application.

Their study was funded by the BBC who would like to have more flexibility in hiring actors who have complicated travel schedules. They wanted to study the ability of actors and directors to rehearse in a VE setting for a real life performance. Pairs of actors with a director met three times virtually before meeting once face to face. The virtual meetings were used to develop and rehearse a dramatic scene. At the face-to face meeting, they performed a fifteen minute acting scene. After the study they surveyed the actors and directors to determine their sense of presence, co-presence and cooperation in the virtual environment. They were able to determine that they could perform this task better over VR than over video conferencing, or mere learning of lines.

ComicChat, is a 2D chat environment in which avatars appear as comic strip characters, and the virtual wold is viewed as evolving comic strip frames[63]. A frame from a ComicChat

session is shown in Figure 2-7. Users have the ability to select emotional bodily expressions

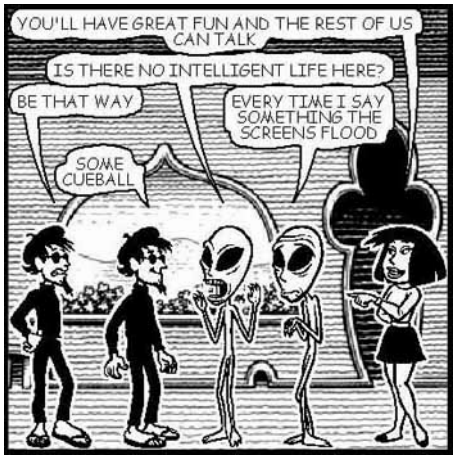

Figure 2-7. Comic Chat avatars.

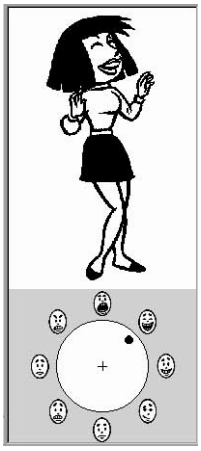

Figure 2-8. Emotion wheel.

using an *emotion wheel,* pictured in Figure 2-8. Icons arranged around the circumference of the circle represent the available expressions: *coy*, *happy*, *laughing*, *shouting*, *angry*, *sad*, *scared* and *bored.* The center of the wheel selects the *neutral* expression. A user selects and modulates an expression by clicking within the circle. The direction selects the emotion and the magnitude of the expression is calculated according to the distance between the center of the circle and the point where the driver clicks. Users are also able to select *actions* for their avatars; that is, they can choose to have their avatars drawn *waving* or *pointing.*

VLNET, the acting interface and ComicChat each add the ability to modulate expression. For a facial expression or posture, magnitude may be the only parameter that needs to be varied. However, an animated gesture, with its spatial, temporal and dynamic extent, can vary in several ways at once. The VLNET interface allows selection of animated actions and gestures, but modulation of only speed.

2.2.2 Avatar as communication agent

A difficulty with managing the NVC through an avatar is coping with the added cognitive load of controlling the avatar motion on top of typing text. This problem is somewhat alleviated in audio-enabled virtual world software such as Traveller and the acting interface. When verbal communication is through voice, then the hands are free to control the avatar. But when the hands are responsible for both the verbal—using words—and nonverbal communication, any kind of manual control can seem cumbersome. To alleviate this problem many graphical chat worlds will automatically change the avatar expression when the user types an emoticon. In this way, verbal and nonverbal communication are handled seamlessly. Modulation, however, remains unaddressed.

In addition to using emoticons, the ComicChat system also analyzes the chat text for keywords, punctuation and sentence structure. Typing a sentence ending in a question mark causes the avatar to shrug. Typing a sentence ending in an exclamation point invokes the *shouting* action.

The *BodyChat* system developed by Hannes Vilhjálmsson at MIT treats avatars as communicative agents that act on behalf of their users to fulfill communicative aims [110]. This project is concerned with *regulators*, the nonverbal signals used in conversation regulation. Users specify high level goals such as initiating, rejecting or ending conversations. The avatars automatically handle the nods, glances and other gestures required to negotiate the interaction. The user carries out verbal conversation through typing.

The *BodyChat* system is an outgrowth of research lead by Justine Cassell. In [20] Cassell and Vilhjálmsson propose that the design of an avatar conversation system should be based on a theory of conversational agents.The theory, they suggest, should account not only for the

various functions of conversational bodily behaviors, but also the user's preference for controlled versus autonomous avatar animations. BodyChat is designed to facilitate the generation of *regulators*, gaze behaviors and facial displays that regulate the flow of conversation. The designers of the system feel that autonomous behaviors are more appropriate for producing regulators because in the real world humans produce these signals without conscious intent. Hence, many of the avatar's behaviors are generated as a result of keywords and punctuation in the (typed) verbal conversation. The avatar uses the state of the environment and conversational rules in order to generate the appropriate displays.

They conducted user studies in which users met in a virtual world [20]. Subjects used one of three versions of BodyChat: the autonomous behavior version, a manual version in which communicative displays were selected from a pulldown menu, and a version with both autonomous and manual behaviors. Subjects who used the autonomous system reported a greater sense of naturalness, expressiveness and control over the conversation than did subjects who used the manual controls alone.

2.2.3 Summary of avatar world controls

The kinds of controls and the controllable behaviors that were described are summarized in Table 2-1

Control	**Behavior type**	**Applications**
Emotion wheel	Affect display	ComicChat
Select emotion from among buttons on panel		VLNet
"Drawing" expression on generic face		Slater
Choose selection on pop-up Menu	Action/Animation	ComicChat
Buttons on panel		VLNet
Sliders on arms		Slater
By navigating avatar around world	Proxemics	All
User selects high level goal	Regulators	BodyChat

Table 2-1. Nonverbal behaviors and their controls in existing avatar systems.

Control	Behavior type	Applications
Select gesture from button on panel	Emblematic gesture	VLNet
Automatic blinking and breathing	Adaptors	VLNet
Automatic blinking, breathing and gaze lock		Slater
Automatic blinking and breathing		BodyChat
Turning body and lifting hand, and turning head	Deictic gesture	Slater
Performance capture	All gesture	Full body motion capture

Table 2-1. Nonverbal behaviors and their controls in existing avatar systems.

2.3 Synthetic actors and scripted animation

Synthetic actors are autonomous animated characters with believable behaviors and expressive movements. Initially this work arose in answer to the question of how to create animations through high level commands instead of the usual painstaking creation of animation through key framing. The user acts as a director who creates and directs a virtual actor. The director writes a script defining action sequences for the actor to perform. The animation system animates the actor's geometry according to the script.

Improv, developed by Perlin and Goldberg at NYU's Media Research Laboratory is one such system for authoring behavior-based animated characters [82]. Authors define the behaviors for the character at several levels. At the lowest level authors define a set of actions, such as *wave* or *sit*, and then specify how each action will be animated. These characters, called *actors* in the Improv system, have articulated bodies whose parts are possibly deformable. Specifying an action means defining how the avatar's parts will move—or its mesh will deform—for the duration of the action. These animation specifications include not only the path of the parts over time, but also a level of noise that will be added to the motion to make the motion look more lifelike.

At a higher level, authors define scripts. Scripts describe higher level behaviors such as *dancing*, or *joking*. They are made up of sequences of actions and other scripts where each action

or script may have a triggering event which initiates it. In addition, authors create relations among actions and scripts so that they will be executed in a reasonable way at run-time. For instance, some actions can be performed simultaneously, while others cannot. The relations specify, among other things, compatible actions.

Authors can also create user interface widgets to give end-users interactive control over the actor. Using the widgets, the user can trigger actions or scripts. In this system, the expressiveness is built into the actor by the author. Actions and gestures are performed differently during each invocation due to the addition of noise signals to the animation.

Blumberg and Gaylean at MIT developed a similar system for creating reactive autonomous creatures [14]. (We will use the word *creatures* since that is the word they use and because their characters are usually animals and not humans.) Like IMPROV, their animation controls are directed by higher level behaviors. A behavior in their system arises from the interaction of a network of separate goal-directed behaviors such as "reduce hunger" or even "chew food." Which of these goal-directed behaviors gains control over the animation is based on the internal state of the creature, stimuli from the environment and the organization of the goal-based behaviors within the network.

The system is loosely based on ethological models of animal behavior. Directability of the creatures is implemented at three levels within the behavior model. At the highest level the behavior system can be tweaked by adjusting the goals and other internal state of the system. At the next level, particular actions that are usually initiated by the behavior system can be forced to execute at any time. At the lowest level, the sensory input to the creature can be modified.

Finally, a director can issue direct motor commands to make a creature walk or lie down. When the behavior system is running, the directed actions will only occur if they are not excluded by a particular goal-directed behavior. If the behavior system is shut down, then the director has complete control over the creature. This model for directing autonomous creatures later became the behavioral engine for creatures in the Alive project [71].

The Jack system at the University of Pennsylvania creates models of humans for a broad range of applications requiring simulated humans [7]. Research with the Jack system emphasizes various fidelity issues depending on the type of application and task being targeted. Because their models can be used in simulations to determine ergonomic factors, cockpit layout for example, their geometry models are built to human dimensions, and the degree of articulation of the joints is set to have human limits.

Like the other synthetic actors, Jack characters can be incorporated in scripted scenarios where they both follow scripted tasks and react to other characters and objects in the environment. In order to adapt the Jack system to applications in which the character can be treated as a controllable avatar, they have developed a method for converting text commands into animated actions. This controllable figure is implemented in a *JackMoo* system, a combination of the Jack system, lambdaMoo (a text-based virtual world[2]), and a widget into which a user can type instructions for their avatar. A typical instruction string is *pick up the remote control.*

[42] describes work by Gibet, Lebourque and Marteau on system for gesture specification and generation. Their intent is to create a system that allows rapid turnaround between the specification of a gesture and its animation. Gestures are specified in their system using a

2. See online reference in Appendix A.

formal language. The language itself is qualitative and text-based. Only the movements of the arms and hands are specified by their language.

Movement primitives in their language are drawn from the features used to describe formal Sign Languages and French Sign Language in particular. These features describe the path of the endpoint of the arm through space, and the configuration and orientation of the hand. For instance, they have identified seven movement primitives in French Sign Language: *pointing, straight-line, curve, ellipse, wave* and *zig-zag*. These movements are parameterized by locations in space. To specify location, the space around the body has been quantized into a set of directions around the body and distances from the center of the body. Also, a number of named points on the body can be used as locations. Similarly, hand orientations are parameterized by these directions and locations.

The language includes operators to specify the temporal aspects of gestural movement. The simplest gesture description is made up of a sequence of movement primitives made by one limb. Synchronization operators specify the timing of the movements within the sequence either in absolute time or relative to the duration of the entire movement. Synchronization operators operate on simple gestures to describe gestures that are made using both arms in parallel. Finally, temporal operators applied to sequences of complex gestures are used to describe gestural phrases.

To generate animations of the gestures, the descriptions in this qualitative language are compiled into quantitative commands. The commands are used to drive the control system for a synthetic human with articulated arms, hands and figures. Motor controllers for each joint calculate the joint trajectory in real-time. They find that the description files required to gener-

ate signs and phrases are compact. For example, they say that 60 lines were needed to describe a particular sentence of four signs.

2.4 Visualizing Computer-Mediated Communication

Avatars are not the only visual representation that can be used to communicate nonverbally. Signals can also be transmitted using motion, colors and abstract geometric shapes. This section describes projects that explore different visualizations that facilitate interpersonal communication.

2.4.1 Chat conversation visualization

Collaboration-at-a-Glance is a program for visualizing the dynamics of online chat conversations developed by Judith Donath [33]. In this visualization, the participants appear as photographic images of the person's face. The user has control over the layout of images within a window. Chat text appears in a separate window. To type in text that is directed at a particular person, the user clicks on the picture of that person. The targeted person, whose image was clicked, sees the image of the first user (the user who did the clicking) as a picture of the first user's face gazing directly out of the screen, as if actually looking at them. Other participants see the respondent's image turn its gaze to the other person. Through this depiction of gaze, one can tell by looking at the window who is speaking to whom, and who is getting the most attention from others in the conversation. People may also be represented as line drawings instead of a photograph. The meaning of having a line drawing representation instead of a photograph depends on the configuration of the program. For instance, it may indicate that someone has not spoken in a while.

Collaboration-at-a-Glance attempts to recreate some of the nonverbal cues that are lost during some kinds of online communication. In particular, it addresses a cue that is lost during

video conferencing, namely gaze determination. This work represents another direction that can be taken with recreating nonverbal communication. In the case of this program, attention cues are visualized through animated images which appear to gaze at one another.

Chat Circles is another program for visualizing online chat that was developed by Fernanda Viegas with Judith Donath [109]. Chat Circles is presented as an alternative to avatar-based graphical chat worlds. One of the most interesting aspects of this program is how it uses graphics to convey signals about the overall level of activity in a chat room—something that is usually conveyed by sound in "real life." Instead of human looking avatars in a virtual world, participants and the virtual space they inhabit are represented as animated circles in a window.

Each user chooses a color for their circle when they log into the system. When a user types something, the text they type appears within their circle, which expands in order to encircle all of the text. At the same time, the circle brightens. The text remains on the screen for a short while and then slowly fades. If the user is silent for some time, their circle begins to dim. Thus, with color and shape, the visualization indicates who is being active in the conversation. The way that circles expand and deflate over time provides information about length of time since an utterance, something that does not exist as an artifact in the physical world. It takes advantage of the affordances of an electronic medium for recording and representing this temporal phenomenon. This work makes up for deficiencies in chat worlds where typing is the sole means for verbal communication.

2.4.2 Expressive abstract communication

Other work in computer science explores means for nonverbal expression that does not accompany verbal communication. The first is a digital baton project by Theresa Marrin [72]. The digital baton is a musical conductor's baton that is instrumented with sensors to

determine its position, velocity, acceleration and orientation. The baton is used for gestural input just as a conductor's baton would be. The result is a live computer music performance that is controlled by the user. In continuing work [73], she developed a conductor's jacket that picks up other physiological signals that might be useful for the synthesis of expressive musical performance. These devices extract physiological data that is expressive and emotional in an abstract musical sense. She is interested in a kind of gesture that must be coordinated with the temporal signal music.

Work by Scott Snibbe and Golan Levin on interactive Dynamic Abstraction addresses the question of using a computer for expressive imagistic communication in an intuitive and yet potentially sophisticated manner [102]. They developed a system that has many features that are desirable in a NVC control interface: tracks continuous input, output is more sophisticated than initial input, non-modal interaction, the "process of interaction is also the product of interaction", and it is used to mediate communication with another person. They asked the questions, does personality come through, and does the tool allow someone greater expressive power with greater adeptness with the tool. They conducted observations of users at a series of art exhibits where gallery visitors acted as naive users and used these observations to find structure in the users' interactions and patterns in the development of users' skills and interactions styles over time.

2.5 Summary

Since their inception, avatar worlds have provided users with some means for nonverbal expression. Designers of these worlds intuitively feel that these behaviors will facilitate interactions in the virtual environment. Interfaces for virtual world software map traditional GUI elements such as menu items and palettes of buttons to a selection of static facial displays

or predefined animations. Both the design of the virtual world graphics and the user interface are constrained by the need to run on a wide variety of possibly older hardware and appeal to a wide audience more interested in online entertainment than technology.

Efforts in research in networked virtual environments try to increase the communicative power of the avatar through two main means. The first is to provide the user with control over the continuous variations in the expression of the avatar in addition to a selection of different expressions. This makes the expressions more life like since real human NVC signals can vary in their quality as well as their type. A second means is to relieve the user of the burden of directly controlling the nonverbal behaviors of their avatar. In these systems, the expressions are generated as a result of the user's verbal communication or of changes in the avatar's environment. The burden of generating appropriate behaviors shifts to the designer of the system software. The system must infer from the user's verbal communication or from other nonverbal cues, such as the navigation of the avatar, the communication intent of the user.

A related effort in computer animation is the development of directable, animated characters. Here the researchers find an algorithmic means of adding personality to the scripted or programmed actions and behaviors of their characters. The goal is to lesson the manual labor required to animate characters.

Other non-avatar visualizations for chat environments explore the ways that the computer can capture and express conversational phenomena that have no physical equivalent in the real world. Such work is perhaps more able to explore the particular affordances of computer mediated communication because their representations are not based on human likeness but on visual abstraction of human interaction behaviors. A last area of nonverbal communi-

cation and gestural interfaces considers communication through musical and animated abstract visual forms.