In *Autonomous Agents and Multi-Agent Systems*, **2**(1): 45-64, 1999.

Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous

J. CASSELL AND H. VILHJÁLMSSON {justine,hannes}@media.mit.edu
MIT Media Laboratory, 20 Ames Street, Cambridge, MA 02139, USA

Abstract: Although avatars may resemble communicative interface agents, they have for the most part not profited from recent research into autonomous embodied conversational systems. In particular, even though avatars function within conversational environments (for example, chat or games), and even though they often resemble humans (with a head, hands, and a body) they are incapable of representing the kinds of knowledge that humans have about how to use the body during communication. Humans, however, do make extensive use of the visual channel for interaction management where many subtle and even involuntary cues are read from stance, gaze and gesture. We argue that the modeling and animation of such fundamental behavior is crucial for the credibility and effectiveness of the virtual interaction in chat. By treating the avatar as a communicative agent, we propose a method to automate the animation of important communicative behavior, deriving from work in conversation and discourse theory. BodyChat is a system that allows users to communicate via text while their avatars automatically animate attention, salutations, turn taking, back-channel feedback and facial expression. An evaluation shows that users found an avatar with autonomous conversational behaviors to be more natural than avatars whose behaviors they controlled, and to increase the perceived expressiveness of the conversation. Interestingly, users also felt that avatars with autonomous communicative behaviors provided a greater sense of user control.

Keywords: Avatars, embodied conversational agents, lifelike, communicative behaviors.

1. Behaviors in avatars

One type of embodied agent that has received much airplay but little serious research attention in the agent community, is the avatar in a graphical chat. An avatar represents a user in a distributed virtual environment, but has until now not been autonomous. That is, it has not had knowledge to act in the absence of explicit control on the part of the user. In most current graphical chat systems the user is obliged to switch between controlling the avatar behavior and typing messages to other users. While the user is creating the message for her interlocutor, her avatar stands motionless or repeats a selected animation sequence. This fails to reflect the natural relationship between the body and the conversation that is taking place, potentially giving misleading or even conflicting visual cues to other users. Some voice-based systems offer simple lip synching, which greatly enhances the experience, but actions such as gaze and gesture have not been incorporated or are simply produced at random to create a sense of "liveliness".

The development of graphical chat environments from text-based IRCs (Inter-Relay Chat) indicates an awareness on the part of designers of the importance of the body, and of different communication modalities. More recently, the creators of multi-user environments have realized that avatars need to be animated in order to bring them to life, but their approach has not taken into account the number and kind 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 largest problem with this approach is that the user has to explicitly control every change in the avatar's state. In reality however, many of the visual cues important to 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.

Finally, when designers 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*. However, lively emotional expression in interaction is in vain if mechanisms for establishing and maintaining mutual focus and attention are not in place [10]. 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 [8]. 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 [12].

Our approach (described more fully in [8], in contrast, begins to develop a theory of embodied conversational agents. As well as relying on findings from the literature on discourse and conversation analysis of human-human conversation, we look at the **function** of communicative behaviors like eye gaze, head nods, and eyebrow raises in conversations between humans and machines, and in conversations between humans that are mediated by machines. Is there a role for interfaces that look like humans, and that generate of their own accord appropriate conversational behaviors? A theory of embodied conversational agents needs to account not only for conversational functions and signals, but also for users' preferences for direct manipulation vs. agent autonomy. Avatars are a kind of agent where this design problem – i.e. the integration of research finding on communicative behavior and issues of interface design – is particularly interesting. That is, avatars are representing other users, who have their own communicative intentions, rather than systems, which have no *a priori* intentions about how and what to communicate. In this paper, we report research on what we have elsewhere called the *conversational envelope* [10] and its role in supporting a natural, conversational experience for the user in an avatar-based system.

2. State-of-the art in avatars

The term avatar has been used when referring to many different ways of representing networked users graphically. 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). Active Worlds Browser (Circle of Fire Studios Inc.), blaxxun Community Client (blaxxun

Interactive) and Oz Virtual (Oz Interactive) are all examples of DVE systems. The ideas presented here are still applicable to other kinds of systems and should be viewed with that in mind. For example, although the popular Palace (The Palace Inc.) and WorldsAway (Fujitsu Systems Business of America) systems place avatar portraits on top of flat image backdrops rather than rendering a virtual environment the basic interaction principles are the same. To give a better idea of the current state of the art and the shortcomings of these current systems, a description of a typical DVE, and some popular commercial systems, is in order.

A client program connects each user to the same server responsible for maintaining the state of the shared world. The client renders a view into the world as seen through the avatar's eyes or through a camera floating above and behind the avatar. The avatars are articulated 3D models that the users choose from a menu of available bodies or construct using a supplied editor. The articulation may include arm and feet motion but rarely facial expression or gaze movement (Oz Virtual does provide some facial expressions). The users can freely navigate their avatars 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 often appears floating above the head of the user's avatar.

When the user wants to initiate a contact with another person, three steps can be taken, of which only the last is necessary. First the user can navigate up to another avatar, attempting to enter the other person's field of view. The user can also select from a set of animation sequences for the avatar to play, 'Waving' being the most appropriate for this situation. Finally, the user must start a conversation by transmitting a sentence into the space, preferably addressing the person to contact. Only this last step is necessary because the user's greeting sentence will be 'heard' by all avatars in the virtual room or space, regardless of their avatar's exact location or orientation. During the conversation, the user keeps typing messages for transmission, switching avatar animations from a set such as 'Happy', 'Angry', 'Jump' and 'Wave' as appropriate. Between the selected animation sequences, idle motions, such as stretching and checking watches are randomly executed.

Upon entry into a world like this, one notices how lively and in fact life-like the world seems to be. A crowd of people that is 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 in the virtual room. One reason for this redundancy is that the bodies in these systems 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 often irrelevant. The manually executed motion sequences allow a few explicit (and somewhat exaggerated) emotional displays, but since they are often 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.

A typical session may look like this:

> 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, several virtual miles away, 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.

Our approach attempts to ameliorate the unnaturalness of this experience by exploring how to more fully integrate the behaviors of the body with conversational phenomena.

3. Making avatar behavior autonomous

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 does not map appropriately onto an avatar standing in a group of other avatars. Thus tracking does not lend itself well to Desktop Virtual Environments.

The approach to avatar design adopted here, in contradistinction to explicit control, treats the avatar as an autonomous agent acting of its own accord in a world inhabited by other similar avatars. However the autonomy is limited to a range of communicative expressions of the face and head, leaving the user in direct control of navigation and speech content. The avatar shows appropriate behavior based on the current situation and user input. One can think of this as control at a higher level than in current avatar-based systems. This approach starts to address the following problems:

- **Control complexity:** The user manipulates a few high-level parameters, representing the user's current intention with respect to conversational availability, instead of micromanaging every aspect of animating a human figure.
- **Spontaneous reaction:** The avatar shows 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 occurrences 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. Control at an intentional level and autonomy at the level of

involuntary communicative behaviors may however allow the avatar to give the cues that are appropriate for the virtual situation.

4. Human communicative behavior

In order to automate communicative behaviors in avatars, one has to understand the basic mechanisms of human to human communication. 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 or modalities 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 that feedback over a secondary channel, such as by facial expression, while receiving information over the speech channel [2]. The channels can also work together, supplementing or complementing each other by emphasizing salient points [12][23], directing the listener's attention [14] or providing additional information or elaboration [20][8]. When multiple channels are employed in a conversation, we refer to it as being multimodal.

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 use of gesture and body posture is also very important, but the required elaborate articulation of a human body is beyond the scope of this current work.

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.

> Paul is standing by himself at a cocktail party, looking out for interesting people. Susan (unaquainted with Paul) walks by, mutual glances are exchanged, Paul nods smiling, Susan looks at Paul and smiles [distance salutation] ([15], 173; [7], 269). Susan touches the hem of her shirt [grooming] as she dips her head, ceases to smile and approaches Paul ([15], 186, 177). She looks back up at Paul when she is within 10' [for initiating a close salutation], meeting his gaze, smiling again ([15], 188; [2], 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] ([15], 188, 193). Then she steps a little to the side to face Paul at an angle ([15], 193; [2], 101). A conversation starts.

During the conversation both Paul and Susan display appropriate gaze behavior, such as looking away when starting a long utterance ([15], 63; [2], 115; [12], 177; [11]), 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 ([3]; [12], 177; [10]), giving feedback while listening in the form of nods, low "mhm"s and eyebrow action ([12], 187; [24]; [10]) and finally giving the floor to the other person using gaze ([15], 85; [12], 177; [3]; [2], 118).

Speakers choose conversational partners but do not choose to raise their eyebrows along with an emphasis word, or to look at the other person when giving over the floor. Yet we attend to these clues as listeners, and are thrown off by their absence [16]. In BodyChat, we have implemented these communicative behaviors as a function of their volitional status. That is, we distinguish between user choices, such as who to speak to and when to end the conversation, and body behaviors, such as meeting the gaze of somebody one has chosen to converse with.

5. Related work

Embodiment in Distributed Virtual Environments has been a research issue in systems such as MASSIVE at CRG Nottingham University, UK, where various techniques and design issues have been proposed [5]. There it is made clear that involuntary facial expression and gesture are important but hard to capture. Avatar autonomy however is not suggested. Popular Internet based chat systems that connect a number of users to graphical multi-user environments, such as the early WorldChat from Worlds Inc., have shown that graphical representation of users is a compelling alternative to purely text-based systems. However these systems have not been able to naturally integrate the graphics with the communication that is taking place.

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 [13], where on-screen participant's gaze direction changes to display their attention, and Microsoft's Comic Chat [17], 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 behaviors.

Creating fully autonomous agents capable of natural multi-modal interaction entails integrating speech, gesture and facial expression. By applying knowledge from discourse analysis and studies of social cognition, we have developed systems like Animated Conversation [9] and Gandalf [26]. Animated Conversation renders a graphical representation of two autonomous agents engaged in conversation. The system's dialogue planner generates the conversation and its accompanying communicative signals, based on the agent's initial goals and knowledge. Gandalf is an autonomous agent capable of carrying out a conversation with a user and employing a range of communicative behaviors that help to manage the conversational flow. Both these systems are good examples of discourse theory and studies of human communication applied to computational environments, but neither is concerned with representations of user embodiment and issues of avatar control.

The real-time animation of lifelike 3D humanoid figures has been greatly improved in recent years. The Improv system [22] demonstrates a visually appealing humanoid animation and provides tools for scripting complex behaviors, ideal for agents as well as avatars. Similarly the Humanoid 2 project deals with virtual actors performing scripts as well as improvising role-related behavior [27]. However, automatically generating the appropriate communicative behaviors and synchronizing them with an actual conversation between users has not been addressed yet in these systems. Real-time external control of animated autonomous actors has called for methods to direct animated behavior on a number of different levels such as in ALIVE [6] and in the OZ Project [4]. In this sense, the goals of BodyChat are similar, but the set of behaviors is different. Here we focus on those behaviors that accompany language. We also introduce, for the first time in this literature, a distinction between *conversational phenomena* and *communicative behaviors*.

6. BodyChat

BodyChat is a system that demonstrates the automation of communicative behaviors in avatars. The system consists of a Client program and a Server program. Each Client is responsible for rendering a single user's view into the Distributed Virtual Environment (see Figure 8). All users connected to the same Server see each other's avatars as a 3D model representing the upper body of a cartoon-like humanoid character. Users can navigate their avatars 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.

6.1 User Choices

The avatar's communicative behavior reflects its user's current intentions and the avatar's knowledge of communicative rules. 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 remote avatars. BodyChat implements three control parameters as described in Table 1.

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 1. 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 welcomes other people that show interest in having a conversation. This has an effect on the initial exchange of glances and whether salutations are performed that confirm the newcomer as a conversational partner. Changing Availability has no effect on a conversation that is already taking place, and is switched ON or OFF through a toggle switch on the control panel (see Figure 8).

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 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, or when one avatar is moved out of conversational range, is the conversation broken [15][25].

6.2 Generated Behaviors

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 1).

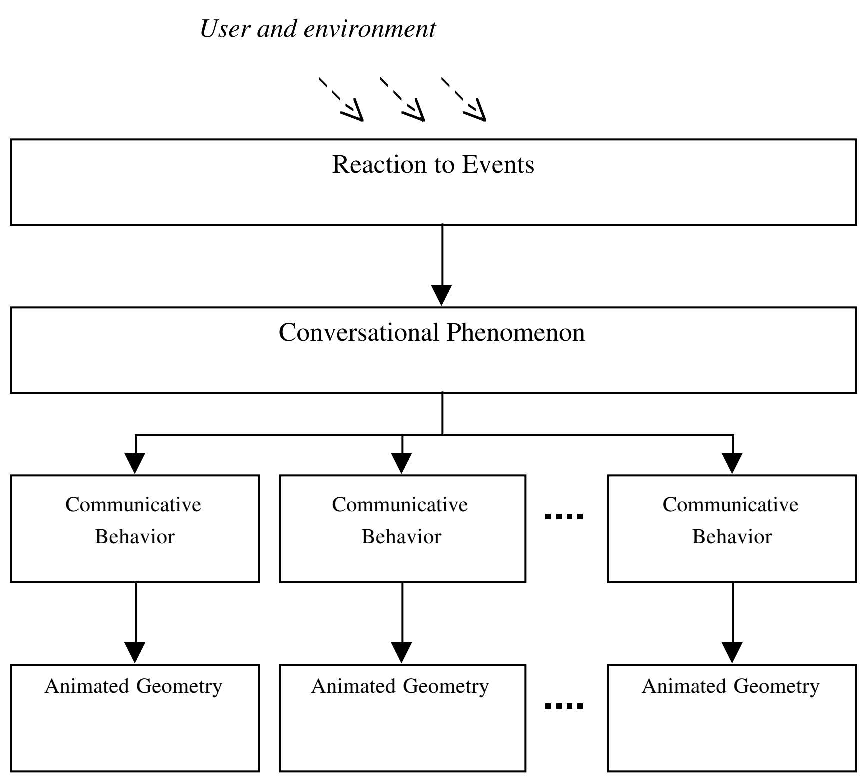

Figure 1. The avatar's behavior control consists of four tiers, where the flow of the execution is from top to bottom.

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 as the environment gets updated. These functions are listed in Table 2. 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 also reflect, as closely as possible, the user's current intentions.

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 2. 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 (see Table 3). This mapping is based on previous work in human communicative behavior, mainly [12] and [15].

Finally, each ***Communicative Behavior*** starts an animation engine that manipulates the corresponding avatar geometry in order change the visual appearance.

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
GiveFloor	Looking, RaiseEyebrows (followed by silence)
BreakAway	GlanceAround
When Leaving:	
Farewell	Looking, HeadNod, Wave

Table 3. The mapping from Conversational Phenomena to visible Behaviors

6.3 Sample Interaction

6.3.1 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.

6.3.2 No interest
User A is scouting out the scene, seeking out someone interested in chatting. After awhile A spots a lone figure that is apparently not occupied. A clicks on the other avatar,

choosing a potential conversational partner (see 6.1). 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 6.1). The automated sequence of glances is shown in Figure 2.

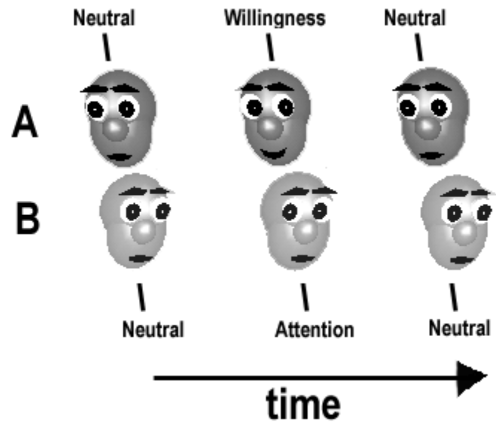

Figure 2. The sequence of glances when user A clicks on avatar B to express willingness to chat while user B is not available**.**

6.3.3 Partner found

User A continues to scout 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 3.

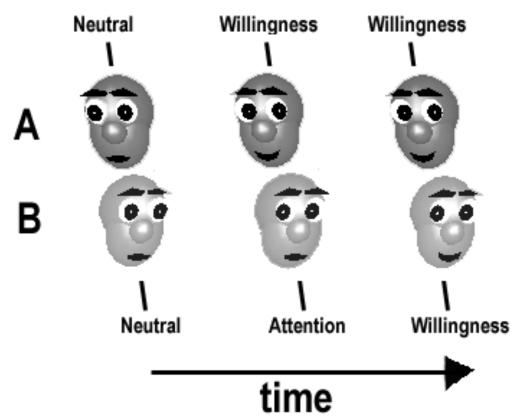

Figure 3. The sequence of glances when user A clicks on avatar B to express willingness to chat and user B is available.

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 4 shows the sequence of salutations.

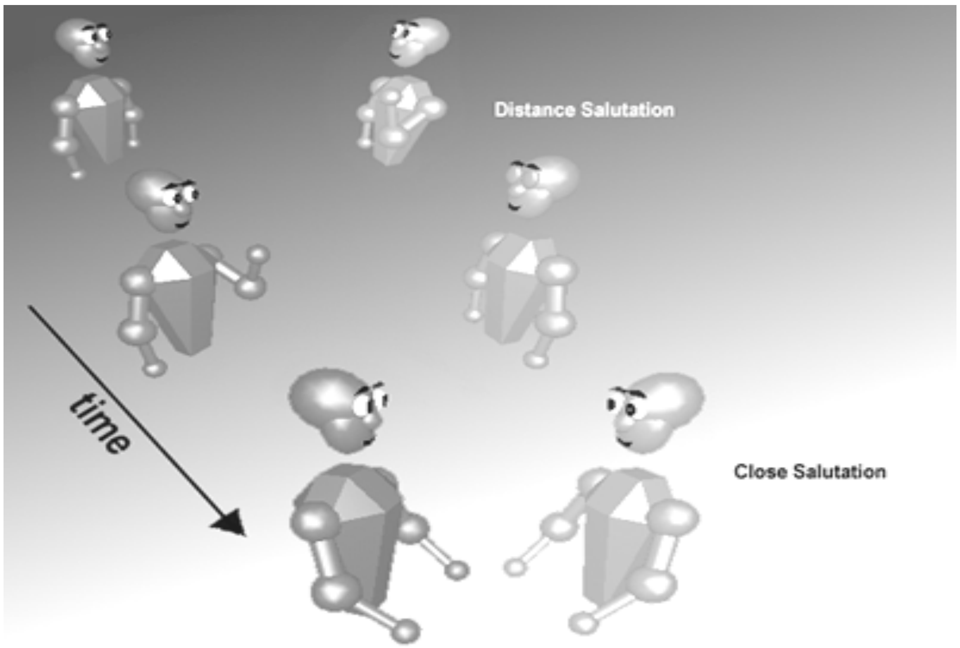

Figure 4. Avatars A and B exchange Distance Salutations when the system registers them as conversational partners. When they get within a conversational range, Close Salutations are exchanged.

6.3.4 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. An example of an animated utterance can be seen in Figure 5.

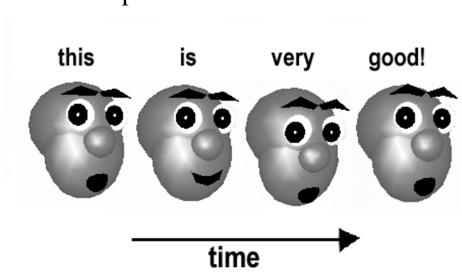

Figure 5. Some words are accompanied with a special facial expression. Here "very" is being emphasized with a nod. The exclamation mark elicits raised eyebrows at the end of the utterance.

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 6.1). This makes A's avatar divert its gaze while reciting the words as shown in Figure 6. 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.

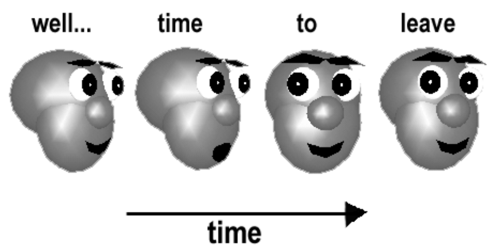

Figure 6. When the user marks a sentence as a Break Away utterance, the avatar displays diverted gaze while reciting the words to give subtle cues to the conversational partner.

7. Evaluation

BodyChat presents a new approach that takes avatars from being a mere visual gimmick to being an integral part of a conversation, from allowing sheer and mere co-presence to allowing embodied conversation. The interaction between user choices and autonomous communicative behaviors allows the user to concentrate on high level control and locomotion, while depending on the avatar to convey the communicative signals that represent the user's communicative intentions.

Regarding the approach in general, however, a few potential limitations are apparent. The first concerns the fact 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 devoid of many real world nuances. This raises concerns about the system's capability to seem like natural human communication.

The second concern relates to the balance between autonomy and direct user control. In the wider field of research on agents, those who promote direct manipulation claim that users have a strong desire to be in control, and to gain mastery over the system (e.g. [19]). Those who promote autonomy, on the other hand, treat the system like any other agent of one's intentions, who will carry out one's instructions to the best of his/her talent (e.g. [18]). In the current case, if the avatar does not accurately reflect the user's intentions, reliability is undermined and the user is left in an uncomfortable skeptical state.

Finally, although the trend in distributed virtual environments is to animate avatars in some way, it is possible that any kind of animation detracts from the communicative power of the system. That is, perhaps avatars *should* simply represent sheer and mere co-presence.

7.1 Methodology

In order to respond to these concerns, we carried out an evaluation of BodyChat that concentrated on exactly these dimensions. We solicited subjects to use the BodyChat system (from within and outside MIT) who had previously used text-based chat, but had never experienced graphical chat. They were told that their goal was to meet the other users in the environment, and to find out as much about each other user as they could. They were told that we wanted them to at least interact briefly with an experimenter in the next room, but that after that they could quit at any time.

In the distributed virtual environment were 4 other users, each a different color (all played by one of the experimenters). Each was scripted with a different personality and life history, except for Avatar #3, which did not respond to attempts to enter into conversation (in order to test the *available* function of the system). Each avatar lightly mirrored the subject's conversation, using similar verbal and non-verbal devices so that users would be exposed to communication of the kind they produced.

In fact, subjects interacted with one of 3 different versions of BodyChat (8 subjects per version). All versions were identical except for the following differences:

- In the **Autonomous** version the avatars had the capabilities described above, with communicative behaviors generated as a function of user intention and in response to other avatars.
- In the **Manual** version the avatars were capable of exactly the same behaviors as those in the Autonomous Condition, except that the behaviors were generated by choosing them from a pull-down menu. The menu selection was listed as follows: nod head, toss head, shake head, wave, glance around, glance away, smile (toggle), raise eyebrows (toggle).
- In the **Both** version, avatars generated autonomous communicative behaviors, **and** users could generate additional behaviors by way of the pull-down menu (which was identical to the menu in the Manual Condition).

In addition, there was a condition without any possibility of animation, which will be compared to the autonomous condition separately:

- In the **None** version the avatars were capable of navigating around the space, but could not gesture, use their face, or produce any other communicative behaviors (like the other three conditions, however, their sternums did move in and out as if they were breathing).

Subjects were taught how to use the system by way of a crib sheet listing the behaviors that their avatar was capable of, and then through a short interaction with an experimenter, who ensured that they were using all of the functionality of the system. They were told that the avatars of the other people in the space were identical to their own, with the same functionality.

At the end of 45 minutes, subjects were told that we needed to go on to the next subject, and they were asked to fill out a questionnaire evaluating

- The overall nature of the experience (with such items as *boring*, *difficult*, *engaging*, *intuitive*, *warm*, *lifelike*).[1]
- The naturalness of the behaviors, and communicative power of the system (with such questions as "how natural did the avatar behavior seem", "in general, how well did

this system allow you to communicate", "how successful was the system at supporting rich conversation").

- Users' control over the system (e.g.. "how much control did you have over the conversation")
- The performance of each of the other avatars encountered (e.g.. "how good was the other person at using the avatar", "how expressive was this person").

In addition, subjects were requested to list all of the facts that they had acquired about each of the other putative users that they had conversed with. This item served as a measure of task performance .

All questionnaire items were measured using 10 point Likert scales.

8. Results

The first set of analyses concentrated on user's perceptions of the naturalness of the autonomous avatars, their judgements of the communicative power of the avatars, and their judgements of their control over the autonomous vs. manual vs. both versions of the system. A final set of analyses compared the avatar without any embodied communicative behaviors to those that were capable of generating those behaviors.

First, some general descriptive statistics: conversations in the autonomous condition were significantly *longer* (a mean of 1111 seconds) than those in the manual (mean of 671 seconds) or both (mean of 879 seconds) conditions. This can be taken as an index of the interest that subjects had in pursuing conversational interaction with people, when they were using the autonomous system. Moreover, subjects in the autonomous condition remembered more facts about the people they interacted with (a mean of 5.2) than did subjects in the manual condition (mean of 3.8) or the both condition (mean of 4.5 facts). This can be taken as an index of how engaged subjects were in the conversation, perhaps because their attention was not divided between participating in the conversation and controlling the avatar.

Next, to address the naturalness of the system, users were asked (a) how natural did the avatar behavior seem, and (b) how natural did the interaction seem. An ANOVA, and subsequent post-hoc t-tests, considering these two questions together showed that users of the autonomous system judged it to be more *natural* than users of either of the other two conditions ($\underline{F}=5.10(2,21)$; $p<.02$)[2]. (see Figure 7).

In order to address the communicative power of the system, users were asked (a) how well do you feel you were able to understand the people you met; (b) how well do you feel you were able to express yourself with the people you met; (c) how well do you think the people you met understood you and understood what you meant to communicate; (d) how well do you feel other users were able to express themselves with you. An ANOVA and subsequent post-hoc t-test of these aggregate data revealed that users of the autonomous system judged it to be more *expressive* than users of the manual system, but not significantly more expressive than the condition with both autonomous and menu-driven communicative behaviors ($\underline{F}=5.94(2,21)$; $p<.01$) (see Figure 7). This indicates that expressivity increases when autonomously generated communicative behaviors are available, whether or not menu-driven behaviors are also available. It should be noted, however, that users of the avatars with both kinds of behaviors found their experience to be more tedious than users of the autonomous system (t=1.9, p<.05, one-tailed).

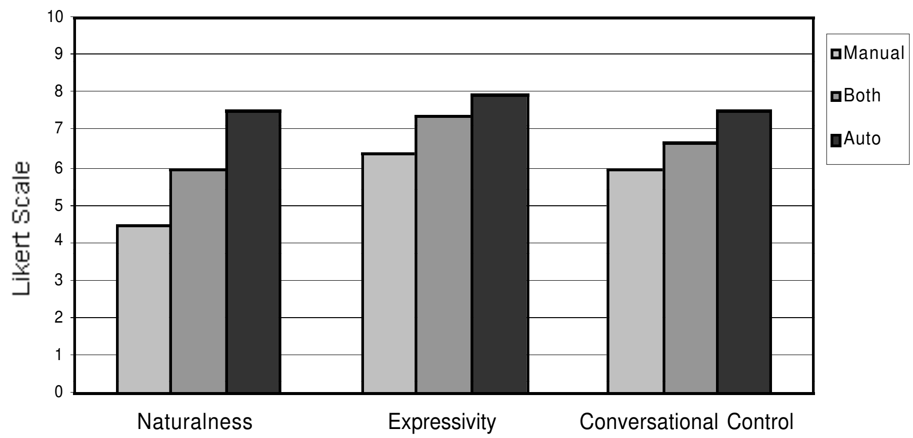

Figure 7. Effects of manual vs. autonomous vs. both avatar interface on perception of expressivity, conversational control, and naturalness**.**

In order to address the issue of control vs. autonomy, users were asked (a) how much control did you have over the conversation, and (b) how much control do you think the other users had over the conversation. An ANOVA, and subsequent post-hoc t-tests, considering these two questions together revealed that users of the autonomous system considered the conversation more under users' *control* than did users of the manual or the both systems (F=6.33(2,21); p<.01) (see Figure 7). This somewhat paradoxical effect may indicate that, when nonverbal communicative behaviors were generated automatically, users were freed from having to worry about them, and they consequently felt that they could better control the course of the conversation. This is, of course, the whole point of autonomously generating embodied conversational actions in avatars.

Recall that, in the autonomous system, communicative behaviors were generated as a function of the user's word or punctuation choices (e.g. an emphasis head nod generated when the user types "very" or an eyebrow raise when the user ends a sentence with '!'), as well as from proximity to other avatars and so forth. One might therefore argue that the autonomous system is really just a thinly veiled manual system, whereby behaviors are chosen via keywords rather than menu choices. In order to test this possibility, the use of these keywords was examined across conditions. In fact, however, neither by category of keyword (*negations* such as "no", "nope", "nah"; *demonstratives* such as "this" "there"; *greetings* such as "hi", "hey", "howdy", and so forth) nor by total keyword was there a significant difference among the conditions. In other words, subjects used the language in the same way across conditions (with a seemingly normal distribution), and did not use the keywords as some kind of control language.

Finally, we turn to the comparison of the autonomous condition with the none condition. Some designers have suggested that simple co-presence exhausts the value of an avatar. On this view, all other features might simply distract from the user's communication. Our final comparison, between the autonomous and the none condition sheds light on this issue.

While the autonomous condition was judged as significantly more natural than the none condition ($t=3.04$; $p<.005$, one-tailed), there was no significant difference between the two conditions on judgments of conversational expressiveness and judgments of users' control over the conversation. Moreover, subjects spent more mean time conversing in the none condition than in the manual condition (mean of 988 seconds), although less time than was spent in the autonomous condition. These findings support the conclusion drawn above, that users may derive a sense of conversational control and pleasure in the conversation primarily from their textual contribution and may feel distracted by having to animate their avatars' communicative behaviors.

But why should users feel that conversation in the chat space is equally expressive whether using an avatar with no embodied communicative behaviors or an avatar with autonomous embodied communicative behaviors? In our view, again, when users are not concentrating on the menu, they are able to concentrate on the conversation itself, whether it is instantiated in text only or in text + animated communicative behaviors. But this leads to another question. Why shouldn't we build systems where avatars simply represent co-presence?. Although this is not the primary issue that we set out to address (since designers of avatar systems seem resolved to add behaviors over-and-above simple presence), this is an issue that deserves more discussion. Our own position is two-fold. First of all, the whole goal of graphical chat systems is to make computer-mediated interactions among humans more *natural*. And, recall, users did judge the autonomous system as more natural than the system without communicative behaviors. Secondly, we believe that one of the contexts in which autonomous communicative behaviors will have the greatest effect in avatar systems, is in multi-party conversations. Recall the example given above where users are unsure of who is speaking to whom, and whether particular users are available for conversation. In a two-party conversation, such as was tested here, these problems do not arise. We believe this is worthy of further investigation.

Finally, there are two caveats to keep in mind in interpreting the results of the evaluation. First, we did not *directly* test whether particular users preferred one system over the other, as it was not the same users who engaged with the different systems. Whereas a direct comparison might be preferable, it is experimentally difficult since a pilot showed that short interactions (under 45 minutes) with the system left users unable to evaluate; and thus a direct comparison would require two periods of 45 minutes each, plus two 20 minute questionnaires, which is a long enough period of time to introduce unwanted fatigue effects into the results, and to introduce an unwanted comfort factor whereby subjects may judge the last system used as the best, simply because they have had time to become comfortable with it. Second, although the manual version of the system was very similar to systems on the market, it is not identical, and thus it is possible that users might prefer a manual system to the autonomous one if the manual system were different in some way. Note, however, that even in the current evaluation, users judged the manual system quite highly (a mean of 7.6 out of 10 on the scale of how *fun* the experience was, for example) – it's simply that they judged the autonomous system even more highly.

9. Conclusions

This paper has introduced a novel approach to the design and implementation of avatars, drawing from literature in context analysis, discourse theory, and autonomous communicating agents. 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, we relied on previous research on multi-modal communication among humans. We used that research to develop BodyChat, a system that employs those findings in the automation of communicative behaviors in avatars.

The system as it stands is a first pass at a repertoire of communicative behaviors, beginning with the most essential cues for initiating a conversation. It is important to continue adding to the model of conversational phenomena, both drawing from psycholinguistic and ethno-methodological literature and, perhaps more interestingly, through real world empirical studies conducted with this domain in mind. In this vein, we are currently examining multi-party conversation, in particular concentrating on issues of floor management. We are also working on better physical models of the arms and hands and, in line with these developments, on how to autonomously generate hand gesture from text or speech.

In terms of the ongoing discussion in the field about direct manipulation vs. autonomous behavior, BodyChat introduces a balance that supports both views. Those that promote direct manipulation claim that users should be allowed to control -- and master -- an interface. However, they also emphasize that the interface actions should stay close to the high-level task domain to minimize the need for a mental decompositon of commands [23]. Therefore, it makes perfect sense that a system that allows users to carry on a conversation spare them the burden of micro-management. At the same time, the execution of the micro-steps involved in, for example, a greeting, depends on the virtual setting and current context, and is therefore not fully deterministic. That type of execution fits the job description of an agent, whose job is to treat the user's actions as instructions whose meaning differs based on the context in which they occur, and the context in which they must be accomplished.

Because of the richness of involuntary behavior in a social situation, relying only on explicit user control will not exploit the function of embodiment in the construction of animated avatars. Regarding an avatar as a personal conversational agent that together with the user is capable of naturally initiating and sustaining a conversation provides a valuable perspective, contributing both to research on avatars, and to a broader theory of embodied conversational agents.

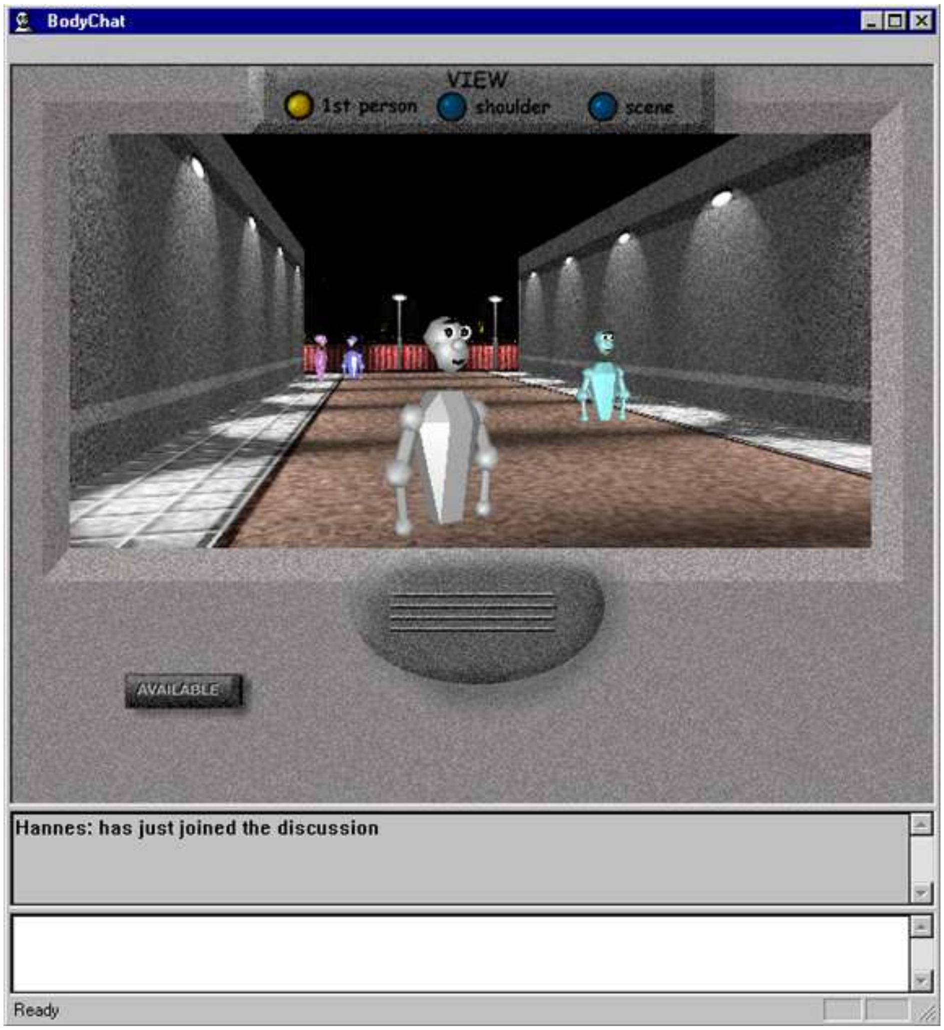

Figure 8. Looking at another user's avatar in BodyChat

Acknowledgements

Thanks to the members of the Gesture and Narrative Language group and the Media Lab community for valuable discussions. This research was supported in part by the National Science Foundation (STIMULATE award 9618939) and by the Media Lab *Digital Life* and *Things That Think* consortia.

Notes

[1] Items adapted from [21]. Questionnaire available on request from the authors.

[2] Tables of means and additional statistics available on request from the authors.

References

1. Anderson, D.B., Barrus, J.W., Brogan, D., Casey M., McKeown, S., Sterns, I., Waters, R., Yerazunis, W. "Diamond Park and Spline: A Social Virtual Reality System with 3D Animation, Spoken Interaction, and Runtime Modifiability." Technical Report at MERL, Cambridge, 1996.
2. Argyle, M., Cook, M. *Gaze and Mutual Gaze*. Cambridge University Press, 1976.
3. Argyle, M., Ingham, R., Alkema, F., McCallin, M. "The Different Functions of Gaze." *Semiotica*, 1973.
4. Bates, J., Loyall, A.B., Reilley, W.S. "Broad Agents." *SIGART Bulletin*, **4** (2), 1991.
5. Benford, S., Bowers, J., Fahlen, L.E., Greenhalgh, C., Snowdon, D. "User Embodiment in Collaborative Virtual Environments." *Proceedings of CHI'95*, 242-249.
6. Blumberg, B. M., Galyean, T. A. "Multi-Level Direction of Autonomous Creatures for Real-Time Virtual Environments." *Proceedings of SIGGRAPH '95*.
7. Cary, M. S. "The Role of Gaze in the Initiation of Conversation." *Social Psychology*, **41**(3), 1978.
8. Cassell, J. "Embodied Conversation: Integrating Face and Gesture into Automatic Spoken Dialogue Systems." In Luperfoy (ed.) Spoken Dialogue Systems. Cambridge, MA: MIT Press, 1999.
9. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M. "Animated Conversation: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents." *Proceedings of SIGGRAPH '94*, 1994.
10. Cassell, J., Thórisson, K. "The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents". *Journal of Applied Artificial Intelligence*, in press.
11. Cassell, J., Torres, O. and S. Prevost. "Turn taking vs. Discourse Structure: how best to model multimodal conversation." In Wilks (ed.) Machine Conversations. The Hague: Kluwer, 1998.
12. Chovil, N. Discourse-Oriented "Facial Displays in Conversation." *Research on Language and Social Interaction*, **25**, 163-194, 1992.
13. Donath, J. "The Illustrated Conversation." *Multimedia Tools and Applications*, **1**, 79-88, 1995.
14. Goodwin, C. "Gestures as a Resource for the Organization of Mutual Orientation." *Semiotica*, **62**(1/2), 1986.
15. Kendon, A. *Conducting Interaction: Patterns of behavior in focused encounters*. Cambridge University Press. NY, 1990.

16. Kendon, A. “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. NY, 1990.

17. Kurlander, D., Skelly, T., Salesin, D. “Comic Chat.” *Proceedings of SIGGRAPH ’96*, 1996.

18. Lieberman, H., Maulsby, D. “Instructible Agents: Software That Just Keeps Getting Better.” *IBM Systems Journal,* **35**, Nos. 3 & 4, 1996.

19. Maes, P., Schneiderman, B., "Direct Manipulation vs. Interface Agents: a Debate." *Interactions*, **4** Number 6, ACM Press, 1997.

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

21. Moon, Y., Nass, C. I. “How "real" are computer personalities? Psychological responses to personality types in human-computer interaction.” *Communication Research*, **23**(6), 651-674, 1996.

22. Perlin, K., Goldberg, A. “Improv: A System for Scripting Interactive Actors in Virtual Worlds.” *SIGGRAPH 1996* Course Notes #25.

23. Prevost, S. “Modeling Contrast in the Generation and Synthesis of Spoken Language.” *Proceedings of ICSLP ‘96*.

24. Schegloff, E. “Sequencing in Conversational Openings.” *American Anthropologist*, **70**, 1075-1095, 1968.

25. Schegloff, E., Sacks, H. “Opening up closings.” *Semiotica*, **8**, 289-327, 1973.

26. Thórisson, K. R. “Gandalf: An Embodied Humanoid Capable of Real-Time Multimodal Dialogue with People.” *Proceedings of Agents'97*, 536-537.

27. Wavish, P., Connah, D. “Virtual actors that can perform scripts and improvise roles.” *Proceedings of Agents'97*, 317-322.