Avatar Augmented Online Conversation

by

Hannes Högni Vilhjálmsson

Bachelor of Science in Computer Science
University of Iceland
Reykjavik, Iceland
1994

Master of Science in Media Arts and Sciences
Massachusetts Institute of Technology
Cambridge, MA
1997

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

at the Massachusetts Institute of Technology
June 2003

©Massachusetts Institute of Technology 2003
All Rights Reserved
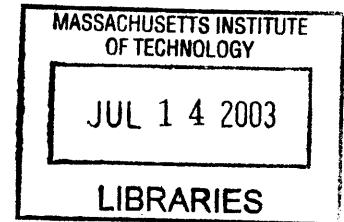
Signature of Author

_____
Program in Media Arts and Sciences
May 2, 2003

Certified by

_____
Justine Cassell
Associate Professor
Thesis Supervisor

Accepted by

_____
Andrew B. Lippman
Chairperson
Departmental Committee on Graduate Students

1

ROTCH

Avatar Augmented Online Conversation

by

Hannes Högni Vilhjálmsson

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning, on May 2, 2003
in Partial Fulfillment of the Requirements of the
Degree of Doctor of Philosophy
at the Massachusetts Institute of Technology

ABSTRACT

One of the most important roles played by technology is connecting people and mediating their communication with one another. Building technology that mediates conversation presents a number of challenging research and design questions. Apart from the fundamental issue of what exactly gets mediated, two of the more crucial questions are how the person being mediated interacts with the mediating layer and how the receiving person experiences the mediation. This thesis is concerned with both of these questions and proposes a theoretical framework of mediated conversation by means of automated avatars. This new approach relies on a model of face-to-face conversation, and derives an architecture for implementing these features through automation. First the thesis describes the process of face-to-face conversation and what nonverbal behaviors contribute to its success. It then presents a theoretical framework that explains how a text message can be automatically analyzed in terms of its communicative function based on discourse context, and how behaviors, shown to support those same functions in face-to-face conversation, can then be automatically performed by a graphical avatar in synchrony with the message delivery. An architecture, Spark, built on this framework demonstrates the approach in an actual system design that introduces the concept of a message transformation pipeline, abstracting function from behavior, and the concept of an avatar agent, responsible for coordinated delivery and continuous maintenance of the communication channel. A derived application, MapChat, is an online collaboration system where users represented by avatars in a shared virtual environment can chat and manipulate an interactive map while their avatars generate face-to-face behaviors. A study evaluating the strength of the approach compares groups collaborating on a route-planning task using MapChat with and without the animated avatars. The results show that while task outcome was equally good for both groups, the group using these avatars felt that the task was significantly less difficult, and the feeling of efficiency and consensus were significantly stronger. An analysis of the conversation transcripts shows a significant improvement of the overall conversational process and significantly fewer messages spent on channel maintenance in the avatar groups. The avatars also significantly improved the users' perception of each others' effort. Finally, MapChat with avatars was found to be significantly more personal, enjoyable, and easier to use. The ramifications of these findings with respect to mediating conversation are discussed.

Thesis Supervisor: Justine Cassell

Title: Associate Professor of Media Arts and Sciences

Avatar Augmented Online Conversation

by

Hannes Högni Vilhjálmsson

Doctoral Dissertation Committee

Thesis Advisor

Justine Cassell
Associate Professor of Media Arts and Sciences
AT&T Career Development Professor
Massachusetts Institute of Technology

Thesis Reader

Cynthia Breazeal
Assistant Professor of Media Arts and Sciences
LG Group Career Development Professor
Massachusetts Institute of Technology

Thesis Reader

Amy Bruckman
Assistant Professor, College of Computing
Georgia Institute of Technology

Dedicated to my parents and grandparents
who I needed to see a lot more often

Acknowledgements

13

16

1 Introduction

1.1 The focus of this thesis

One of the most important roles played by technology is connecting people and mediating their communication with one another. Remote conversations, were inconceivable before the introduction of the telegraph in 1837, but are now routinely conducted with devices ranging from two-way pagers to videoconference systems.

Building technology that mediates conversation presents a number of challenging research and design questions. Apart from the fundamental issue of what exactly gets mediated, two of the more crucial questions are how the person being mediated interacts with the mediating layer and how the receiving person experiences the mediation (see Figure 1). This thesis is concerned with both of these questions and proposes a theoretical framework of mediated conversation by means of automated avatars.
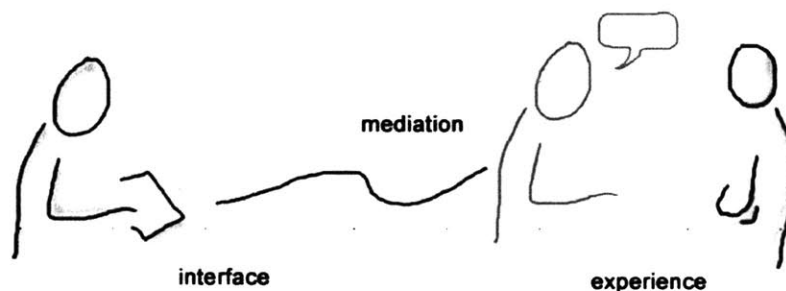
Figure 1: The person being mediated interfaces with the mediation layer, that in turn produces a communication experience for the recipient

The fundamental assumption is made that the process of face-to-face conversation represents the ideal, where the participants are free of any interfacing overhead and the experience is as rich as it gets. The goal of this thesis is to show how a model of face-to-face conversation can be used as the basis for the proposed framework.

1.2 The process of conversation

Establishing and maintaining a channel of communication with other human beings face-to-face is an ability that has evolved since the dawn of humanity. The coordination of a conversation is not merely a person's spoken transmission of thought, but rather it is a dynamic process involving exchanges of gesture, gaze, facial expression and body posture, carefully coordinating the mutual understanding about what is being shared and how to proceed with the conduct. The process is woven into the fabric of discourse context. This context is both what the participants

bring with them to the conversation and what accumulates during the conversation itself.

When communicating, we expect everyone to adhere to a shared protocol. The protocol allows us to interpret everyone's words and actions, in the current context. While ensuring that the conversation unfolds in an efficient and orderly fashion, this elaborate process does not ask for much conscious effort beyond what is required to reflect on the topic at hand.

Figure 2: Nonverbal behaviors, such as gesture and gaze, play an important role in coordinating face-to-face conversation.

Conversation, and the processes that contribute to its successful execution, have been studied extensively in the fields of discourse analysis, conversation analysis, sociolinguistics, social psychology and computational linguistics. While the number and roles of these processes are both many and varied, four major categories resurface throughout the literature, outlining some of the most crucial elements of conversation. These categories of conversation processes are "awareness and engagement management," "interaction management," "discourse structure management" and "information management."

Awareness and Engagement Management

Potential participants need to be aware of each other's presence before a conversation can start. Once awareness has been established, they typically need to negotiate whether to engage in conversation or not. Likewise, when a participant wishes to leave a conversation, the intent to depart is announced and the other participants have to agree to it, before

leave is taken. Managing awareness and engagement is the process by which people initiate and break a conversation.

Interaction Management

Once conversation has started, the participants have to maintain an open channel of communication between them. They take turns speaking, so that everyone can be clearly heard, and those listening show speakers signs of attention, to confirm the clarity of the channel. The process of interaction management deals with the control and maintenance of the conversation channel.

Discourse Structure Management

The organization of conversation often takes participants through various topics that in turn can contain sub-topics. Each topic provides a context that contributes to the successful interpretation of what is being said. This hierarchical organization of relevant context is termed discourse structure. Discourse structure management is the process of announcing and negotiating shifts within this structure.

Information Management

At the core of conversation, information sharing is taking place. A speaker's utterance is, in part, meant to update what listeners know. This update often involves the things around us in the world. The way that a speaker presents new information and the way they refer to the relevant entities in the world is part of information management. In order to ensure proper uptake of information by listeners, the speaker may need to look for or request signs of understanding, which also is a part of the management process.

1.3 Online Conversation

Technologies that mediate conversation between two or more people have of course continued to evolve since the telegraph and today networked computers play an essential role.

1.3.1 Characteristics

Computer Mediated Communication or CMC is an extensive field of study that has been rapidly gaining in interest and importance. CMC refers to communication that takes place between human beings via the instrumentality of computers (Herring 1996). While computers have found their way into most of human communication systems, the term usually refers to communication that relies on the classic desktop machine as the terminal device and user interface. There are two broad categories of CMC systems, those that support synchronous communication and those that support asynchronous communication. The former category

21

includes applications such as chat rooms and video-conferencing while the latter includes technologies such as email and newsgroups.

This thesis will concern itself with synchronous CMC (S-CMC). It is a pervasive medium that potentially provides a convenient way to build and maintain social relationships online, while delivering less than it promises when it comes to coordinated activities such as teaching or business meetings. People attribute to it qualities of face-to-face conversation because of how responsive and spontaneous this medium is (Werry 1996; Garcia and Jacobs 1998), but a number of limitations make the medium unsuitable for many of the tasks traditionally solved face-to-face.

1.3.2 Applications

One of the first synchronous CMC systems was a system that allowed the users of ARPANET, precursor of the Internet, to write messages onto each other's consoles. While first conceived as an administrator's tool, the system gained immediate popularity among other users. A derivative of this simple system is still today one of the most widely used synchronous communication tools, in the guise of the instant messenger (Nardi, Whittaker et al. 2000; Herbsled, Atkins et al. 2002). Instant messengers such as AOL Instant Messenger, MSN Messenger or ICQ allow users to transmit a line of text to any of their friends that have registered with the same service. The text appears in a window on their friends' display and the recipient can then decide to immediately reply, in which case a conversation ensues and lines of text are exchanged in rapid succession.

Other forms of real-time textual exchange also emerged, two of the most influential being the Multi-User Domain (MUD), first developed in 1979 and the Internet Relay Chat (IRC), released in 1988. While the messenger systems are most frequently used to carry one-on-one conversations, the MUD and the IRC were originally built to support group conversations. MUDs provide textual chat rooms structured around locations in a fictitious world of shared adventure, and IRC provides thousands of worldwide chat rooms organized around user-selected topics.

The largest traffic that these systems carry may be recreational or casual in nature, but one has to be careful not to dismiss their effect on modern society as insignificant or frivolous since they contribute to social connectedness and community building that has impact beyond the console (Turkle 1995). Synchronous CMC systems have also been widely employed in more structured settings, such as in distance learning and training. They are then often used in conjunction with asynchronous media such as the web and newsgroups, and both relay in-class activities as well as out-of-class discussions.

1.3.3 Limitations

Beyond the perhaps obvious limitation that typing speed imposes on the pace of text based conversation, there are a few major factors that would

seem to make text-based synchronous CMC ill-suited as a replacement for face-to-face. These include overlapping threads, limited turn negotiation mechanisms, no dedicated feedback channels and no way of visually establishing referents or focus of attention. Let's look at each conversational process in turn.

Awareness and Engagement Management

Exchanging brief glances of recognition and negotiating whether to engage in a conversation before a single word has been exchanged is not straightforward in a text based medium. Each CMC system approaches the issue of how conversation is initiated differently. In regular chat rooms nothing is subtle and you simply have to speak up to get someone's (and then usually also everyone's) attention. Sometimes people describe nonverbal actions taken before addressing someone (especially true for MUDs where actions are expected), for example by typing: "Eldark notices the unusual attire of the newcomer." Such actions are still deliberate and therefore expose the initiator just as much as speaking would do.

Instant messengers provide signs that users can place next to their names declaring whether they are available for chatting or not. For the one placing the sign, this allows a certain negotiation before they have to say anything, but other people are forced to break the silence to make contact. Apart from the difficulty in exchanging initial nonverbal signals, people sometimes find it more difficult to actually recognize other participants by their textual nicknames alone than by visual appearance (Vronay, Smith et al. 1999; Smith, Cadiz et al. 2000).

Similar to initiating contact, breaking away from a conversation requires delicate negotiation. The lack of being able to exchange subtle signs of interest to leave a conversation (for example by briefly diverting attention elsewhere) forces online participants to be more explicit and abrupt about this intent. This may lead to people essentially getting stuck because they shy away from making the action too explicit and one-sided.

Discourse Structure Management

For group conversations, such as those that occur on IRC, it is often hard to keep track of conversation topics because all contributions made in the same chat room or channel end up in a single scrolling window, temporally ordered rather than topically or by what is a response to what. In other words, each contribution is essentially an announcement to all present and it is the receiver's responsibility to understand who the announcement is meant for and how it may or may not fit into any of the ongoing conversation topics. This often leads to confusion among participants, especially those that are not extensively familiar with the medium (Lindeman, Kent et al. 1995; Werry 1996; Cherny 1999; Vronay,

Smith et al. 1999; Smith, Cadiz et al. 2000; Barnett 2001; Suthers 2001; Russell and Halcomb 2002).

Interaction Management

Taking turns sometimes proves to be a non-trivial matter in CMC. More than one participant can simultaneously construct a reply to the last contribution; therefore, some confusion can occur as all the replies get displayed in the sequence they are finished while each is meant to immediately follow the original statement. No turn negotiation is possible - either you grab it or leave it; either you transmit or you don't. In order to guide your contribution to particular participants, you must explicitly name them in your message, which may be fine if only one or two need to be named, but to address a certain sub-group, this starts to become awkward (McCarthy 1993; Werry 1996; Russell and Halcomb 2002).

It is often difficult to have a sense of who is actively participating in the current conversation because participation status cannot be immediately gleaned from the list of those sharing the chat room. It is not clear who is taking on a speaker role by typing a message, who is listening and who is only logged in but not attending to the conversation at all (Garcia and Jacobs 1998; Vronay, Smith et al. 1999; Smith, Cadiz et al. 2000).

An important limitation is the lack of subtle listener backchannel. Since there is only one modality available, i.e. the text channel, both the contributor's content and the recipient's spontaneous feedback have to occupy the same channel. You essentially have to take the floor in order to produce even the simplest cues of feedback, such as "mhm." In fact, if you are not actively messaging, you are practically invisible to the conversation (Smith, Cadiz et al. 2000; Donath 2002). While expert users produce such back channel turns with great frequency (Cherny 1999), non-experts use them less frequently than they do in spoken interaction and end up having a less efficient interaction (Oviatt and Cohen 1991).

Information Management

Many CMC systems provide a rich context and resources for the interaction, such as learning materials or objects and documents that are being collaborated on. It then becomes important how these resources get integrated into the conversational flow. For example, MUDs describe a virtual setting textually, naming any objects that can be handled by those sharing the room. The objects can be manipulated simply by typing in actions that involve the object. For instance, one could type "pick up folder" and then "read folder" to access a document. However, just like contributions to the conversation, all actions have to be explicit and discrete. Users have to directly address an object one at a time. This limits how seamlessly the conversation and object manipulation can be woven together. Some collaboration systems include a shared whiteboard, placed directly above the text chat window. While this allows

collaborators to use illustration and visual annotation, the link between the text and the board is not seamless and miscommunication is frequent when attention is being paid to the wrong window (Damianos, Drury et al. 2000).

1.3.4 Adaptation

Veteran users of synchronized CMC systems have adapted to the medium and created a number of textual conventions that try to overcome these limitations (Cherny 1995; Werry 1996). These conventions include inserting labels to refer to other participants, using punctuation and capitalization to simulate intonation, using abbreviations to reduce response time, and special codes to describe nonverbal actions and reactions. The result is a highly saturated, multi-layered, fast flowing text, which may evoke the feeling of a near face-to-face in certain conversations for trained users, but appears as almost random garble to those that have no prior experience. New users have to learn to recognize and skillfully wield the conventions of this new online language in order to gain full participation status. However, even for veterans, the lack of visual cues can cause frustration when the conversation relies on manipulation of and frequent references to a shared collaborative environment.

1.4 New Approach

This thesis presents a new approach to synchronized CMC that aims to provide the experience of face-to-face interaction without requiring users to learn new conventions or burden them with a complex interface and unfamiliar controls. The approach is based on graphical chat rooms where users are represented as graphical figures termed *avatars*. It departs from traditional graphical chat rooms by proposing that the avatars should not merely be pictures or puppets that passively represent their users, but that they should be designed as an integral part of the mediation layer. There they can help overcome deficiencies in the communication channel by monitoring the conversation and automatically supply missing nonverbal cues supporting the processes described in 1.1.

The approach involves automatically interpreting and encoding what the user means to communicate, drawing from an analysis of the text message and the discourse context. It then uses the model of face-to-face conversation to suggest how the delivery can be augmented through the appropriate real-time coordination of gaze, gesture, posture, facial expression, and head movements in an animated avatar.

Because the avatars of all participants occupy the same virtual environment, all the nonverbal cues are rendered with a shared point of reference, so that the approach not only improves textual communication but also addresses a fundamental limitation of Video Mediated Communication, which is the lack of a fully shared space.

25

1.5 Contributions and Organization of Thesis

First the background about human face-to-face conversation is reviewed in chapter 2 along with a review of computer mediated communication and computational models of conversation. Then the **theoretical framework** describing the augmentation of online conversation based on a model of face-to-face conversation is introduced in chapter 3. This model lists the essential processes that need to be supported and how nonverbal behavior could be generated to fill that role. The theory is taken to a practical level through the engineering of an online conversation **system architecture** called Spark described in chapter 4 and then an actual implementation of this architecture in the form of a general infrastructure and a programming interface is presented in chapter 5. A working **application** for online collaborative route planning is demonstrated in chapter 6 and the implementation and approach evaluated in chapter 7. Possible follow-up studies, application considerations and interesting issues are discussed in Chapter 8. Finally future work and conclusions in chapters 9 and 10 place the approach in a broader perspective, reflecting on general limitations as well as on the kinds of augmented communication this work makes possible.

2 Related Work

2.1 Face-to-face Conversation

The goal of this thesis is to present a theoretical framework of how a system for augmenting online conversation can be built from a model of face-to-face conversation. Here conversation refers to any type of real-time conversation, ranging from casual chatting to conversations that occur when groups solve a particular task together. The model therefore has to identify the general processes that characterize and support all face-to-face conversations. Once the processes have been identified, the nonverbal behaviors that support them have to be described so that they can be replicated online by automated avatars. As mentioned in the introduction, the literature on face-to-face conversation has identified four fundamental categories of conversational processes. Each will be reviewed in this section.

2.1.1 Awareness and Engagement Management

Processes

(Goffman 1963) describes a theory where communicative behavior relates to either unfocused interaction or focused interaction. The former refers to a state where participants are aware of each other, but are not committed to any interaction beyond the management of sheer and mere co-presence. In the latter state, participants openly cooperate to sustain a focus of attention. A crucial process identified by this theory is the transition from unfocused to focused interaction. According to Goffman's theory, this transition can happen in two ways depending on the situation. A very smooth, and a relatively automatic, transition happens if the participants are already acquainted and/or their roles in the interaction are well defined. However, if they are unacquainted and/or if their roles have not yet been defined, they need a reason to engage and therefore a process of interest negotiation is called for. In support of Goffman's theory, (Cary 1978) has observed that a stranger who receives a signal of interest is far more likely to engage in a conversation than a stranger who does not receive a signal beyond mere awareness. Goffman's theory has been adopted by computer supported collaborative workspaces such as (Dourish and Bly 1992) and various derivatives.

Behaviors

Establishing and maintaining participation in a conversation is largely dependent on appropriate body orientation and gaze direction. To engage people in a conversation, one has to show them visual attention beyond what would be considered a passing glance according to (Goffman 1963; Cary 1978). Subject to the other people's reciprocal action and

acceptance, salutations are exchanged. Finally it is possible to move closer and everyone re-orients themselves such that they have clear access to each other's field of attention (Kendon 1990).

2.1.2 Interaction Management

Process

Once a focused interaction is underway, participants have to coordinate a successful exchange. Two processes are central to this coordination: turn-taking and feedback. The former is the way by which participants ensure everyone is not speaking at the same time, and can thus be clearly heard. Turns are requested, taken, held and given using various signals, often exchanged in parallel with speech over nonverbal channels such as gaze, intonation and gesture (Duncan 1974; Goodwin 1981). When there are more than two people interacting, it is not enough to simply indicate an end of turn, but the next speaker also needs to be chosen (Goffman 1983; Rosenfeld 1987). When turn taking is hindered by limiting available channels, chaos may ensue. For example, voice conferencing with multiple participants, where little can be exchanged over and beyond the speech channel, has often been reported as troublesome (Vertegaal 1999).

Furthermore, the speaker's ability to formulate efficient messages is critically dependent on dynamic listener attentive feedback. For example, (Krauss and Fussell 1991) have shown that in the absence of backchannel feedback, speakers progress more slowly from using long descriptive names to using compact referring expressions.

Behaviors

During and between turns, listener and speaker exchange many kinds of nonverbal signals that act as conversation regulators (Rosenfeld 1987). Although taking a turn basically involves starting to speak, some nonverbal behavior usually coincides with that activity. The most common behavior, whose likelihood increases with the increasing complexity and length of the utterance about to be delivered, is looking away from the listener (Argyle and Cook 1976). Hands are often being raised into gesture space as well, in preparation for gesticulation (Kendon 1990).

These nonverbal signs of looking away and raising the hands may be employed by a listener at any time to indicate to the current speaker that they wish to receive the turn. According to (Duncan 1974), a speaker gives the turn by stopping the utterance, looking at the listener (if there is more than one listener, the speaker looks at the listener whose turn it is to speak next (Kendon 1990)), and resting the hands. Sometimes the hands turn over with open palms towards the selected next speaker as they are brought down to rest (Bavelas, Chovil et al. 1995).

During turns, some exchange of signals usually occurs around junctures between clauses, where each clause often corresponds to an intonational phrase. Lets use the term *gaps* for these within-turn junctures. At gaps, speakers often request feedback from listeners. The basic feedback request typically involves looking at the listener and raising eyebrows (Chovil 1991). To request a more involved feedback, this behavior can be supplemented with pointing the head towards the listener or conducting a series of low amplitude head nods prior to the gap, and raising the head at the juncture (Rosenfeld 1987). Where gaps occur because the speaker is hesitant and is searching for words, the speaker is often seen to either elicit the listener's help by looking at the listener while producing some sort of "cranking" gesture or to avoid listener involvement by looking off to the side (Bavelas, Chovil et al. 1995).

Listener feedback can take on a variety of forms depending on the desired impact, and mostly occurs around the end of a speaker's turn or around gaps (Chovil 1991). Brief assertion of attention, with or without a speaker's explicit feedback request, may be given by the dropping of the eyelids and or a slight head nod towards the speaker. A stronger attention cue, typically given after a speaker's request for feedback, may involve a slight leaning and a look towards the speaker along with a short verbal response or a laugh. A more pronounced feedback request seems to increase the likelihood of nodding on top of that. Another style of attention feedback has been observed, that does not differ functionally, but has a more serious tone, involves raising the eyebrows and closing the eyes, while pressing the lips together with the corners of the mouth turned down.

2.1.3 Discourse Structure Management

Process

Conversations go through different phases, at the very least an entry, body and an exit (Schegloff and Sacks 1973; Clark 1996). Furthermore, the body itself may take participants through various topics, that again can divide into sub-topics. This organization of contributions into a hierarchy of topics has been termed discourse structure (Polanyi 1988). By structuring the discourse into parts that each has a clear topic or a goal, the participants ensure relevant contributions. Each topic section provides a context that becomes the active focus of attention. References, such as pronouns, are interpreted in that context (Grosz and Sidner 1986). It is important that everyone follows the discourse structure in order to stay on the same page so to speak, and therefore transitioning to a new topic is often announced or negotiated (Grosz and Sidner 1986; Hirschberg 1990; Kendon 1990; Clark 1996).

Behaviors

Discourse structure and the transitions within it are clearly reflected in the accompanying nonverbal stream (Kendon 1987; Chovil 1991; McNeill 1992). Behaviors typically involve motion and a number of body parts proportional to the impact of the shift on the ongoing discourse (Kendon 1990). For example, changing the topic of the conversation altogether is usually preceded by a change in overall posture, whereas a digression from a main point is often accompanied by a slight gesture to the side (Bavelas, Chovil et al. 1995). Gestures that go with starting a topic or introducing new segments as a part of a speaker's elaboration, e.g. anecdotes, explanations and quotes, are often metaphorical gestures that present or (Clark 1996) offer the upcoming discourse through a conduit metaphor (McNeill 1992). Holding both hands as if presenting a package while saying, "Let me explain..." is an example of the hands forming a conduit for the upcoming explanation as it is being offered to the listener. Returning from a digression, such as when ending a story or explanation and returning to the main topic, is usually signaled by momentarily raising the eyebrows (Chovil 1991).

2.1.4 Information Management

Process

How information gets shared over the course of the entire conversation can be described by a discourse model (Grosz 1981; Prince 1981; Allen 1995). Each utterance corresponds to an instruction from a speaker to hearers on how to update the their discourse model. Discourse entities, corresponding to noun phrases such as "a green cat" in "I saw a green cat yesterday", are added to the model as they are introduced. At any given point in the conversation, the speaker has certain assumptions about what entities are salient in a hearer's model, and can therefore tailor new utterances to maximizing efficiency. For example, just after "a green cat" has been introduced, it may be referred to in an abbreviated format as "it." Discourse entities can be introduced and referred to nonverbally through being pointed at, placed a certain way or by being acted upon (Clark 2001), such as when one picks up a pocket watch and says "9 minutes fast!"

The packaging of information within each utterance has been described as information structure: a structure that accounts both for a new contribution, the rheme, and for the anchoring of that contribution in the ongoing discourse, the theme (Halliday and Hasan 1976; Brown and Yule 1983). As an example, consider the utterance "I am the suspect" in response to the question "Who are you?" Here "I am" serves as the link to the original question and would be considered the theme. The latter half, "the suspect," however is the new piece of information being shared and corresponds to the rheme. Had the question been "Who is the suspect?",

the very same reply of "I am the suspect" would this time have had "I" as the rheme and "am the suspect" as the theme. Information structure highlights the process of contributing something new while preserving local cohesion, resulting in a naturally flowing conversation.

The exchange of information also requires updating the shared knowledge or common ground, a process called grounding (Clark and Brennan 1991; Kendon 1996). A speaker can only be certain that the common ground has been updated, once listeners have given some evidence of understanding.

Behaviors

Nonverbal behavior associated with information packaging serves primarily one of three main functions: emphasis, reference and illustration. Emphasis signals to listeners what the speaker considers to be the most important contribution of the utterance. Reference is a deictic reference to an entity, often used to disambiguate what is being talked about when the spoken utterance is not explicit. Illustration is an iconic or a metaphorical gesture, that together with speech, redundantly or complimentarily describes objects, actions or concepts.

Emphasis commonly involves raising or lowering of the eyebrows that reaches maximum extent on the major stressed syllable of the emphasized word (Argyle, Ingham et al. 1973; Chovil 1991). As the emphasis increases in intended prominence, vertical head movement synchronized with the eyebrow movement becomes more likely (Argyle, Ingham et al. 1973; Chovil 1991). A short formless beat with either hand, striking on the same stressed syllable, is also common, especially if the hands are already in gesture space (McNeill 1992).

Deictic reference can be accomplished with the head or foot but is most commonly carried out by a pointing hand. The reference can be made to the physical surroundings such as towards an object in the room, or to a previously mentioned entity being assigned a specific spot in gesture space. The latter is particularly common when the speaker's narrative revolves around interaction among various characters; the characters often get their own invisible place holders in gesture space, to which the speaker can then point in order to avoid the need to fully disambiguate between them in speech (McNeill 1992).

A kind of a deictic reference to the non-visible discourse entities can also be made without a spatial placeholder. Sometimes the speaker makes a pointing gesture towards the listener when mentioning entities that were an item of discussion previously in the discourse as if to remind the listener to search for them and bring them back into play (Bavelas, Chovil et al. 1995). Similarly, a speaker may sometimes point towards a listener when referring to an entity previously introduced by that listener, as to acknowledge their contribution (Bavelas, Chovil et al. 1995).

Illustration is the spontaneous portrayal of some semantic features of the current proposition. The particular features may lend themselves well to be portrayed by modalities other than speech, and therefore what the nonverbal expression presents may be important complementary information (Kendon 1987). A good example is the depiction of path and manner through gesture when the speaker wishes to communicate a particular kind of movement, such as bouncing. While the utterance may refer to the movement as "goes down the hill," the accompanying gesture can express the rest of the idea by tracing a bouncing path. This is an example of an iconic gesture, a class of gestures that deal with describing concrete objects or events, often by attempting to replicate their visual appearance or behavioral characteristics (McNeill 1992). Illustration may also involve other parts of the body including the face (Chovil 1991).

As a part of grounding behavior, listeners can display nonverbally how they evaluate or how well they understand what the speaker just conveyed (Chovil 1991). Such evaluation typically involves a facial expression. A frown can signal confusion. A smile, sometimes accompanied by a series of small head nods, can signal a clear understanding. Disbelief or surprise can be signaled through an appropriate emotional expression, often held throughout the clause being evaluated. And a sincere appreciation of the situation being described by the speaker can elicit motor mimicry where the listener mimics the speaker's own facial expression (Chovil 1991).

Negative feedback can be in the form of looking away from the speaker (Kendon 1987) or simply ignoring a request for feedback and showing no reaction at all (Rosenfeld 1987). Typically after a failure to elicit positive feedback, or any feedback at all, from the listener, the speaker displays a keeping turn signal, consisting of looking away from the listener while keeping hands in gesture space (the area in front of a speaker where most spontaneous gesture activity occurs). This signal may also occur if listener feedback seems premature by taking place just before normal feedback gaps (Duncan 1974).

2.1.5 Summary

All face-to-face conversations have certain things in common that boil down to four fundamental processes. To ensure that a conversation is successfully conducted, these processes have to be supported. Nonverbal behaviors play an important role in supporting them face-to-face, but in online chat environments such cues are absent, and therefore problems can arise as seen in section 1.3.3. A technology that attempts to bring the nonverbal cues into the CMC will be reviewed next.

2.2 Video Mediated Communication

As early as 1926, scientists at Bell demonstrated a telephone that transmitted a video image along with the audio. Termed the Picturephone, this contraption was considered a logical next step for communication

technologies; seeing as well as hearing the person you were talking to would bring the experience closer to being face-to-face. However, today video-mediated communication (VMC) devices have not become as commonplace as expected and many early studies on the contribution of video to remote collaborative task solving showed no benefits over audio-only connections.

A number of recent studies attempting to explain the slow adoption of VMC have shown that today's VMC devices provide many important benefits over audio-only but are also hampered by important limitations and in some cases may introduce negative artifacts that compromise the interaction.

2.2.1 Benefits

Some of the benefits provided by VMC include the availability of nonverbal feedback and attitude cues, and access to a gestural modality for emphasis and elaboration (Isaacs and Tang 1994; Doherty-Sneddon, Anderson et al. 1997; Isaacs and Tang 1997). When there are lapses in the audio channel, the visual channel shows what is happening on the other side, providing important context for interpreting the pause (Isaacs and Tang 1994). This ability to continually validate attitude and attention may be the reason why VMC has been shown to particularly benefit social tasks, involving negotiation or conflict resolution. However, benefits for problem-solving tasks have been more evasive (Doherty-Sneddon, Anderson et al. 1997). People are more willing to hold delicate discussions over video than over the phone, and for many, being able to establish the identity of the remote partner is important (Isaacs and Tang 1997). Groups that use VMC tend to like each other better than those using audio only (Whittaker and O'Conaill 1997).

2.2.2 Limitations

Many important limitations of VMC prevent it from achieving the full benefits of face-to-face. Turn-taking and floor management is difficult in groups because it relies on being able to judge exact gaze direction, something that most VMC systems don't support (Isaacs and Tang 1994; Whittaker and O'Conaill 1997). Judging a collaborator's exact focus of attention when observing or helping with a task is difficult for the same reason (Neale and McGee 1998). Side conversations cannot take place and any informal communications have been shown to be extremely difficult to support (Nardi and Whittaker 2002). Pointing and manipulation of actual shared objects is troublesome (Isaacs and Tang 1994; Neale and McGee 1998). Many VMC systems buffer the audio signal so that it can be synchronized with the video; however, the introduced delay can be highly disruptive and work against many natural communication processes (Isaacs and Tang 1997; O'Conaill and Whittaker 1997). Compared to desktop systems, the process of scheduling teleconference rooms and

sitting in front of a large TV screen, contribute to an unnatural passive or formal style of interaction (Isaacs and Tang 1997). Some systems fail to properly provide cues to the social context of interaction, such as whether a conversation is public or private (you cannot see who is in the room outside the view of the camera), which prevents users from framing their interactive behaviors (Lee, Girgensohn et al. 1997).

2.2.3 Evaluation difficulties

It is not a simple matter to evaluate the impact of VMC and one should be careful not to take some of the early results on limited task performance gain as indicating the unimportance of visual information in general. First, the range of VMC systems and their properties such as display size, presence of delays, synchronization of channels, half or full duplex and possibility for eye contact all have effect on the supported communication processes and may well account for inconsistent findings (Doherty-Sneddon, Anderson et al. 1997; O'Conaill and Whittaker 1997). Secondly, it is important to carefully consider the appropriateness and role of video for different kinds of task contexts and how it should display more of the shared environment than just "talking heads" for collaborative working (Neale and McGee 1998). Lastly, researchers have pointed out that a lot of what the video provides is process oriented rather than product or problem oriented, and that the most visible effects may be long term in nature (Isaacs and Tang 1994). Studying VMC needs to be focused on how it can be usefully integrated into people's work practice and needs to employ combined methodologies (Isaacs and Tang 1997).

2.2.4 Design guidelines

The research on VMC provides useful insights and design guidelines for developing any tools for synchronous CMC. One of the most important concerns is to enable behaviors associated with particular collaborative tasks and take advantage of users' existing collaboration skills (Isaacs and Tang 1994). Implementation of directional audio and video may prove crucial for approaching face-to-face performance (O'Conaill and Whittaker 1997; Taylor and Rowe 2000). Video is often more effective when combined with other means for interaction such as graphics, text and computer applications, essentially broadening the users' shared environment (Isaacs and Tang 1994; Isaacs and Tang 1997; Lee, Girgensohn et al. 1997). For example, video of the face has been shown to assist with collaborative coordination of activity displayed on a different screen (Neale and McGee 1998). When such an integration is provided, the seamlessness of transitions made between various spaces will affect usability. It is important for supporting primarily casual or social interactions to strike a balance between very short connection times and the ability for the "receiver" to negotiate with the "caller" whether to proceed with the connection. Providing ways to protect privacy is always an issue (Isaacs and Tang 1997) and allowing users to control how they

appear, possibly blurring the video or replacing their image completely is an often requested feature (Lee, Girgensohn et al. 1997).

2.2.5 Innovative VMC systems

A number of variations on the classic video conferencing system have been developed, each attempting to address some of the limitations mentioned above. For instance, to provide correct gaze cues, the Hydra prototype developed at the University of Toronto displays each participant on a separate LCD screen with an embedded camera. The HERMES system does not attempt to display correct gaze, but in order to better integrate remote participants with a FTF meeting, arranges video monitors around a circular meeting table so that each local participant directly faces a monitor. This allows the people around the table to shift their gaze from the monitor to the others around the table with little effort – a configuration that, as it turns out, encourages local participation more than when everyone is lined up in front of a single monitor (Inoue, Okada et al. 1997). These systems do not provide a shared working area. A landmark system that combined gaze awareness and a shared computer application was ClearBoard, a system that displayed the application on a translucent surface, through which one could see the other participant on the other side (Kobayashi and Ishii 1993).

These systems all rely on a relatively complex equipment and infrastructure, and especially in the case of ClearBoard, don't scale very well with increased number of participants. To address this, a number of systems construct a virtual shared space on a regular desktop machine and project images of participants into this space, using computer graphics techniques. Both InterSpace (Sugawara, Suzuki et al. 1994) and Free Walk (Nakanishi, Yoshida et al. 1996) are examples of live video images mapped onto icons in 3D space that can be moved around to form arbitrary discussion groups. Taking the idea of mapping video into a different space, (Paulos and Canny 1998) have built robots, carrying a two-way live video and audio feed, that a user can remotely drive around a physical environment. While the orientation of a mounted video image or icon can hint at the user's focus of attention, the expression on the image itself does not necessarily map correctly onto the image's configuration in the remote space. Also it is not very natural for participants to have to manually rotate their image or robot while engaged in a discussion. A few systems have tried to measure where a participant is looking in a desktop virtual environment and based on that automatically rotate an icon. The GAZE groupware system was one of the first systems to do this, but it only provides static images (with variable orientation) (Vertegaal 1999) while a system demonstrated by (Taylor and Rowe 2000) implements live video icons with automated orientation. These systems only deal with faces, but providing important gestural capability, such as pointing, is still to be solved.

Video Mediated Communication represents attempts at creating a window between separate geographical locations through which nonverbal behavior can be gleaned. The research in this area helps us to understand the role of nonverbal behavior in conversation and collaboration, while also pointing out many of the hard problems presented by remoteness. One of the hardest problems is how to give the impression that participants are sharing the same space. An approach emerging in some of the innovative systems described above is to bring participants into a shared virtual environment.

2.3 Avatar Mediated Communication

An avatar is a user's visual embodiment in a virtual environment. The term, borrowed from Hindu mythology where it is the name for the temporary body a god inhabits while visiting earth, was first used in its modern sense by Chip Morningstar who along with Randall Farmer created the first multi-user graphical online world Habitat in 1985 (Damer 1998). Habitat was a recreational environment where people could gather in a virtual town to chat, trade virtual props, play games and solve quests. Users could move their avatars around the graphical environment using cursor keys and could communicate with other online users by typing short messages that would appear above their avatar. Habitat borrowed many ideas from the existing text-based MUD environments, but the visual dimension added a new twist to the interactions and attracted a new audience (Morningstar and Farmer 1990). Avatar-based systems since Habitat have been many and varied, the applications ranging from casual chat and games to military training simulations and online classrooms.

2.3.1 Graphical Chat

Inspired by the vision of science fiction, such as Neuromancer (Gibson 1994) and Snowcrash (Stephenson 1992), and fueled by the sudden appearance and growth of the World Wide Web, many embraced the idea of cyberspace, a visual representation of the global network where all its users would roam as avatars, going about their electronic business or just stroll down the virtual commons. It was believed that virtual environments were a natural extension of web pages, they would be online places you could visit, but unlike browsing the web, you would actually be able to see other users flock to the same locales, providing possible chance encounters and giving you a sense you were not surfing alone (Curtis 1992; Damer 1997). The race to build online cities and communities has been well documented and researched (Suler 1996; Braham and Comerford 1997; Damer 1997; Waters and Barrus 1997; Dodge 1998; Dickey 1999).The first Internet based virtual environment employing avatars was Worlds Chat in 1995 (Worlds Inc.). Many others quickly followed such as AlphaWorld (now ActiveWorlds) (Worlds Inc.), The Palace (Time Warner), Worlds Away (Fujitsu Software Corp.), V-Chat (Microsoft) and Black Sun Passport (Blaxxun).

36

However, these places have not received the amount of general acceptance as alternatives to face-to-face socialization as was expected, in part because the avatars tended to do a poor job of exhibiting social activity (Vilhjalmsson 1997). Whereas the avatars were meant to give you a sense of being among other people, their static stares and abrupt movements would instead fill you with a strong feeling of alienation. Some systems offered the users ways to animate their avatars in various ways by pressing buttons or selecting entries from menus, such as in World Inc.'s ActiveWorlds. However, this added yet more interface controls to worry about along with the already cumbersome movement control, and since the behaviors were explicitly initiated, natural spontaneous behaviors were still missing, such as reactive glances and expressions of recognition.

While most of the systems, like Habitat, had their users communicate via typed text, some systems such as OnLive! (Electronic Communities) and SmartVerse (SmartVR) integrated voice communication. Being surrounded by spatialized audio certainly heightened the sense of presence, but again the associated body and head motion was missing, including appropriate gaze. At best, the avatars would exhibit automated mouth movement based on the intensity of the speech.

2.3.2 Multiplayer Games

A popular category of avatar-based online socialization is massively multiplayer online role-playing games (MMORPG). The history of these games is firmly rooted in the tradition of text-based MUDs, which in turn traces its roots to tabletop pen and paper role-playing games such as Dungeons and Dragons. These are entire evolving worlds, usually mythological or futuristic, that contain ecologies, economies and evil emissaries. Unlike multiplayer first-person shooters (MFPS), such as Unreal Tournament, where players bring their avatars to engage in short skirmishes with a dozen or so other players, players of MMORGPs take their avatars on lifetime journeys through persistent lands inhabited by thousands of other players, engaging in politics and intrigue that span across a number of online sessions.

The first MMORGP of this sort was Meridian 59, released in 1996, soon followed by the popular Ultima Online in 1997 (Electronic Arts), EverQuest (Sony) and Asheron's Call (Microsoft) in 1999 and a whole score of other persistent universes released in the last couple of years. Obviously social interaction plays an important role in MMORPGs, yet the avatars they offer do not provide convincing conversational behavior. Head and face are usually not articulated and gestures are very limited. More effort has been spent on creating flashy spell effects and colorful attire.

Game developers have done a remarkable job of bringing shared virtual environments to the desktop and filling them with breathtakingly realistic vistas and fully interactive props, but the avatars they provide have

naturally always been designed to help players carry out primary game objectives rather than to reflect general human capacity and behavior. Game worlds inspire, but they quickly break down when taken out of context and applied to other domains such as online collaboration and learning because the game objective is no longer relevant.

2.3.3 Online Learning

Shared online environments are increasingly being used to support learning (Lehtinen, Kakkarainen et al. 1998). Such environments can provide benefits that include continuous informal access to a community of other learners and instructors (Bruckman 2000), a remote presence at in-class lectures and discussions for those unable to attend physically (Isaacs and Tang 1997) and the possibility of carrying out a variety of activities, experiments and explorations in virtual worlds that would be too costly, dangerous or simply impossible in the physical world (Roussos, Johnson et al. 1998).

Text-only chat rooms have been successfully employed for outside-class group discussions where students get together in an informal setting to discuss class material or have cheerful conversations (Lindeman, Kent et al. 1995; Barnett 2001; Spears 2001). The sense of the class as a community often coalesced in the chat environments (Lindeman, Kent et al. 1995). MOO environments are a more advanced form of textual chat rooms that allow their participants to program and interact with various artifacts that respond to and generate textual messages. A MOO can consist of hundreds of interconnected persistent rooms that each contains a set of artifacts. MOOs therefore allow its users to not only chat, but also engage in various shared activities such as planning, constructing and testing artifacts that then become props in an evolving community. This sort of an environment can provide a great setting for learning because it integrates a supportive social context with a problem solving context. One very successful implementation of this is MOOSE Crossing, an environment designed to help eight to thirteen year old children learn how to write computer programs (Bruckman 2000).

As for the use of avatars in learning environments, the University of Colorado-Boulder conducted an entire Business Computing course online, relying on various CMC tools including the Web, video-conferencing, and also the avatar-based Active Worlds shared online virtual environment. The environment essentially provided a virtual campus, where students could access resources located in various buildings. Walking paths and shared patios next to these "context" buildings naturally grouped students working on related things and provided opportunities for discussion and unplanned encounters (Dickey 1999). Also using the Active Worlds environment, the Active Worlds Academy provides regular 3D modeling classes, but these are more in the form of lectures along with demonstrations inside the environment itself (Dickey 1999). Virtual Cell

and Geology Explorer are two avatar-based graphical environments built on top of a classical MOO. They both allow college students to explore and conduct experiments inside a simulated environment, a living cell in the former case and an island full of interesting geological sites in the latter. The students are given various virtual instruments that operate on the artifacts that are found and are encouraged to share their findings with fellow travelers. These environments have been found to produce higher scores on special scenario-based assessment tests than non-interactive WWW activities (McClean, Saini-Eidukat et al. 2001).

Examples of children's learning environments, include ZORA, ExploreNet and NICE. ZORA is designed to support the exploration of identity and values through the building of personally meaningful artifacts that the children can reflect upon and share with others in a community. The system is built on top of Microsoft's VChat platform and allows the children to construct their own environments, objects and avatars. ZORA was found to facilitate the exploration of powerful abstract ideas by making them almost visible and malleable (Umaschi Bers 1999). The idea behind ExploreNet is that under the leadership of teachers, students could learn to construct their own virtual worlds that teach specific concepts to other students. The creators of a world could interact in real-time with guests to their world through various characters. By participating both as mentors and learners, the students are engaged in co-construction of knowledge (Hughes and Moshell 1997). NICE (Narrative-based, Immersive, Collaborative Environment) provides both a fully immersive Virtual Reality interface (3D Cave) and an animated Web interface to a virtual garden that children have to construct, cultivate and tend together. The children can interact with each other via speech, but only the ones using the VR interface have their head and gesture movements tracked and applied to their avatars. Initial results show that the presence of avatars were a strong spur to social interaction, but learning goals were obscured by lack of directedness combined with novelty and usability issues (Roussos, Johnson et al. 1998).

Many of the existing online learning environments show potential, but most of them also presented some problems that relate to the lack of adequate communication mechanisms. It is common that the users of textual chat systems complain that it is hard to keep track of multiple conversation strands and in the case of MOOs, that the environment tends to overload students with activities overwriting and interrupting the text-flow (Lindeman, Kent et al. 1995; Barnett 2001). In systems that provide avatars, limited nonverbal behavior causes difficulties in using "traditional methods of maintaining control and signal turn-taking" (Dickey 1999), the interface was found confusing and typed messages would get ignored (Hughes and Moshell 1997), and children would find it difficult to organize themselves so everyone was heard (Umaschi Bers 1999). Both in ExploreNet and NICE, children felt that other children's avatars were there to compete as opposed to collaborate with them on the learning tasks

(Hughes and Moshell 1997; Roussos, Johnson et al. 1998). This confusion may well relate to cues given off by the avatars that were inappropriate in a collaboration context.

2.4 Innovative Avatar Control

Most avatar-based systems ask that users control the behavior of their avatars by selecting a motion from a set of pre-defined animations, either presented as menu options or activated by key presses. It then requires conscious effort to activate any avatar motion. This is fine for high-level actions such as "dancing" or "eating." However, more fine-grained behavior, especially the kind of behavior that needs to be synchronized with speech such as gesturing, nodding or glancing, cannot be produced that explicitly because of their spontaneous nature. The number and complex sequencing of these behaviors would also burden the users with excessive control (Cassell and Vilhjalmsson 1999).

Other ways of controlling an avatar have been proposed by a variety of researchers. Most of these control methods fall into one of three categories: Text driven, device driven or performance driven.

2.4.1 Text Driven

Comic Chat (Kurlander, Skelly et al. 1996) automatically generates a comic strip depicting the participants of a conversation from the text messages passed between them and a user controlled "emotion wheel." The characters in the comic strip, the avatars, are automatically framed to give the impression of a face-to-face group interaction and their expressions reflect an emotion set by their user prior to transmitting each message as well as keywords in the message text itself. Similarly the Illustrated Conversation creates an animated performance of a group interaction by automatically choosing an avatar representation, in this case a portrait from a set of portraits, that reflects whether the user is active or not and to whom they are attending (Donath 1995). While highly innovative in their rendering of the conversation, Comic Chat does not provide a continuous embodied presence and Illustrated Conversation delivers a headshot that only changes because of a few control events but remains static otherwise. Just using the text messages themselves to drive continuous avatar gesture has only been attempted in Signing Avatars (Vcom3D, Inc.), where the text is translated on the fly into American Sign Language.

2.4.2 Device Driven

Device driven control employs specialized input device and maps its manipulation into avatar motion. For example, VOES (Lee, Ghyme et al. 1998), maps modified Korean Sign Language produced by the user wearing a CyberGlove into control parameters for avatar motion. Different signs correspond to different motion types, such as "bow" and

"walk," and the direction that the sign is given in announces the direction towards which the action is taken. Similarly, Cursive, uses pen gesture to drive avatar motion. The symbol that is sketched indexes a pre-recorded avatar gesture, which is played back, modulated by the drawing style of the stroke (Barrientos 2000). Specialized devices that resemble the actual avatar have also been used, such as a stuffed chicken fitted with a number of sensors used to control an animated chicken in an interactive cartoon (Johnson, Wilson et al. 1999). Most of the device driven control schemes require that the user learn a set of commands and how they map onto various avatar movements. Learning the commands introduces an overhead that may discourage some users, and the control only captures explicit actions and therefore the avatar may not exhibit more fine-grained spontaneous behavior expected in face-to-face interaction.

2.4.3 Performance Driven

Performance driven control elaborates on the simple idea of having the avatar mimic exactly the behavior of its user, including then of course any spontaneous movement. This requires that the user's every movement be tracked in some way, either by having the user don an instrumented suit or by having computer vision trace specially marked or otherwise salient features on the users body. An example of the former is the MotionStar motion capture device from Ascension Tech, Inc. that places up to 18 sensors on a users body. Each sensor reports its position and orientation relative to an electromagnetic field generated in the vicinity. An example of the latter has been demonstrated with ALIVE, where a user's silhouette is automatically extracted from a static background and used to detect a user's pose. The pose was then used to select a graphical portrait of that same pose, morphing between the two closest portraits if an exact match was not found (Darrell, Basu et al. 1997). The main problem with directly mapping behavior from a user's body to the avatar's body is that the avatar exists in a world that is drastically different from the user's. In a common scenario, the user is sitting in front of a desktop computer, while the avatar is strolling up and down a street in a virtual city. If that avatar were to take on the user's posture and gaze pattern, it would it would appear very out of place.

2.4.4 Abstract Visualization

Not all graphical chat systems strive for re-creating face-to-face behaviors in anthropomorphic avatars, in fact it has been suggested that given the flexibility of the medium, designers of chat systems could go beyond reality when augmenting the conversation experience (Donath 2001). Combining computational analysis of the discourse with methods from aesthetics and visual design, both ChatCircles (Viegas and Donath 1999) and Coterie (Donath 2002) have pioneered the abstract visualization of conversation with dynamic shapes and colors representing participants, activity, topics and interaction history. This is an ambitious approach that,

41

unlike approaches that model actual human behavior, has to invent a whole new visual language, which may or may not turn out to be intuitive.

2.4.5 Automated Avatars

Treating avatars as autonomous agents under the user's influence, as opposed to being directly driven by them was first proposed in BodyChat (Vilhjalmsson 1997), described in more detail in section 2.6, but related approaches are emerging. In particular, several research groups are now looking at automating gaze behavior in avatars and evaluating the effect this has on users.

Microsoft Research has built an algorithm for controlling the amount of gaze between participants, based on whether they are in a speaker or listener role and statistics drawn from gaze behavior studies (Colburn, Cohen et al. 2000). They conducted an experiment where subjects spoke with a remote experimenter, represented by an animated gaze avatar, a static avatar and a blank screen. The subjects looked significantly more at the screen when there was an avatar there than when the screen was blank. The animated avatar was looked at a lot more than the static avatar when subjects were listening to the experimenter, though this did not reach significance. The relatively weak conclusion was drawn that the animated behavior of the avatar was having some effect on the user behavior.

Follow-up work involved studying groups of subjects interacting with each other in four different conditions: audio only, an icon interface, an avatar interface and face-to-face. The icon interface provided static pictures of everyone present and highlighted the picture of the current speaker. In the avatar interface, photo-realistic avatars of all participants were seen sitting around a table. The speaking avatar would be shown speaking while all the avatars followed the gaze algorithm. More pauses and shorter utterances were found in the audio only condition than in the other conditions. A survey showed that people felt they could express themselves significantly better in the avatar condition than in the audio only condition. Furthermore, moving from audio only, to the icon interface and to the avatar interface the subjects felt it was increasingly easier to know who was talking and when to talk themselves (Colburn, Cohen et al. 2001).

Another experiment involving avatars with automated gaze behavior is described in (Garau, Slater et al. 2001). This time two subjects interacted with each other in an audio only condition, random avatar gaze condition, algorithmic avatar gaze condition and through a video tunnel. Similar to the previously described work, the timings for the gaze algorithm were taken from research on face-to-face dyadic conversations and based on who was speaking and who was listening. A questionnaire assessing perceived naturalness of interaction, level of involvement, co-presence and attitude toward the other partner showed that the algorithmic gaze outperformed the random case consistently and significantly. This

suggested that for avatars to meaningfully contribute to communication it is not sufficient for them to simply appear lively. In fact, the algorithmic gaze scored no differently than the video tunnel with regard to natural interaction and involvement, demonstrating at least subjectively that even crude and sparse (only gaze) but appropriate behavior in avatars brings the interaction closer to a face-to-face experience.

Neither of these two groups has attempted to show an objective improvement of automated avatar behavior on online collaboration. The tasks performed by the subjects were designed to elicit interesting discussions, but not for evaluating success or failure with a task. However, (Vertegaal and Ding 2002) performed a study where a random gaze avatar was compared to an algorithmic gaze avatar in a task setting. A subject had to collaborate with two double-blind actors on constructing as many meaningful and syntactically correct permutations of sentence fragments. Interestingly, the subjects in the algorithmic gaze condition gave significantly more correct answers than in the random gaze condition.

This recent work has essentially been confirming the validity of the original BodyChat approach and the results have been consistent with the results from the BodyChat study (see 2.6). The automated behaviors have still only been restricted to gaze and mouth movement, and have not at all relied on any analysis of the conversation itself. This thesis, however, takes the idea all the way and integrates a full set of essential conversational behaviors.

2.5 Embodied Conversational Agents

Researchers in the field of Human-Computer Interaction (HCI) have been interested, from the very beginning of machine computation, in the idea of an interface that is both intuitive to use and powerful in its expressiveness. Some believe that the human natural ability to communicate with other humans holds the key to such an interface. Humans intuitively use language to engage in interaction with each other to, for example, delegate tasks or collaboratively solve problems. Imbuing computers with some sort of a natural language interface has therefore been pursued by a number of HCI researchers hoping to leverage off more than a million years of human-to-human social interfacing.

Recognizing that language interaction not only involves the use of spoken language, but also proper coordination of nonverbal behaviors, a subset of these researchers has aimed at developing autonomous interface agents that have the ability to produce and respond to both verbal and nonverbal behavior. These agents, that are meant to have the same properties as humans in face-to-face conversation, have been termed Embodied Conversational Agents (ECA).

2.5.1 Face-to-Face Interfaces

One of the first ECAs to come to "life" was Gandalf who served as an interface to a database of facts about our solar system. He could detect the movements and gestures of a user wearing a motion-tracking suit, as well as understand a set of spoken queries. In response, Gandalf, represented by a cartoonish head and a hand, would speak and gesture towards a large animated display of planets. While Gandalf's ability to interpret and generate natural language was limited, he was able to smoothly take conversation turns with a naïve user, demonstrating the power of nonverbal signals for turn regulation. In fact, it was shown that turning off some of Gandalf's nonverbal displays resulted in a less orderly interaction (Thorisson 1996).

REA the real estate agent was Gandalf's successor (Cassell, Vilhjalmsson et al. 1999). She improved upon Gandalf's user experience by replacing the motion-tracking suite with unobtrusive computer vision. More importantly, REA was given a genuine natural language generation engine that could construct responses in real-time by drawing from a domain knowledge base and a grammar of English. Having access to the language generation process allowed REA to use rules about distribution of semantic content across modalities, as observed in human discourse, to produce natural conversational gesture to appropriately complement the speech (Cassell 1999).

Bringing ECAs into the physical arena, Kismet is a robot, represented by an articulated head and face, which engages a human in a social interaction. The interaction model is that of caretaker-infant, where the human is in the role of a parent introducing Kismet to the world around it. While it does not have natural language skills, it can hear and see the person in front of it and respond to social stimuli with facial expressions, head movement and vocalizations resembling a child's babble. Similar to Gandalf, it can take conversation turns and generates nonverbal cues to regulate the flow. These cues have been observed to naturally entrain naïve users to Kismet's somewhat slower than human pace. Futhermore, Kismet's ability to make eye contact and visually attend to objects in the environment, gives it the ability to establish joint attention, something that heightens the sense of human-like social behavior. Unlike Gandalf, Kismet's behaviors stem from a model of drives and emotions, giving it the ability to form its own social agenda (Breazeal and Scassellati 1998).

2.5.2 Embedded Interfaces

Cosmo, an ECA embedded in a desktop learning environment, was built as an automated tutor that would help a student to learn about complex concepts, such as network routing, by giving demonstrations and supervise exercises. A principal feature of Cosmo was its ability to produce spatial deixis, e.g. pointing, based on the ongoing discourse and the dynamic problem-solving context. It would essentially know when a pointing

gesture or moving next to an object on the screen was needed in order to prevent ambiguities during an explanation (Lester, Voerman et al. 1999). Similarly, STEVE, an ECA living inside a fully immersive virtual environment, was built as a virtual tutor capable of having a task-oriented dialogue with a student while providing a "hands-on" experience inside an interactive simulation of, for example, an engine room. STEVE was able to demonstrate the proper operation of the simulated equipment while also allowing a student to take over and then answering questions or providing helpful comments when detecting hesitation or wrong moves. STEVE relied on a model that combined a representation of the task context and the dialogue context, to produce both relevant and timely information (Rickel and Johnson 1998).

Focusing more on developing a relationship between users and their personal desktop assistants, the researchers that built Peedy the parrot worked on the computational modeling of emotion and personality. Peedy could recognize a user's spoken commands and then give verbal replies as well as carry out requested tasks on the desktop. Influenced by the attitude expressed in the user's input as well as its own personality, Peedy would choose appropriate speech tone (controlled by parameters such as rate, energy and pitch), language style (such as strong, terse or formal) and gesture form (size and rate of gesture) (Ball and Breese 2000).

2.5.3 Contribution

From the standpoint of HCI, it is important to know whether giving an interface voice and a body, improves the actual interaction. A number of studies (Takeuchi and Naito 1995; Koda and Maes 1996; Andre, Rist et al. 1998; Moreno, Mayer et al. 2000) have shown that animated interface agents have been perceived as more helpful, entertaining and engaging than non-anthropomorphic interfaces. These results suggest that in many situations humans would choose a social interface, which is not surprising since people seem naturally inclined to relate to technology in social terms (Reeves and Nass 1996). However, no study has yet shown that switching to an ECA interface has improved, or hindered, task performance. It may be that the mere presence of a face and a body positively affects a user's perception of the interface, but in order to show a task performance gain, the ECA has to put the face and body to truly skillful use, something that only a few ECAs are starting to accomplish.

From the standpoint of CMC, ECAs provide an opportunity to create and test computational models of human social interaction in a real-world social context. Architectures have been developed that bring together a number of processes that mimic aspects of human communication skills. These architectures and how they perform, increase our understanding of what is minimally required to uphold a conversation. Furthermore, as ECAs become more competent they may start to appear in remote places on behalf of a real human, carrying out business that requires face-to-face

contact. Until then, the idea of computational models of communication and automated embodiment can be applied to avatars, which still are under human control.

2.6 BodyChat

For my master's thesis I created a system called BodyChat where a few communicative nonverbal signals were automatically generated in avatars, based on the proximity of other avatars, some user actions and settings (see Figure 3). The focus was on gaze cues associated with the process of awareness and engagement management as described in 2.1.1. The novelty here was that the avatar was not only waiting for its own user to issue behaviors, but was also reacting to events in the online world according to preprogrammed rules (Vilhjalmsson 1997).

The set of rules active at each moment was determined by the user's overall communicative intent as indicated by high-level user choices. For example, users could set a switch that indicated that they were not interested in chatting with anyone that approached them. This setting would result in their avatar automatically engaging in avoidance behavior whenever someone else showed interest in interacting.

Figure 3: The first version of BodyChat (left) explored in particular support for Awareness and Engagement and a later version (right) focused on turn-taking as part of Interaction Management

A second version of BodyChat built in 1999, focused on group conversations and the process of interaction management. It introduced an algorithm that automatically generated turn-taking behavior, such as raising the arms to request the turn, and giving the turn to another participants by means of gaze, based on keyboard activity, who was being addressed, and who was the last speaker (see Figure 3).

A user study evaluated the approach in detail (Cassell and Vilhjalmsson 1999). Three different versions of the first BodyChat system were compared, one where users could select nonverbal signals directly from a menu, one where all nonverbal signals were automatically generated, and one where menu choices and automation were both available. The results

46

of the study indicated that the avatars in the unmodified BodyChat, i.e. the automation only version, were judged more natural and more expressive than their manually controlled counterparts.

The most controversial result, and perhaps the most important one, was that the users of the unmodified BodyChat felt more in control of their conversations than the users of other versions. This was surprising because the nonverbal conversational behaviors were not under their direct control. However, one could argue that since the users were freed from the overhead of managing nonverbal behavior, they could concentrate on steering the course of the conversation itself.

Other results, while not statistically significant, indicated that the users of BodyChat could better recall information gathered during conversations and that they engaged in longer chats with the strangers they met. This experiment showed that the fundamental approach was strong and well worth pursuing further.

3 Theoretical Framework

3.1 The Big Idea

So far, the behaviors that are key to the coordination of face-to-face conversation have been described, and it has been explained how video mediated communication attempts to transmit these behaviors across long distances. The direct video approach has been shown to be problematic, one of the largest issues being the lack of a shared point of visual reference. Representing people as avatars in a shared virtual environment starts to address this issue and is widely used to support online social interaction. However, the limited repertoire of available avatar behaviors and complete lack of spontaneous conversational behaviors contribute to frustrations with this emerging medium. Having then looked at autonomous agents that are able to exhibit conversational behavior based on what they are saying and on the social situation they are in, it is now a logical next step to see if automation can play an important role in mediating conversation between people. BodyChat was a start that demonstrated how some aspects of human conversation could be successfully automated.

In general, mediating conversation presents a number of challenging research and design questions. Apart from the fundamental issue of what needs to be mediated, two of the more crucial questions are how the person being mediated interacts with the mediating layer and how the receiving person experiences the mediation. This thesis is concerned with both of these questions and proposes a theoretical framework of mediated conversation augmented through automation.

3.1.1 Automated Augmentation

As discussed earlier in this thesis, when people have conversations with each other face-to-face, the mediation layer - their own bodies - rarely require a conscious effort for smooth operation. When technology becomes a part of this layer, we risk introducing additional control overhead that can distract from the communication experience. Any mediated communication system should therefore be designed to minimize this overhead, taking into account any constraints that the channel itself may impose, such as limited bandwidth. One approach to limiting control overhead is to introduce automation. Automation has already been employed for mediated communication where it completely stands in for the person being mediated, such as in answering machines or email vacation messages. It is of course obvious that automation is called for when the person cannot operate the communication channel at all, because they are not there. However, using automation to augment a poor channel of communication is a less explored area.

49

From the perspective of the mediated person, the automation could for the most part simply eavesdrop and then insert helpful signals, such as channel maintenance signals, when appropriate according to predefined and accepted rules. Content should not be replaced or modified, but given support through enriched context, such as by adding visualization. From the perspective of the receiving person, the automation should not be seen as competing with the interlocutor for attention, but rather, it should create one seamless augmented representation of the originator's message, fully integrated and synchronized with the ongoing conversation. In fact, the recipient may not need to know how much of the experience was contributed by an augmenting mechanism and what originated as the sender's input, as long as the experience is consistent with the original intent. In some way this is analogous to some of the more advanced music compression schemes that simply store crucial control parameters and then effectively re-synthesize the music on the receiving end.

3.1.2 General Framework

The input, or the original signal, needs to be in a form that can be interpreted by the augmentation mechanism. It does not have to be a single input channel, but if there is more than one, they need to be brought together and represented by structures that can be correlated in time, because the sender's intent could be encoded in the temporal interaction between channels. For example, a camera and a microphone may be picking up head nods and speech respectively, and knowing which word gets a deep nod allows that word to be marked and augmented for emphasis.

Two important parts need to contribute to the augmentation mechanism itself: a model and a discourse context. The model in essence describes when and how to augment. It models certain communication processes that allow it to interpret the input and to see where the input fails to fully support these processes, or where elaboration may be needed. The context provides the resources to draw from when generating supplementing material. Such resources can be specific to the particular interaction, for example a meeting agenda, or they can represent a more extensive ontology that may for example associate various media types. The context also needs to contain anything generated during the communication session itself, because that material is likely to continue to play a role in the unfolding process.

The output mechanism needs to coordinate a seamless presentation to the recipient. It needs to be aware of some of the limitations inherent in the channel, such as time delays, and then try to compensate, for example through buffering. This mechanism should sit as close to the recipient as possible, so that it can give the impression of being highly reactive in face of recipient action such as replying or pausing. An overview of this framework is shown in Figure 4.

```
                    ┌──────────────┐
                    │   Context    │
                    ├──────────────┤
              ┌ ─ ─▶│   Model      │──┐
   monitoring │     └──────────────┘  │ augmentation
             │                        │
             │                        ▼
   ──────────┼────────────────────────────────▶
   message
```
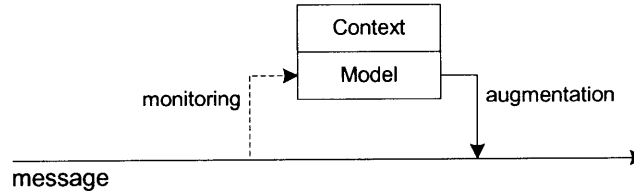
Figure 4: A model, using context, augmenting a channel of communication

3.1.3 Avatars as Agents

This thesis proposes using face-to-face conversation as a model for augmenting online chat or instant messaging. The input is the text message and the output is a graphical rendition of a conversation, taking place in a virtual environment. An avatar delivers the message through speech and gesture while other avatars, representing any other people present, seem to attend and react. All the supporting nonverbal behaviors and the tone of the synthesized speech are suggested by the model and discourse context.

In this case, the output needs to create the illusion of a continuous presence in a shared virtual place, even though the input is only in the form of discrete messages. This means that the output mechanism, or rather each avatar, needs to behave continuously in a convincing human-like manner in order not to break the illusion. One way to approach this is to treat each avatar as an autonomous agent that is imbued with enough intelligence to sustain minimal conversation participation in the absence of input, essentially an embodied conversational agent. When input becomes available, such an agent then needs to seamlessly transition into a mode of delivering the message and generating the behaviors that augment it and making it seem delivered in person.

3.2 The Model

3.2.1 Hypotheses

The model has the role of examining the input and then suggesting behaviors that in this case would be nonverbal behaviors to be carried out by an avatar. If these nonverbal behaviors serve an important communicative function in face-to-face conversation, they should be serving the same communicative function in the online environment. If successfully employed, the following hypothesis should hold true:

Hypothesis 1: process hypothesis

> Compared to synchronous text-only communication, adding avatars that automatically animate the nonverbal behaviors that in face-to-face conversation support (a) Awareness and Engagement Management, (b) Interaction

51

Management, (c) Discourse Structure Management, and (d) Information Management, will improve the overall process of conversation.

If these processes are being improved, the outcome of the online conversation, that is, the lasting impact, should also be improved. An important outcome, that is especially relevant to collaborative environments, is how well the participants performed on a task they were working on. Another outcome that can be impacted is the social relationship between participants. If the first hypothesis is true, then this second order hypothesis should also hold true:

Hypothesis 2: outcome hypothesis

> Compared to synchronous text-only communication, adding avatars that automatically animate the nonverbal behaviors listed in hypothesis 1, improves the (a) task outcome and the (b) social outcome of the online conversation.

The animation has to be synchronized with the words being delivered because nonverbal behaviors are interpreted in their immediate linguistic context, so if the message is displayed as text, it has to show temporality through scrolling or a "bouncing" marker.

However, even though the delivery of the text is synchronized with the motions of the avatar, it is likely that reading the words diverts attention from looking at the avatar itself, and thus nonverbal behaviors may slip by unnoticed. Of course, reading what one is saying is not what we naturally do in a face-to-face situation. This leads to a third hypothesis:

Hypothesis 3: matched modality hypothesis

> Animating nonverbal behaviors in avatars will have a greater impact when they are synchronized with a speech modality than temporal text.

It may therefore be better to synthesize the text message and synchronize the avatar's behaviors with the audio playback than to write the message on the screen.

3.2.2 Monitoring processes online

The model needs to monitor the input into the conversation and based on the status of each process suggest relevant supporting nonverbal behavior such as those described in section 2.1. What follows are suggestions about how each of the processes can be monitored and marked.

Awareness and Engagement Management

If the participants are allowed to move their avatars around the virtual environment and pick their own conversation partners, proximity to other

people's avatars can be used as a trigger for exhibiting minimal awareness. An explicit action, such as clicking on another avatar, may be used to indicate to the model that the user wishes to initiate contact. Based on the context, which includes a person's availability parameter, a reaction to contact initiation can be automated in the receiving avatar. This process, including the exchange of greetings, was extensively explored in (Vilhjalmsson 1997).

Interaction Management

When a participant starts to type a message, the keyboard activity is a good indication that the turn is wanted by that participant, especially when the number of typed letters exceeds typical feedback responses such as "hmm." When a message actually is sent, that participant has then taken the turn. In case of a feedback response, however, the turn does not have to be taken from the current speaker.

When a speaker has finished transmitting messages and no one else has indicated that they want the turn, and the speaker has not explicitly addressed another participant by name, then a good guess is that the turn should be given back to the participant that spoke before the current one (Clark 1996). This would be treating the current contribution as a response to something said earlier. A state machine that keeps track of everyone's participation status, including speaker, hearer, addressee and overhearer status (termed Participation Framework by (Goffman 1983)) can be helpful for determining default transitions when explicit ones are not available.

The message may contain some phrases that indicate feedback elicitation, such as "you know," but punctuation can also serve as a good indicator of feedback eliciting behavior. Commas for instance, are used to mark intonational phrase boundaries, a feature used to locate the "gap" described in section 2.1.2. Longer pauses may be represented by ellipses and can be an indication of speaker hesitation, especially if preceded by an utterance that is not grammatically complete.

In the absence of the more explicit feedback elicitation markers, the timing of feedback elicitation and then the corresponding feedback can also be predicted based on information structure. Speakers tend to look away at the beginning of a theme and then look back at the beginning of a rheme – a place where feedback may be important to a speaker since this is where the new contribution is being made. While this was reported as highly probable behavior, looking back at the beginning of a rheme that also coincided with the end of a turn was found to be an absolute predictor (Torres, Cassell et al. 1997).

Discourse Structure Management

It has been observed that movements between discourse topics frequently occur with the aid of connective expressions, termed cue words or

discourse markers such as "anyway," "that reminds me" and "so". These markers seem to serve a variety of functions, such as marking general topic shifts, digressions, contrast, elaboration and inferential relations (Schourup 1999). Those discourse markers that serve as shifts between segments of discourse, or discourse level cue words, are most often found at the initial position of utterances (Schiffrin 1987). This assumption and the lexical classification of discourse markers has been used to identify discourse topic shifts with relatively high reliability (Hirschberg and Litman 1993).

Information Management

A computational discourse model can be built as a structure that keeps track of discourse entities. When a noun phrase is encountered in the input, that phrase is interpreted as a reference to a discourse entity that should update the discourse model. Since one can refer to the same object in multiple ways, the discourse model has to attempt to map each noun phrase to all entities listed in a discourse context structure, that includes domain ontology and a scene description, and pick the best possible match. The discourse model can mark parts of the input as creating new discourse entities, referring to old entities or increasing the salience of already shared entities (such as parts of a shared scene). Furthermore, by using both the context and a database of semantic relations, such as WordNet, relations between entities, such as contrast, can be marked in the input.

Monitoring information structure is about trying to spot what part of an input message contains a new contribution. This can only be done in light of the discourse history: the utterances accumulated so far. A heuristics developed by (Hiyakumoto, Prevost et al. 1997), splits an utterance into clauses and then each clause into candidate parts for theme and rheme. Based on the number of lexical items in each part that are not found in the discourse history, theme and rheme are assigned.

3.2.3 Behavior Mapping

After the model has identified what communicative process is either associated with a transmitted message or being requested through a user action, applying the collection of approaches above, the nonverbal behavior that supports that process has to be chosen. The literature reviewed in section 2.1 describes behaviors that serve particular communicative functions in face-to-face conversation, and can therefore be a reference for the mapping process. Table 1 lists 14 examples of communicative functions, each belonging to one of the four categories of crucial conversation processes. For each entry, there is a short description of what kind of monitored event helps to identify the function and a description of the corresponding nonverbal behavior along with a picture of avatars engaged in that behavior.

Awareness and Engagement Management

	Func: Avoidance *Trig:* User selection, recipient settings (interrupt not wanted) *Beh:* Brief glances, no re-orientation		*Func:* Invitation *Trig:* User selection, recipient settings (interrupt wanted) *Beh:* Look, smile, re-orientation
	Func: Close Salutation *Trig:* Approach after distance salutation, proximity *Beh:* Look, dip head, smile, handshake		*Func:* Break Away *Trig:* User selection *Beh:* Averted gaze

Interaction Management

	Func: Take turn *Trig:* Start of message delivery *Beh:* Gaze away, ready arms		*Func:* Give turn *Trig:* End of message delivery *Beh:* Gaze at next speaker, relax arms
	Func: Request feedback *Trig:* Punctuation marks *Beh:* Gaze at listener(s), raise eyebrows		*Func:* Signal attention *Trig:* Feedback requested *Beh:* Gaze at speaker, slight head nodding

Discourse Management

	Func: Shift topic *Trig:* Discourse marker *Beh:* Change posture		*Func:* Offer explanation *Trig:* Typical phrases *Beh:* Metaphoric conduit gesture

Information Management

	Func: Emphasis *Trig:* New lexical item within rheme *Beh:* Beat gesture		*Func:* Refer to earlier mention *Trig:* Discourse entity already in discourse model, but not visible *Beh:* Point to placeholder in space
	Func: Refer to visible object *Trig:* Discourse entity introduced or contrasted, mutually observable *Beh:* Point towards object		*Func:* Illustrate object *Trig:* Discourse entity introduced, not mutually observable, depict-able feature *Beh:* Iconic gesture of feature

Table 1: Examples of important communicative functions, how they can be detected and depicted in behavior

4 The Spark Architecture

4.1 From theory to practice

This chapter will describe a general architecture, called Spark[1], that formalizes the model described in section 3.2 in a way that makes it easy to implement actual communication applications that build on it. To recap, the architecture should take as input some communicative event initiated by one online participant, monitor and understand the event in the context of the current interaction, and then produce nonverbal behaviors that supplement the event and finally coordinate a performance that simulates a face-to-face delivery of the communicative event to the other participants. The performance needs to take place in a shared graphical environment.

The design criteria for the Spark architecture reflect the lessons learned from looking at human face-to-face conversation, from reviewing systems that mediate human conversation and from building computational systems that model face-to-face conversation, such as the work on embodied conversational agents. The criteria can be divided into the requirements that the face-to-face paradigm places on the architecture and interface, and the sound software engineering design considerations that make the architecture flexible and useful in a range of applications. The next three sections describe these criteria in more detail.

4.2 Conversation Requirements

4.2.1 Multiple Timescales

The architecture needs to accommodate human communicative behaviors that occur over multiple timescales, ranging from near instantaneous reactions, such as quickly glancing towards something being pointed at, to behaviors that represent relatively stable state such as attending to a task. In conversation all these behaviors interleave, creating concurrent action-reaction loops that span various units of discourse including words, clauses, turns, topics and entire conversations.

Consistency is important because humans expect other humans to act in a certain way, adhering to social and conversational protocols. Consistency across timescales is a part of that. An animated behavior can be expected to stay in motion after it is initiated even after explicit user input has ceased. A behavior expected to occur as a spontaneous reaction may need to be generated before explicit user input is even possible, because

---

[1] When I started my research at MIT, I was intent on finding how a person's "spark of life" could be transmitted across long distances, thus the name.

translating a reaction into keyboard or mouse input plus the network travel time for events, can delay actual delivery too long.

4.2.2 Multi-modal Synchrony

The interaction between various communicating modalities, such as gesture, gaze, head movement, eyebrow movement and speech, is in itself significant and needs to be properly coordinated or the original communicative intent may get lost. For example if one says "I think Joe drove that car" with the pointing gesture towards the car happening with the word "Joe" the meaning is that "Joe" is being suggested as the most likely driver of the car. If however the pointing doesn't happen until the word "that," the meaning is that this particular car is being suggested as the most likely car that "Joe" drove.

Speech and gesture in particular need to be coordinated and fully synchronized. Synchronizing modalities in the output, i.e. when a message is being delivered, is particularly relevant to Spark. However, it is also important to consider the synchrony of multiple input channels in the sender's interface. For example, using mouse or pen movement along with spoken input would require input fusion to take place before communicative intent can be properly annotated. Although multiple input modalities are not dealt with explicitly in this thesis, the architecture should not restrict such an extension.

4.2.3 Shared Discourse Context

The generation of all communicative behaviors meant to augment a message relies on a discourse context that represents the common knowledge backdrop against which the behavior will be interpreted. It is important that this context remains shared and synchronized with respect to all recipients. Even though a message may reach recipients at different times, the augmentation of that message must occur in the same discourse context or the message may end up being interpreted in different ways causing confusion. This context needs to include both a dynamic portion, such as the discourse history that contains all that has been said so far, and a static portion, such as a knowledge base that describes the domain of discussion.

4.3 Interface Requirements

4.3.1 Multiple Levels of Control

Various factors determine how much or little direct control users can or want to have over their own avatars. Outside factors such as network lag or poor input devices can prevent users from exerting complete control. It is also possible that a user's attention is divided between controlling their avatar and some other task. The user may wish to be able to delegate control to the system, based on circumstances.

It has also been pointed out elsewhere in this thesis that manipulating the avatars at the level of individual motor skills would simply place too much responsibility on the users, which could distract from the actual act of communicating. Participants should not have to take on the roles of animators.

It therefore needs to be possible to give an avatar high-level, or intentional level, instructions that are then automatically broken down into a series of motor skills or a single fine tuned motor skill. The system needs to be able to be persistent with regard to these behaviors, so that a momentary distraction or lag won't break the execution of a communicative behavior sequence.

Another kind of control is the one that the shared virtual environment exerts on the avatars by being interactive. Events originating in the dynamic environment may need to access avatar behaviors to produce believable spontaneous reaction. For example, an object that collides with an avatar may need to produce a momentary loss of balance ad a startled look.

Flexible level of control is both a question of being able to span the spectrum from fully controllable puppets to the avatars becoming autonomous agents responsible for carrying out appropriate behavior, but also how other processes within the system can get involved. There have to be multiple entry points as well as paths to the control mechanism. This view of multiple levels of control for avatars is inspired by (Blumberg and Galyean 1995).

4.3.2 Shared Visual Space

One of the lessons learned from the use of video conferencing in collaboration is that having the users share an environment is important. Using virtual environments is one way of addressing this. The architecture has to ensure that communicative performances of all avatars are coordinated both within a single environment and across multiple copies of that environment, in order to maintain a common point of reference. Behavior such as eye gaze or pointing rely on this reference to be meaningful and to intuitively depict shared attention and action.

4.4 Design Considerations

The design considerations reflect sound software engineering practice and address how well the architecture supports flexibility in implementation, variety of applications and scalability. In addition, the architecture is a demonstration of a theory in practice and should therefore closely reflect the theoretical model. What follows is a summary of some of the most important design criteria.

4.4.1 Modularity

- *Domain Independent*
  One should be able to use the architecture to build communication systems that support online conversation in many forms, regardless of topic.

- *Common Module Interface*
  It should be possible to add, expand and exchange modules as needed without having to change the interfaces to adjacent modules.

- *Extendible Representational Language*
  The messages being augmented and the discourse context are both inherently open ended and therefore need to be represented using a representational language that can easily be extended to describe new concepts. Compatibility with existing messaging protocols, behavior descriptions and knowledge representation languages would be a plus.

4.4.2 Scalability

With regard to:

- *Model Improvement*
  Modest changes to the computational part of the model, for example due to improved discourse processing techniques, should not affect the rest of the architecture.

- *Number and Types of Behavior*
  It needs to be easy to add new behaviors as needed and describe those behaviors in enough detail. For example a "head tilt" might need to be added when new empirical data becomes available about its role in the conversation process.

- *Number of Participants*
  As long as the model supports the number of participants, the architecture should not have to be modified to accommodate increased numbers.

- *Number of Conversations*
  It should be straightforward to expand the architecture to cover multiple groups having conversations at the same time.

4.4.3 Abstraction

- *Functional description*
  The messages that are being processed need to be represented in the system at an abstract level so that rules for augmentation are not bound to the surface form alone. It is more robust and scalable to apply a few high level rules based on a functional description of

the message than to use a large numbers or rules specific to the exact wording of messages.

- *Functional morphology*
  There needs to be a clear separation between the functional description and the behaviors that are chosen to help carry out those functions in the end. One reason is that the choice and surface form of the communicative behaviors relies on a number of factors that are highly permeable compared to a description of the underlying meaning. These factors include culture, available display resources (such as available degrees of freedom on an articulated animated body or even whether an articulated body is being used in the first place) or anything that may personalize the behaviors (such as current mood). The morphology of communicative intent needs to be decided on in a module that is both accessible and exchangeable without changes to the model.

4.5 Components

The Spark architecture, shown in Figure 5, consists of a client part, sitting on the computer of each participant, and a server part. The client contains a user interface, where users compose new messages and experience the animated delivery of augmented messages. The client also contains a set of agents, one for each participant, responsible for delivering messages through the user interface. The server receives messages from individual clients, augments them and then broadcasts them back out to all clients. It contains the model and discourse context that allows it to annotate each message with a rich description of communicative intent, as explained in section 3.2.2. Apart from the client/server structure and multiple points for generating behavior (inside various avatar agents), the message processing pipeline derives from the previous work on BEAT (Cassell, Vilhjalmsson et al. 2001).

61
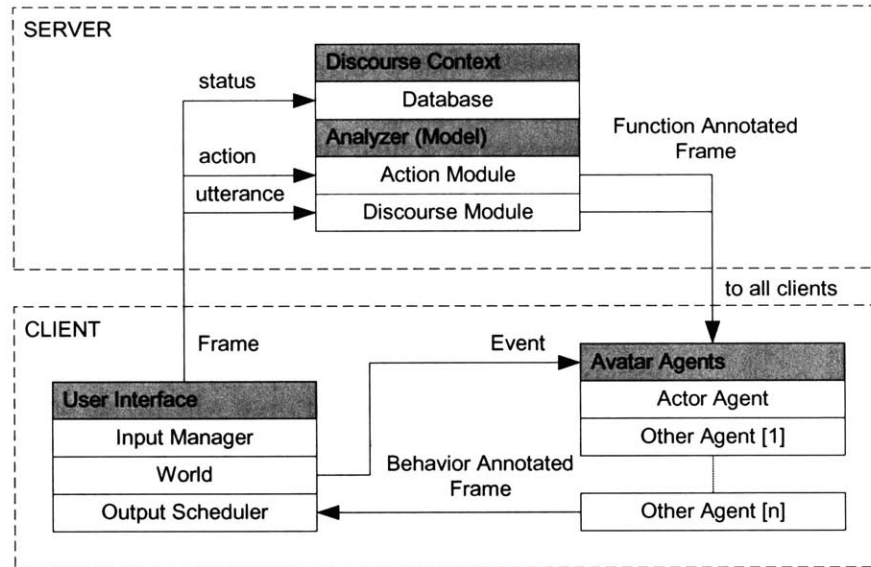
Figure 5: The Spark Architecture

4.5.1 User Interface

World

The simulated environment in which the conversation takes place is managed by a World component. This component keeps tracks of all visual representations of objects and avatars, rendering them from any chosen perspective. The World relies on a scene manager and a rendering engine to deliver a continuous graphical update as users interact with each other and any interactive objects.

Input Manager

An Input Manager is a component that gathers any type of user input and prepares a message for further processing. This preparation involves adding information about the user that caused the event, termed the actor, and identifying any other objects and users, involved. For example if a user clicks on the avatar of another user in the World, it will send a message to the Input Manager saying that a particular avatar was clicked on. The Input Manager turns that into a message saying that user A has selected user B.

When only using text and mouse input, the Input Manager will receive most of its messages from the World, but if other input modalities are available, the Input Manager is responsible for gathering them as well, and integrating them using a standardized representation. For example, if input is spoken, the Input Manager may receive an audio recording, a text string from a speech recognizer and a prosody contour from an intonation tracker. These would all be integrated into a single utterance message where each channel would line up on a single timeline.

62

Output Scheduler

A Scheduler is the converse of an Input Manager as it is concerned with directing all coordinated output within the World. It takes as input a description of one or more behaviors that have to be executed by objects in the World. These descriptions only use relative timing information; such as behavior A has to be executed by avatar X immediately after word B has been spoken by avatar Y. The Scheduler constructs an behavior timeline for each object, preserving the overall timing constraints, and delivers these to the objects as scripts.

4.5.2 Frames

An event is a moment in time associated with some change in user or world state. All interaction related events that pass through the architecture are represented by a data structure termed a frame[2]. A frame holds the actual event description along with some additional context that can be used to interpret the event. A frame is a dynamic entity that can have its event description expanded and an interpretation added as it passes through a sequence of analyzing processes.
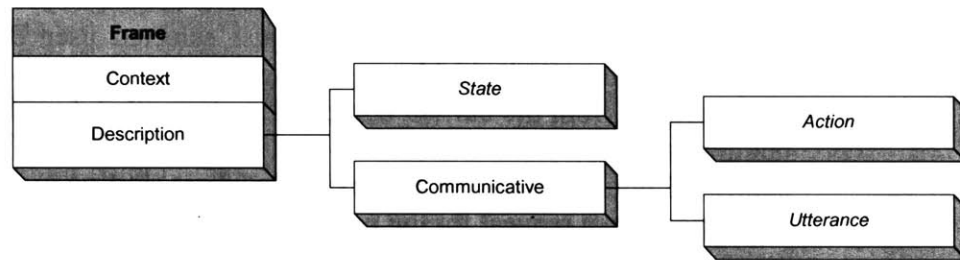
Figure 6: Frame types

There are two categories of frames: those that carry a communicative message to be received by other users and those that only manipulate the discourse context, which in turn affects how future frames are interpreted by the model. These are communicative frames and state frames respectively (see Figure 6). The communicative frames are further divided into two categories, action frames and utterance frames. An action is a communicative event that occurs in the absence of a verbal message, such as nodding in agreement or selecting an on-screen object. An utterance contains words and any nonverbal behavior that are associated with the delivery of those words. State frames don't pass through the communication model, but rather, set parameters within it. For example, when users signal to the system that they are busy and will only respond to important messages, this is announced to the system by a state frame.

---

[2] This structure is loosely based on Marvin Minsky's notion of a *frame* in that it is a structure describing a certain event that is taking place and contains a set of attributes (slots) and values (fillers) to describe everything associated with that event.

XML provides a good way to represent the content of a frame. The outermost tags indicate the type of frame being passed (STATE, ACTION or UTTERANCE), with some of its context described in the tag attributes (for example who is the speaker of the utterance). As the message contained in the frames is being processed and augmented, it can be annotated by adding more XML tags, specifying important functional units. Finally, XML provides tools that allow new tags to be generated from patterns of existing XML tags, which is a powerful feature for generating associated behaviors (see below).

4.5.3 Analyzer

The Analyzer interprets all incoming messages and annotates their communicative function. It consists of two modules, one to process action frames and one to process utterance frames. Both modules have access to the discourse context to help with the interpretation.

Action Module

The action module interprets the action described in the frame and maps it from an interface event to a communicative action. For example, if it receives a frame saying that user A just selected user B, the module replaces that description with one saying that user A is inviting user B to talk, drawing from the current context that shows that A and B are not yet talking and a pre-defined semantic binding for a "selection" event in this context. Figure 7 shows another example where starting to type a message is mapped into a request for the turn.

FRAME IN

`<ACTION><ACTIVITY TYPE="TYPING"/></ACTION>`

Action Module

FRAME OUT

`<TURN TYPE="REQUEST"><ACTION><ACTIVITY TYPE="TYPING"/></ACTION></TURN>`
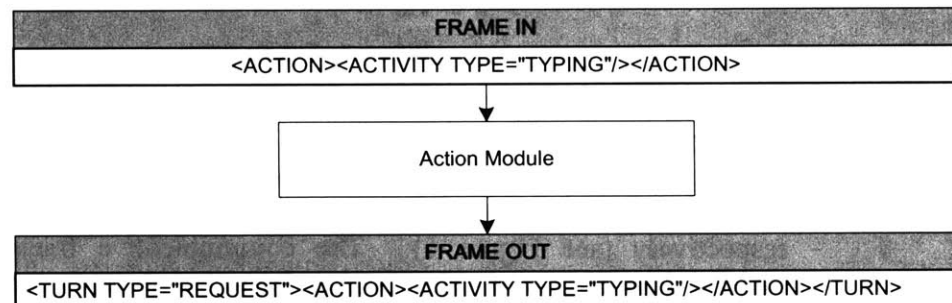
Figure 7: Sample Action Frame Annotated by Action Module

Discourse Module

The discourse module carries out a series of linguistic and discourse analyses to identify and label how the various units of discourse within the text, such as words, phrases and clauses, contribute to the conversation processes described in 2.1. For example, after it has parsed and chunked the utterance into clauses, it annotates each clause for information structure according the heuristics developed by (Hiyakumoto, Prevost et al. 1997). This annotation places a THEME tag around the thematic part of a clause and a RHEME tag around the rhematic part. Other tags include

turn-taking events and discourse entity descriptions. Figure 8 shows how the Discourse Module annotates a short utterance. See Appendix A for a full list of discourse function tags.

```
FRAME IN
<UTTERANCE SCENE="MAP1" SPEAKER="GREEN">give him some gold</UTTERANCE>
```

Discourse Module

```
FRAME OUT
<UTTERANCE SCENE="MAP1" SPEAKER="GREEN"><CLAUSE>
<THEME><TURN TYPE="TAKE"><ACTION><NEW>
<W LEM="give" POS="VB" SYN="VBA">give</W></NEW></ACTION><OBJECT>
<W LEM="he" POS="PR" SYN="NNHD">him</W></OBJECT></TURN></THEME>
<RHEME><TURN TYPE="GIVE" TARGET="BLUE">
<EMPHASIS TYPE="PHRASE"><REFERENCE TYPE="VISUAL" TARGET="MAP:mine">
<REFERENCE TYPE="TEXTUAL" SOURCE="ORANGE"><OBJECT ID="MAP:mine">
<W LEM="some" POS="DT" SYN="NNPE">some</W><EMPHASIS TYPE="WORD"><NEW>
<W LEM="gold" POS="NN" SYN="NNHD">gold</W></NEW></EMPHASIS></OBJECT>
</REFERENCE></REFERENCE></EMPHASIS></TURN></RHEME>
</CLAUSE></UTTERANCE>
```
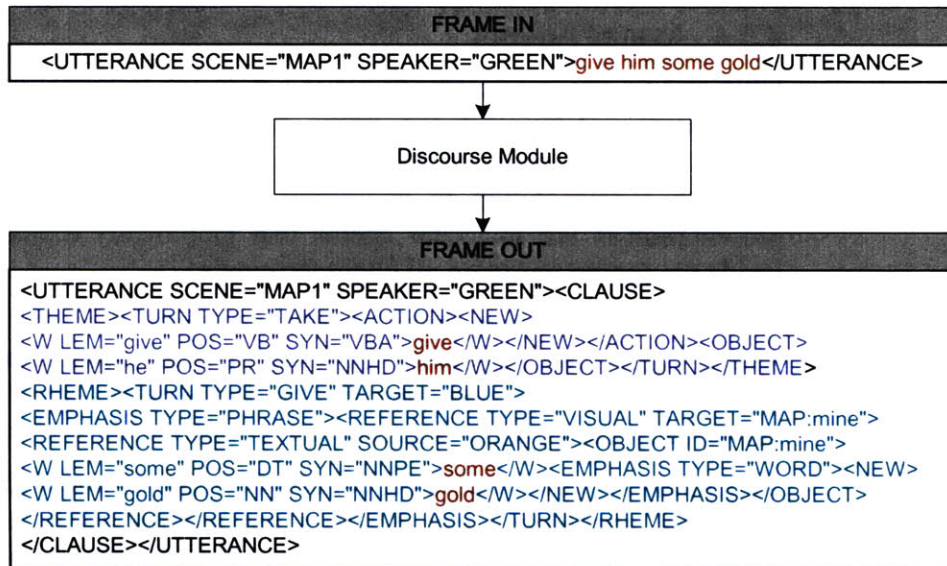
Figure 8: Sample Utterance Frame Annotated by Discourse Module

Discourse Context

Discourse Context is important to the analysis. It is represented by three main data structures: Discourse Model, Domain Knowledge and Participation Framework.

The Discourse Model is a dynamic structure that reflects the state of the ongoing conversation. Central to the Discourse Model is the Discourse History that lists what has been said so far, in particular which discourse entities have been introduced and how recently they have been referred to. Finally, the Discourse Model contains a description of the visual context, more specifically, what objects in the environment are mutually observable by participants and are therefore discourse entities with a certain "given" status.

The Domain Knowledge is a static structure that describes the ontology of a particular domain that relates to the conversation. This ontology can help the Discourse Model to track discourse entities, since there is often more than one way to refer to the same entity – an ambiguity that an ontology may help resolve.

The Participation Framework is a dynamic structure that keeps track of who is present and what their roles currently are in relation to the current utterance. These roles include who is the current speaker, who may the speaker be responding to, who are the other listeners and who are within hearing range while not being active participants.

When a frame leaves the analyzer, it contains a detailed description at a functional level. What began as an isolated event is now a rich description of a communicative action in the context of the ongoing conversation.

4.5.4 Avatar Agents

Ultimately a communicative frame is a message to be delivered to remote participants. Now that the frame has been analyzed and annotated according to the communication model, the delivery itself can draw from this rich representation to coordinate an effective presentation. This presentation is left to avatar agents that graphically represent each user inside the shared world on all client terminals.

When a communicative frame arrives at a client, it is first handed to the Avatar Agent that represents the actor of that communicative message. The actor's job is to now annotate the frame with actual behaviors that nonverbally carry out the communicative functions described.

Annotating a frame with visual behaviors is simply a matter of translating functional annotations into behavior annotations according to a set of translation rules. In essence, this step defines the morphology of the communicative functions, that is, it takes an abstract representation of intent and generates realization into surface form. This is not necessarily a one-to-one mapping because there can be more than one way to realize the same intent. The realization may for example depend on the availability of certain resources such as limbs or time for completion.

The Avatar Agent performs this translation by passing the frame through a small network of Behavior Modules. A Behavior Module takes as input an annotated frame, applies a set of transformation rules, and returns the resulting frame. A basic Avatar Agent contains four Behavior Modules. Utterance frames are handed to the Speaker Module that specializes in co-verbal behavior. Action frames are handed to the Action Module that handles stand-alone behaviors.

Once the acting Avatar Agent has had the chance to populate a frame with behaviors, either through a Speaker Module or an Action Module, the frame is then passed around to all other Avatar Agents that then get a chance to add reacting behaviors. Utterances are processed by Listener Modules and actions get processed by Reaction Modules. This way, everyone represented in a scene can have their avatars spontaneously react to what is going on. In fact, other agents than Avatar Agents could also participate in this reaction phase, for example a Camera Agent could add camera moves to the frame, based on what is happening in it.

4.5.5 Delivery

The output from the Avatar Agents is a frame that now is a detailed description of a performance that involves one or more avatars. This performance has to be carried out in the World. The frame is given to the

Output Scheduler (see World section above) that hands out scripts to the individual world objects. One type of a world object is an Avatar Puppet. Each Avatar Agent has a corresponding Avatar Puppet inside the World. The puppet receives and executes behavior scripts. Puppets can maintain behaviors that have been assigned to them in order to appear to be continuously animated. For example, a script may ask an Avatar Puppet to maintain eye focus on a target object, even if the object moves around the scene, by automatically adjusting head and eye angles. The Avatar Puppet is therefore an advanced graphical object that has a set of motor skills than can be turned on and off as dictated by incoming scripts.

4.6 Innovative Concepts

The Spark architecture introduces two new fundamental concepts to online conversation systems. The first one is functional markup and the other is continuous agency. These two concepts warrant some further discussion.

4.6.1 Functional vs. Behavioral Markup

XML was conceived as a markup language that would be used to describe the structure and content of information, not how it should be displayed, that was the role of formatting languages like HTML. The idea was simple but powerful: by separating the description from the rendering, it would be easy to render different views of the same data. The rendering would be accomplished by applying transformation rules, also written in XML. Being able to generate different views is particularly helpful when the information is complex and those viewing it are interested in a certain subset or particular associations. One can think of the rendered views as filtering the data. Views are also helpful when dealing with constraints inherent in the rendering mechanism. That is, one can tailor the view to fit a certain output device, for example, underlining can be used instead of color on monochrome display devices to represent the same thing.

Spark takes this idea and applies it to messaging. Instead of the typical use where XML encodes the results of a database query that then gets displayed using HTML, Spark uses XML to describe the structure and content of a message as it is being sent from a person to one or more recipients. The model on the server side as described above adds this description. After the message has been annotated, essentially with XML representing functional markup, it is up to each client and the avatar agents within them, how the XML gets rendered. In collaborative virtual environments, the XML is rendered as a performance, strung together of behaviors that carry out the various functions embedded in the message.

As mentioned earlier, the transformation rules can in fact differ from client to client or from avatar to avatar. For example, a client running in Japan could apply transformation rules that convey the messages in a performance that adheres to Japanese social conventions and behavioral

traits, while the same functional description of the message could be subject to Icelandic transformation rules in a client in Iceland.

Taking this idea even further, collaborative groups that include clients that don't sport virtual environments and animated avatars, could apply different kinds of transformation rules that render the conversation as annotated web pages, dynamic abstract 2D visualizations (like Chat Circles) or illustrations (like Comic Chat).

The functional markup is therefore a device-independent representation that supports augmented visualization of the conversation through any means possible. While this thesis argues that articulated avatars, mimicking human nonverbal behavior, are the most intuitive and appropriate visualization, there are certainly times when other views make more sense or are necessary. Spark naturally supports this flexibility by separating functional markup from behavioral markup.

4.6.2 Autonomous Avatars

The idea of automating communicative behaviors in avatars was introduced in (Vilhjalmsson 1997) and is represented by the avatar agent objects in the Spark architecture. Previously avatars were only considered puppets whose control strings would literally always have to be in the hands of their users. Not only did this limit the avatar ability to show spontaneous reaction, but it would also burden the user with too much micro management of behavior. By treating the avatar more as an autonomous agent, it can exhibit programmed reactions and can offload the micro management from the user by accepting instructions at a higher level that it can then break down into the appropriate series of behavior.

4.7 Fulfillment of Requirements

Now that the Spark architecture has been introduced, it is important to explain how this architecture specifically addresses all the requirements and considerations presented in 4.2 through 4.4.

4.7.1 Conversation Requirements

Multiple Timescales

The reaction module in avatar agents can automatically provide an immediate reaction to a speaker's message, and action frames, with their relatively direct path from users to avatars, can provide a near immediate deliberate reaction. Within the utterance frame itself, behaviors can be timed to discourse units of different sizes such as individual words, discourse entities, rhematic parts and entire clauses. Behaviors that span longer timescales can be set through toggling states, both in the discourse context (for example setting participation status) and in the avatar agents

themselves (for example telling the agent it is in an idle state and it should be exhibiting idling behaviors until further notice).

Multi-modal Synchrony

On the output side co-verbal behaviors are generated from and then inserted into the same temporal structure that gives rise to the functional interpretation of the message. A special scheduling module ensures that the precise timing of the generated behaviors is maintained through the actual performance.

On the input side, user interface events are encapsulated in a communicative frame that then gets interpreted. Such a frame could contain a description of more than one input event that occurred concurrently. It would then be the job of the action module to look at all the events in a frame in context to derive communicative function. The frame representation is a logical grouping of related events, which along with the context and model provide a sufficient framework for implementing multi modal fusion.

4.7.2 Interface Requirements

Shared Discourse Context

The discourse context is kept track of in a single place on a server. Interpretation of messages only happens once, in that one particular discourse context. All clients receive the same interpretation of communicative intent.

Multiple levels of control

Users control their avatar agents through frames that describe communicative intent. The frame structure places no constraints on how low or high level this intent is. Frames can result in direct action (action frames) or entire performances (utterance frames). The avatar agents provide autonomy when it is called for, even in the absence of any frame input. The environment itself can affect the avatar agents directly through their perception, which provides an additional control path.

Shared Visual Space

The model centrally describes any communicative function that involves the environment and because its parameters remain constant across clients the resulting behavior is perfectly aligned. For example the current focus of attention is described by the server in terms of a target so that all the avatars are seen attending to the same visual object. Non-communicative behaviors however, such as random idling behavior, can be coordinated locally by the avatar agents and does not have to be identical in all clients. There is an un-avoidable lag involved when communicating to clients, but

the frames are guaranteed to reach their destination so that eventually the different clients catch up and should then provide identical environments.

4.7.3 Design Considerations

Modularity

- *Domain Independent*
  The model of conversation is a general model and the set of communicative functions being annotated should be applicable to any conversation. Domain specific information is kept in a domain knowledge base separate from the model.

- *Common Module Interface*
  All modules receive frames and produce frames. As long as this capability and the general format of frames is preserved, modifications to the plumbing should be straightforward.

- *Extendible Representational Language*
  XML is already gaining widespread support as the knowledge and messaging representation language of choice.

Scalability

- *Model Improvement*
  The action and discourse modules in the model contain modular methods that each annotates a particular communicative function. These methods can be modified without necessarily affecting other methods. For example, the "markContrast" method in the discourse module could be improved to look for more kinds of contrasts than it currently does, without touching the rest of the module. New methods could also be added to the discourse module to produce new tags describing new discourse functions. Entire new modules can also be added to the pipeline to supply other types of analyses and annotation.

- *Number and Types of Behavior*
  Frames or processing modules place no constraints on what tags are added. New tags would not break anything. As long as the new behaviors have corresponding motor skills in the avatar puppet, generating new behaviors is just a matter of adding new generation rules to the behavior modules in the avatar agents.

- *Number of Participant*
  Participants and their avatar agents are simply kept track of by lists that can grow as needed.

- *Number of Conversations*
  The participation framework in the discourse context keeps track of multiple conversations. This database can be consulted to know what conversation is being addressed when a new frame arrives.

Multiple instances of the discourse model can be kept for analyzing each conversation. Annotated frames arriving at the clients specify what avatar agents are a part of each performance.

Abstraction

- *Functional description*
  This is the functional markup added by the analyzer.

- *Functional morphology*
  The behavior markup is separate from the functional markup. The generation rules that produce the behavior markup from the functional markup are treated as modular plug-ins under the control of each avatar agent.

5 The Spark Implementation

5.1 Overview

Spark has been implemented as a collection of C++ and Java classes that together form a functional graphical chat system. This system can then be used for specific applications by populating its world and domain knowledge base and by adding more functional annotations and behavior generation rules as needed. Figure 9 gives an overview of the client side of the implementation. The rest of the chapter will go into more detail.
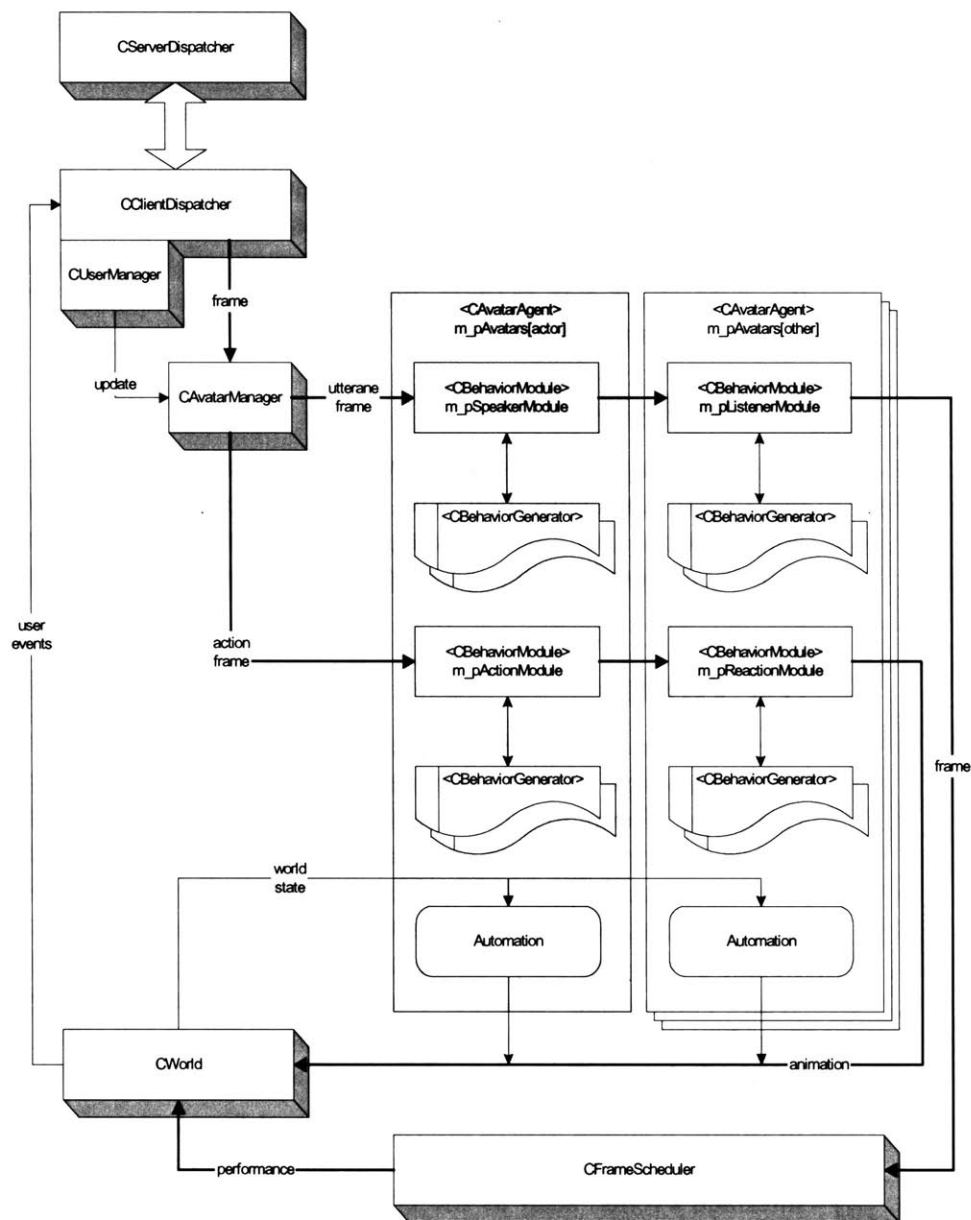
Figure 9: Instances of CAvatarAgent in the clients annotate utterance and action frames received from the server. They can also react to events in the world.

5.2 Networking

The networking is a simple client-server model where each client communicates directly with a central server. The system uses the DirectPlay component of the DirectX 8.1 API to handle this. When clients initially connect to the server, the server places the log-in information in a database that allows it to track who is present and where they are connecting from.

Dispatchers route frames in Spark. A Dispatcher can receive a frame, determine what module can handle it and then send it on. A Dispatcher can also be handed a frame for transmitting it across the network to another Dispatcher, which in turn can get the frame to appropriate remote modules.

In a typical scenario the Client Dispatcher passes new frames across the network to a Server Dispatcher that passes it on to the appropriate processing module. When server side processing is done, frames are given back to the Server Dispatcher that then broadcasts the frames to all clients. Client Dispatchers catch the annotated frames and hand them to the right client side module, usually the avatar manager that lets all the avatar agents process it.

5.3 Management

A user manager keeps track of who is logged into the system, holds their profile and associates them with an avatar agent that represents them in the conversation. An avatar manager manages the set of active avatar agents and is responsible for passing frames to the appropriate agents for processing.

5.4 World

The World's scene manager and animation engine is Pantomime, a high-level graphical object manager developed in Gesture and Narrative Language Group to handle real-time interactive characters. Pantomime allows multiple objects, animate and inanimate to co-exist in a 3D environment and provides a common messaging structure for all of them. New objects, with special functionality can be built and added to the world, as long as they implement a rudimentary World Object Interface. While Pantomime can receive and dispatch messages to any objects in its world, it can also produce its own messages in response to direct user manipulation such as when objects are selected with a mouse click.

In Spark, a single World interface is used for communicating with all the objects in Pantomime. The World interface is also used to monitor user interface events, such as the typing on the keyboard, and to display text messages as overlay on top of the virtual environment rendered by Pantomime. The World interface is built using Open Inventor so that it

can seamlessly integrate Pantomime that also uses Open Inventor for rendering.

5.4.1 Pantomime

Pantomime is written in C++ and was designed to be highly modular, so that extending it to fit the animation requirements of particular research projects would be straightforward. At the highest level, Pantomime consists of a world shell and a set of world objects, loaded and manipulated through the shell. Manipulating world objects involves executing a KQML performative in the world shell. A standard performative looks like this:

```
(tell :recipient "name" :content (command :key value ...))
```

The performative specifies a particular object as the recipient of a command. Each command has the same format: Command name followed by any number of key-value pairs. All world objects are required to implement a command handler. By far the most sophisticated world object, that also implements the largest number of commands, is the Pantomime Humanoid. This object represents an interface to a fully articulated human figure with a wide variety of motor skills.

In the Pantomime Humanoid (Figure 10), each motor skill is implemented as a modular plug-in called a driver. A driver manager relays incoming commands to the drivers that can handle them. The drivers in turn manipulate the various degrees-of-freedom of the humanoid through an arbitrator that acts as a resource manager. The Pantomime Humanoid architecture was the thesis work of Kenny Chang (Chang 1998). Currently implemented drivers include for example a simple "headnod" driver that responds to a command such as (headnod :amt 0.5) by tilting the chin halfway to the chest and back.
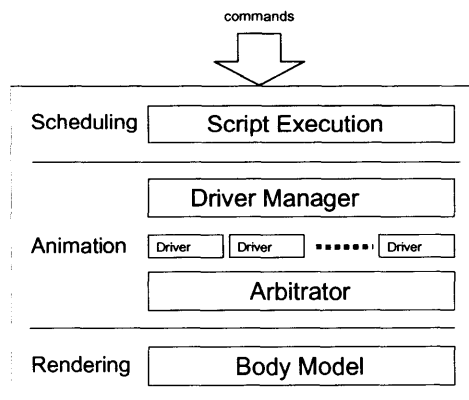
Figure 10: The Pantomime Humanoid modular construction

One of the most important and elaborate drivers is the gesturing driver. This driver, co-developed with Ivan Petrakiev and Vitaly Kulikov, provides commands to construct an entire timed sequence of various gestures such as pointing (using IK), reaching (also IK), drawing a shape

(using key frames) or emphasizing (using displacement). The stroke time of each gesture (i.e. time of the most effortful part of the gesture) can be specified and the driver will calculate a smooth trajectory seamlessly connecting all the stroke points. Being able to precisely time the stroke is important when synchronizing the gesturing with speech.

In Spark, instances of the Pantomime Humanoid object are used as the visual representations of the animated avatars. To distinguish this lower level object from the higher control level of the avatar agent, these instances are called avatar puppets.

5.4.2 Models

The geometry of the avatar puppets and in fact all of the world geometry is specified in VRML files that get loaded when the world is initiated or when new avatars are added. When a user logs into the system, their username is used to retrieve a humanoid model that becomes their avatar puppet on all clients. Humanoid models are expected to adhere to a certain format, based on h-anim (H-Anim 2001), so that all the necessary degrees of freedom can be found and manipulated.

5.5 Server

While the server communication part is written in C++ the processing of incoming frames happens in a Java application called FrameAnalyzer (run from within C++ using the Java VM API). This application contains an action module and a discourse module as well as instances of all the context structures. The FrameAnalyzer sends communicative frames to the corresponding module for processing and gathers the resulting frames to be sent out to the clients.

5.5.1 Action Module

Action frames are handled in a very straightforward manner by simply mapping an incoming action, described in the frame, to a communicative function using a mapping table. This mapping represents bindings between interface events, usually deliberate user actions, and certain semantics that can be arbitrarily defined for each application. A couple of basic bindings are shown in Table 2.

User Action	Action Communicative Function[3]
Press ENTER alone	GROUNDING @TYPE='AFFIRM'
Start typing	TURN @TYPE='REQUEST'

Table 2: **Action mapping table. Here pressing ENTER by itself has been mapped to explicitly giving affirmative feedback, typically resulting in a head nod by the avatar. The act of beginning to type a message creates a frame that gets interpreted as a request for the turn, typically resulting in the avatar raising its arms.**

5.5.2 Discourse Module

The first step of processing incoming utterance frames deals with tagging some of the very basic units such as words, phrases and clauses (see Appendix A for a full list of tags). Marking words, as well as punctuation, is the role of the tokenizer. Once the tokenizer has marked all words with a W tag, it consults a part-of-speech tagger (currently the EngLite tagger from Conexor) to fill in attributes for each word. The first attribute is the actual part-of-speech, such as noun or a verb. The second attribute is the lemma of the word, i.e. the basic root form of the word. An example of a lemma is "be" for the word "were." The third attribute is a light syntax identifier that describes where the word stands in relation to the words around it. This generally marks words as either as the head of a phrase, such as a noun phrase or a verb phrase, or modifiers to such a head. After all the words have been marked and classified, the next processing stage is the chunker. The chunker groups the words together into phrases and clauses based on punctuation and word classes. Noun phrases get an OBJECT tag, verb phrases an ACTION tag and clauses a CLAUSE tag.

When an utterance has been chunked it is ready for the actual discourse processing which attempts to describe the communicative function of the parts that make up the message. The discourse processing is handled by a number of annotation methods, each applied in turn to the utterance. These methods use the discourse context, existing annotation and heuristics supported by the literature to progressively enrich the description. What follows is a summary of each of the currently implemented methods, in the order they are applied. The summaries start with a short description of the discourse phenomenon the method is annotating and the conversation process category is mentioned. Then "Uses" lists the tags and attributes that already need to be present in the text for the method to work (the format is comma separated tag names

---

[3] The function annotation is described by giving the name of the tag first followed by each associated attribute (labeled with an @ symbol) and value pair.

with any needed attributes identified with "@" and immediately following the tag they describe). "Creates" lists the tags that get inserted into the text as a result of running the method and finally the algorithm itself is described.

MarkNew

Lexical givenness. Whether a certain lexical item, i.e. a word, has been seen before in the current discourse (Information Management).

Uses	W @POS @LEM
Creates	NEW
Method	Tag every W element whose POS attribute indicates an adjective, noun or a verb (words belonging to any *open class* except the adverb class) and whose LEM attribute is not identical to the LEM attribute of any W element in the discourse history.

MarkTopicShifts

Movement within the discourse structure. Seeing the discourse structure as a stack of topics, where topics can be pushed onto the stack and popped off later (Discourse structure management).

Uses	CLAUSE, W @LEM @SYN
Creates	TOPICSHIFT @TYPE=(NEXT|PUSH|POP|DIGRESS|RETURN)
Method	Tag the first W of a CLAUSE (skipping to the second if the first W is a connective) if its LEM attribute matches any of the topic shift discourse markers listed by Clark (1996) (see below listed by type). For multi-word discourse markers, the subsequent W elements are also checked for a match.

Next - and, but, so, now, then, speaking of that, that reminds me, one more thing, before I forget

Push - now, like

Pop - anyway, but anyway, so, as I was saying

Digress - incidentally, by the way

Return - anyway, what were we saying

78

MarkInformationStructure

The thematic and rhematic components of a clause. The theme is the anchor to a previous clause and the rheme is the new contribution (Information Management).

Uses	CLAUSE, OBJECT, ACTION, NEW
Creates	THEME, RHEME
Method	Groups together all OBJECTs that occur before the first ACTION in a CLAUSE, calling that the pre-verbal group. Similarly the group of any OBJECTs or ACTIONs occurring after that first ACTION gets called the post-verbal group. If a group or the ACTION contains a NEW element, it is marked as focused. If the pre-verbal group is the only focused group or element, it gets tagged as RHEME and the post-verbal group as THEME, otherwise the post-verbal group gets the RHEME tag and the pre-verbal the THEME tag. If there is only one group, it gets tagged as a RHEME regardless of focus status. If the post-verbal group is focused, the ACTION gets counted with the pre-verbal group, otherwise the post-verbal.
	This follows the heuristics described in (Hiyakamoto, Prevost and Cassell 1997)

MarkEmphasis

Particular attention is drawn to this part of the utterance (Information Management).

Uses	RHEME, ACTION, OBJECT, NEW
Creates	EMPHASIS @TYPE=(PHRASE, WORD)
Method	All numbers get tagged (TYPE = WORD). Every ACTION or OBJECT within a RHEME and that contains a NEW element gets tagged (TYPE = PHRASE) and all the NEW elements also get tagged (TYPE = WORD).

MarkContrast

Two or more items are being contrasted with each other (Information Management).

Uses	W @POS @LEM
Creates	CONTRAST @ID
Method	For each W that is an adjective, tag if its LEM attribute equals the lemma of any antonym or any synonym of that antonym of an earlier adjective W (using WordNet). If a match is found within the current utterance, both W elements get tagged and get an ID number identifying the pair.

IdentifyClauses

The general communicative purpose of the clause. Essentially speech act category, but currently limited to what punctuation reveals (Information Management).

Uses	CLAUSE, W @SYN
Creates	*CLAUSE* @TYPE=(EXCLAMATION, QUESTION)
Method	All clauses ending in a question mark get TYPE = QUESTION and all clauses ending in an exclamation mark get TYPE = EXCLAMATION.

IdentifyObjects

Find the particular discourse entity that a noun phrase refers to (Information Management).

Uses	UTTERANCE @SCENE, OBJECT, W @LEM
Creates	*OBJECT* @ID
Method	For all OBJECTs try to find a match in the set of instances listed in the domain knowledge base (KB) and in the discourse history.
	If a match is found in the KB, then the OBJECT gets the ID set to the unique ID of the matched instance. If a KB match is not found, then the discourse history is searched for a matching OBJECT. If a match is then found, the ID of that OBJECT is used. If no match is found, a new

unique ID is assigned to the OBJECT.

A match score between an OBJECT and an instance in the KB is the number of instance features that are identical to any W LEM attributes contained in the OBJECT. A match score between two OBJECTs is calculated as the number W LEM attributes they contain that are identical. The match that scores the highest is picked as the match. If there is a tie, no match is reported.

IdentifyActions

Talking about an action often calls for descriptive complementary iconic or metaphoric gesturing. Here, verb phrases are linked to action descriptions in the knowledge base (Information Management).

Uses   ACTION, W @POS @LEM

Creates   *ACTION* @ID

Method   For all ACTIONs, try to find a match in the set of action descriptions listed in the KB.

It is a match if the lemma of the head verb in the ACTION's verb phrase is identical to an action description identifier. If no match is found then the search is repeated with the set of all hypernyms of the head verb[4].

Any matching identifier is used as the ID value of the ACTIONS. The ID is left blank if no match is found.

markReference

Find whether a discourse entity is brought in (or *evoked*) through a visual or textual reference (Prince 1981) (Information Management).

Uses   UTTERANCE @SCENE, OBJECT

Creates   REFERENCE @TYPE=(VISUAL, TEXTUAL) @TARGET @SOURCE

Method   Every OBJECT that matches any of the instances listed in the scene description is tagged and the TYPE set to

---

[4] No attempt is made to identify the correct sense of a verb. A match is only checked with the first sense that WordNet returns (generally the most common use).

VISUAL and the ID to the instance ID.

Every OBJECT that matches any of the OBJECTs in the discourse history is tagged and the TYPE set to TEXTUAL, the ID set to the matched OBJECT's ID and the SOURCE set to the ID of the person who last contributed the OBJECT to the discussion.

markIllustration

Indicate a feature of a discourse entity that should be illustrated through an iconic gesture.

Uses   OBJECT, ACTION

Creates   ILLUSTRATE @DESCRIPTION

Method   Every OBJECT within a RHEME and that contains a NEW element gets checked against the KB using the object ID. If this instance of an object has an unusual value assigned to an object feature, as determined by the definition of a typical instance in the KB, the a description of the atypical feature and value are assigned to DESCRIPTION as a string.

Every ACTION within a RHEME and that contains a NEW element gets checked against the KB using the action ID. If a description of the action, or any of its hypernyms (a more generic verb) as shown by WordNet, is found in the KB, that description is assigned to DESCRIPTION.

markInteractionStructure

Attempt to infer who is being addressed (Interaction Management).

Uses   UTTERANCE @SPEAKER @SCENE

Creates   *UTTERANCE* @HEARER

Method   If the HEARER attribute of an UTTERANCE is not already set, first all OBJECTs in the UTTERANCE are examined to see if there is a match with any instance of a person in the set of participants for the scene identified in the SCENE attribute. If a match is found, that person's ID is set as HEARER. If no match is found, then HEARER is set to the person who was the last speaker. If there was not last speaker (this is the first utterance of

a conversation), HEARER is left undefined.

If no SCENE attribute is given, the default scene "LOCAL" is used when looking up participants.

MarkTurntaking

How the floor is negotiated (Interaction Management).

Uses	THEME, RHEME
Creates	TURN @TYPE=(TAKE,KEEP,GIVE), GROUNDING @TYPE=(REQUEST,...) @TARGET
Method	Tag all RHEMEs that are at the end of an utterance with TURN of TYPE GIVE and TARGET set to HEARER. If the RHEME is not at the end of an utterance, tag it 73% of the time with GROUNDING of TYPE REQUEST and set TARGET to HEARER.

Tag all THEMEs that are at the beginning of an utterance with TURN of TYPE TAKE. If the THEME is not at the beginning of an utterance, tag it 70% of the time with TURN of TYPE KEEP.

This implements the algorithm presented in (Torres, Cassell et al. 1997)

5.5.3 Domain Knowledge Base

The Domain Knowledge Base (KB), essentially an ontology, is the part of the discourse context that describes the set of things that are likely to be referred to and talked about in the conversation. The KB is in the form of an XML file loaded by the server at startup. The entries in the KB are of three different types: object type, object instance, and action description. Each will be described in turn.

Object Type

Type definitions associate features and their typical values with generic object types. These object types serve as templates for specific object instances, or discourse entities, that need to be recognized in the discourse (usually as noun phrases).

The feature list of an object type is a set of attributes shared by all objects of that type. Each feature is given a descriptive name, such as "cost," "weight" or "color." Features are either numeric or symbolic, the former referring to a feature whose value is described numerically and the latter to a feature whose value is described by any text. For each feature named for an object type, typical or normal values have to be given. This is because

an unusual feature of an object is an important piece of knowledge when generating behaviors that co-occur with the introduction of that object.

The format of a type definition is as follows:

```
type       ::= <TYPE NAME="string" CLASS="class"> { feature }* <TYPE>
class      ::= OBJECT | PERSON | PLACE
feature    ::= symfeature | numfeature
symfeature ::= <SYMFEATURE NAME="string" TYPICAL="typicalsym"/>
typicalsym ::= string{,string}* | ANY
numfeature ::= <NUMFEATURE NAME="string" TYPICAL="typicalnum"/>
typicalnum ::= float{-float}
```

An example of a type definition would be:

```
<TYPE NAME="STAIRS" CLASS="OBJECT">
  <NUMFEATURE NAME="STEPS" TYPICAL="4-30" />
  <SYMFEATURE NAME="SHAPE" TYPICAL="STRAIGHT" />
</TYPE>
```

This defines a generic STAIRS type and names two features that stairs in general share, namely that they have a certain number of steps and that they can be described having a certain overall shape. Typical values have been provided for both features.

Object Instance

Instance definitions describe particular instances of a particular object types. Each instance gets a unique ID that will be used to track references to it in throughout the conversation. In linguistic terms, an instance is a *discourse entity*. The instance definition assigns values to the features listed in the corresponding, and previously defined, object type. The format of an instance definition is as follows:

```
instance    ::= <INSTANCE OF="typename" ID="string"
                          {featurename=featurevalue}* />
typename    ::= name of a previously defined type
featurename ::= feature defined for this particular type
featurevalue ::= string | float
```

An example of an instance definition would be:

```
<INSTANCE OF="STAIRS" ID="STAIRS1" STEPS="15" SHAPE="SPIRAL" />
```

This describes one particular staircase in the world and assigns a unique identifier to it, STAIRS1. Values are given to both features named in the type definition of STAIRS. The first one, STEPS="15" falls within the typical range, but the second value, SHAPE="SPIRAL" identifies an unusual trait that may warrant an iconic elaboration when a reference is made to it in an utterance.

Feature Description and Action Description

These descriptions describe which configuration of the hands and which movement of the arms would visually illustrate a feature or action, either iconically or metaphorically. Each description is associated with a particular lexical value, either some possible value of an object feature

(such as "tall") or an action verb (such as "run"). The format of a descriptions is as follows:

```
description ::= <DESCRIPTION TYPE="gesturetype" VALUE="string">
                    rightarm* leftarm* </DESCRIPTION>
rightarm ::= <RIGHTARM HANDSHAPE="string" TRAJECTORY="string"/>
leftarm ::= <LEFTARM HANDSHAPE="string" TRAJECTORY="string"/>
```

An example of a feature description would be:

```
<DESCRIPTION TYPE="ICONIC" VALUE="SPIRAL">
  <RIGHTARM HANDSHAPE="point_up" TRAJECTORY="spiral" />
</DESCRIPTION>
```

This description can eventually map onto an iconic gesture that is associated with the concept "spiral."

5.5.4 Participation Framework

The participation framework structure is the part of the discourse context that describes the participation status of every person in a particular gathering. Participation status can currently be any of HEARER (ratified), ADDRESSEE (focus of speaker's attention) or SPEAKER. When no one is speaking, a HEARER status is assumed for everyone.

A gathering is the group of people in a found in a particular visual scene that have their role attribute set to "participant" (see visual scene description below). Technically over hearers are also a part of a gathering and may be included in future implementation of participation framework (extending participation status to include non-*ratified* status as well).

When the status is set for a person in the participation framework, the structure automatically updates the status of the other gathering members if necessary. In particular, if person A is currently a SPEAKER and person B gets SPEAKER status, then the person A gets ADDRESSEE status if a new addressee was not named, otherwise a HEARER status. This implements the turn taking rule from the second version of BodyChat (see 2.6).

It is possible to store multiple participation frameworks simultaneously in a special participation framework database. Particular frameworks can then be referred to by the name of the scene in which it is occurring.

5.5.5 Discourse Model

The discourse model is the part of the discourse context that keeps track of the dynamic state of the overall discourse through a discourse history and a visual scene description.

Discourse History

There are two parts to the discourse history. The first part is simply a list of all tagged utterance frames processed so far. Leaving them tagged allows the history to be searched both by lexical items and discourse

85

function. The second part is a recency list of discourse entities. This is a list of discourse entities that have been created during the course of the discourse, with the most recently referred to entity on the top. Only one instance of each entity is allowed in the list, so when an entity is referred to a second time for example, it gets promoted to the top.

Visual Scene Description

The scene description simply associates a scene name with a list of object and person instances that make up important parts of that scene. Each object and person instance must have been defined in the knowledge base so it can be referred to by it's ID. Any reference to a person instance also has a ROLE attribute set, which determines whether that person is a participant in the current gathering or not. Here is an example of an initial scene description file:

```
<SCENE ID="PUB">
    <OBJECT ID="BURGER1"/>
    <OBJECT ID="FRIES1"/>
    <OBJECT ID="PINT1"/>
    <PERSON ID="WAITER1" ROLE="OVERHEARER"/>
    <PERSON ID="PETER1" ROLE="PARTICIPANT"/>
    <PERSON ID="OLAF1" ROLE="PARTICIPANT"/>
    <PERSON ID="NED1" ROLE="PARTICIPANT"/>
</SCENE>
```

5.6 Avatar Agent

5.6.1 Behavior Generation from Frames

After the server has processed and annotated a communicative frame, it is distributed to all connected clients. In each client there is an avatar agent representing each participant in the conversation. One of these avatar agents represents the actor that initiated the frame and the others represent the audience. The actor agent is the first one to receive and process the frame, but then the frame is passed around to all the other agents get that then get a chance to process it as well. The idea is that the ensuing performance incorporates actions performed by the actor as well as automated reactions from the audience.

Behavior Modules and Behavior Generators

An agent processes a frame with a behavior module. A behavior module takes a frame as input, applies a series of behavior generators on it, and provides as output the same frame, but now annotated with newly generated XML tags that describe behaviors. A behavior generator generates behavioral markup as a function of the incoming XML tags, which in this case is the functional markup from the server. Each generator stands for a rule that associates a behavior with a communicative function. The rule inserts behavioral markup where it finds a certain pattern of functional tags.

A behavior generator can be described in an XML transformation language such as XSLT. An example of a simple behavior generator written in XSLT follows:

```
<xsl:transform>

<!-- Nod head on word emphasis. -->
<xsl:template match="EMPHASIS[@TYPE='WORD']" priority="10">
<HEADNOD>
    <xsl:copy>
        <xsl:apply-templates select="@*|node()"/>
    </xsl:copy>
</HEADNOD>
</xsl:template>

<!-- DEFAULT RULE: Any non-matching tags just get copied -->
<xsl:template match="@*|node()">
    <xsl:copy>
        <xsl:apply-templates select="@*|node()"/>
    </xsl:copy>
</xsl:template>

</xsl:transform>
```

This generation rule looks for any tag with the name of EMPHASIS that furthermore has the value of its TYPE attribute set to WORD (see the highlighted "match" expression) and then surrounds that tag with a new HEADNOD tag (see the highlighted tags). The discourse function EMPHASIS is therefore getting realized here through the precisely placed HEADNOD behavior.

A behavior generator can also be written in C++ simply by sub-classing a generic behavior generator. This is useful when the transformation requires more computation than matching on a pattern, for example if the transformation only occurs part of the time as predicted by a stochastic model.

Processing Utterance Frames

The actor agent processes utterance frames with a behavior module called the speaker module. This module contains a set of behavior generators that produce co-verbal behaviors. When the speaker module is finished with the frame, each of the other agents processes it with a behavior module called the listener module. This module produces behaviors that automatically respond to any of the speaker's behaviors, such as by generating visual attention in response to a speaker's visual reference or generating attentive feedback in response to feedback requests.

Discourse Function	Speaker Behavior	Listener Behavior
EMPHASIS @TYPE='WORD'	HEADNOD GESTURE_RIGHT @TYPE='BEAT'	
EMPHASIS @TYPE='PHRASE'	EYEBROWS	
GROUNDING @TYPE='REQUEST' @TARGET={@TARGET}	GAZE @TYPE='GLANCE' @TARGET={@TARGET}	(only if target) GAZE @TYPE='GLANCE' @TARGET={ACTOR} HEADNOD EYEBROWS
CLAUSE @TYPE='EXCLAMATION' or @TYPE='QUESTION'	EYEBROWS	
TURN @TYPE='GIVE' @TARGET={@TARGET}	GAZE @TYPE='LOOK' @TARGET={@TARGET}	(only if NOT target) GAZE @TYPE='LOOK' @TARGET={@TARGET}
TURN @TYPE='TAKE'	GAZE @TYPE='AWAY'	GAZE @TYPE='LOOK' @TARGET={ACTOR}
TURN @TYPE='KEEP'	GAZE @TYPE='AWAY'	
TOPICSHIFT[84%]	POSTURESHIFT @BODYPART='BOTH' @ENERGY='HIGH'	
W[16%+no top.shft in claus]	POSTURESHIFT @BODYPART='BOTH' @ENERGY='LOW'	
REFERENCE @TYPE='TEXTUAL' @SOURCE={@SOURCE}	GAZE @TYPE='GLANCE' @TARGET={@SOURCE}	
CONTRAST[elements=2]	GESTURE_RIGHT @TYPE='CONTRAST1' GESTURE_LEFT @TYPE='CONTRAST2'	
CONTRAST[elements>2]	GESTURE_RIGHT @TYPE='BEAT'	

Table 3: The basic set of behavior generation rules, executed by the speaker and listener avatar agents, that turn functionally marked up messages from the server into an animated performance

Table 3 summarizes the basic set of generation rules that reflect the empirical data presented in section 2.1 on the crucial processes of face-to-face conversation and the nonverbal behaviors that support them. The first

column contains the discourse function markup that is embedded in the utterance when it comes back from the discourse module in the server. The second column contains the speaker behavior markup that gets added in the speaker module as a transformation of the first column by a generation rule. The third column contains the listener behavior markup that gets added in each listener module based on the markup so far. Note that the generation of some listener behavior depends on who is the target of the speaker behavior (All these tags are explained in Appendix A).

After all avatar agents have processed an utterance frame, it is sent into a scheduling pipeline that turns the frame into an animation script where every behavior is synchronized to the production of the utterance words. This pipeline simply consists of the last few modules in the BEAT processing pipeline (Cassell, Vilhjalmsson et al. 2001), starting with the filtering module. The BEAT Pantomime compiler had to be updated to allow multiple synchronized Pantomime scripts to be generated, one for each participant, from a single XML representation. When the Pantomime scripts are received back from BEAT, they are sent into Pantomime where each of them gets interpreted by an avatar puppet for execution.

Processing Action Frames

Action frames are similarly first processed by the actor in an action module and then by the other agents in modules called reaction modules. Again, this allows the agents to automatically respond to an actor with reactive behavior. Unlike with utterance frames, the action and reaction modules don't add annotation to the action frame. Instead, they initiate immediate action in their respective avatar puppets. This is because actions are considered to be single instantaneous events that need not to be synchronized with anything else such as speech. The behavior representing the action and any reaction behavior all occurs as quickly as possible after the original action was initiated.

Action Function	Actor Behavior	Reaction Behavior
GROUNDING @TYPE='AFFIRM'	Headnod Eyebrows	
TURN @TYPE='REQUEST'	Gesture(READY) GlanceAway	Eyetrack(ACTOR)

Table 4: Behavior mapping table. The functions here correspond to the ones shown as the outcome in Table 2. Actor Behavior describes how the avatar representing the user that initiated the action behaves as a result of it and Reaction Behavior describes how other users' avatars behave in response.

Table 4 shows a couple of examples how the actor's avatar agent transforms functional markup from the server (first column) directly to avatar puppet motor commands (second column). The last column shows

the commands sent to the non-actor avatar puppets as the action frame passes through their reaction modules.

5.6.2 Behavior Generation from World Events

Users always initiate frames, either by typing a message or manipulating the interface. The behavior modules allow the avatar agents to properly act upon those frames. However, everything that happens around the avatars doesn't have to come from a user; the world itself can be dynamic. In order for the avatars to preserve the illusion of being fully immersed in the virtual environment, they need to be able to react to it.

The automated reaction to the virtual environment is implemented at two levels. The lower level is at the motor skill level. An example of this kind of automation is the motor skill associated with following a visual target with your eyes. The eye tracking motor skill itself, inside the avatar puppet, has access to the world geometry in order to turn the eyes to face any named target. This skill updates itself at regular intervals, so that the eyes can re-orient themselves if the target moves. Automation at this level is hard-coded into the motor skill and therefore behaves exactly in the same way across all avatar puppets.

At a higher level, each avatar agent runs its own even loop that can receive events from the world and execute arbitrary code in response, with full access to individual user profiles. A simple example of automation at this level is the execution of various idling behaviors, such as scratching one's neck, after the user has not been typing for a certain amount of time. Although all the avatar agents currently share the same set of idling behaviors, it would be trivial to base the selection of behaviors on some user characteristic (other than just their typing action), such as how excited they are (perhaps as measured by their mouse using skin conductivity).

6 MapChat Application

6.1 The Task

The Spark architecture is meant to support a wide range of applications that involve online avatar interaction, especially those where social contact and conversation are primary. However, an application needs to supply specific domain knowledge, so behaviors can refer to it as part of the discourse context, but specifying this knowledge doesn't have to be a difficult task. In fact, many applications already contain resources that can be easily converted into accessible discourse context. For example, an environment that allows architects to discuss planned buildings can provide labeled 3D models as part of the context, or a complex online game world could make its entire database of objects, quests and occupants available as domain knowledge.

Figure 11: Three users collaborating on route planning in MapChat. The animated behavior of the avatars is synchronized with the textual message delivery at the top of the screen as well as with synthesized speech.

Collaborative route planning was chosen as the conversation domain to evaluate the theoretical approach and Spark implementation. Planning a trip is an activity that relies heavily on verbal negotiation as well as a shared visual environment containing a map. An application based on Spark, termed MapChat, was built to allow three people to log into a shared virtual map-room and collaborate on route planning. Each person is represented by an avatar, standing in the center of the room by one of the edges of an instrumented table. The table, sort of a holographic projector, can be loaded with arbitrary 3D maps or scenes. MapChat

automates appropriate nonverbal behavior in the avatars as the users discuss the map display. The map itself feeds into the Spark discourse context by providing domain knowledge in the form of path and landmark descriptions.

To evaluate the quality of the conversation in terms of collaboration performance, a particular route-planning task was designed. The basic task was for the participants to negotiate and choose the quickest way to get from a starting landmark to an end landmark on the map in front of them, using only the supplied roads. Two things were introduced to make the task challenging and rely on good interaction among participants. The first were the various landmarks that sit along the roads on the map. Each of them had a special function that either hindered or opened passages in that spot or other places on the map. Simply looking at the length of the paths was therefore not enough, the actual sequence in which landmarks were visited was also important. The second complication introduced was that each participant was briefed on different landmark and terrain properties before they joined the discussion. Each therefore only had partial knowledge to begin with, but complete knowledge was required to pick the best route. The quality of the discussion was therefore a factor in completing the task. Each participant had to complete the task twice, using a different version of MapChat each time (see description of study in chapter 7.3), so two different sets of maps, roads, landmarks and briefings had to be ready to be loaded into the system.
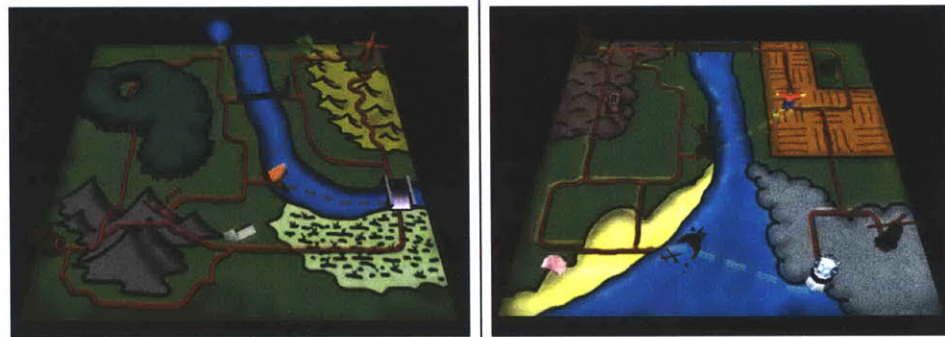
Figure 12: The two 3D maps that were created for the route planning task

6.2 Interactive Map

MapChat required a virtual map-room with an interactive shared map. It was relatively simple to build such a map using Pantomime's flexible world object structure. A new world object was created that responded to mouse clicks on any of its special hot spots (in this case the paths) by changing the color of the clicked geometry (the paths become bright red) and calling an event handler outside Pantomime[5]. The Spark World

module catches the event and creates an action frame specifying which path has been selected and who selected it. Once this frame reaches the other clients, the maps get updated everywhere to reflect the new selection.

6.3 New Behaviors

While the theoretical model explains what sorts of nonverbal behaviors are important and how they relate to the underlying conversation processes, it does so in general terms. Once a domain has been picked, it is therefore important to take a closer look at some of the supporting behaviors and refine the model.

MapChat's conversation setting was easy to set up in the physical world, and gathering data on how three people behave when planning a route in each other's physical presence was a matter of installing proper video and audio capture equipment. Video and audio data from 6 minutes of collaboration were transcribed using the Anvil multi-modal transcription system (see Figure 13).

Figure 13: Three people solve the route planning task face-to-face, the speaker points while the others attend with gaze

Two of the most commonly observed behaviors, not fully predicted by the basic model, were looking at the map and pointing at features on the map. The modularity of Spark made the process of adding these behaviors to the already rich basic set of behaviors straightforward. In addition, the selection of paths chosen as a part of the solution and the idle behavior the subjects engaged in when doing nothing had to be handled by MapChat. What follows is the description of these phenomena and how they were incorporated into MapChat using the Spark framework.

6.3.1 Looking

The most striking characteristic of the subjects' behavior while working on the task was that they rarely took their gaze off the map in front of them. By default they were staring at the map and only bothered looking up at the others when something was going on, such as when listener feedback was being requested from them with a direct gaze (the light blue bars represent gazing at the map in Figure 14 and the green represent gazing at each other).
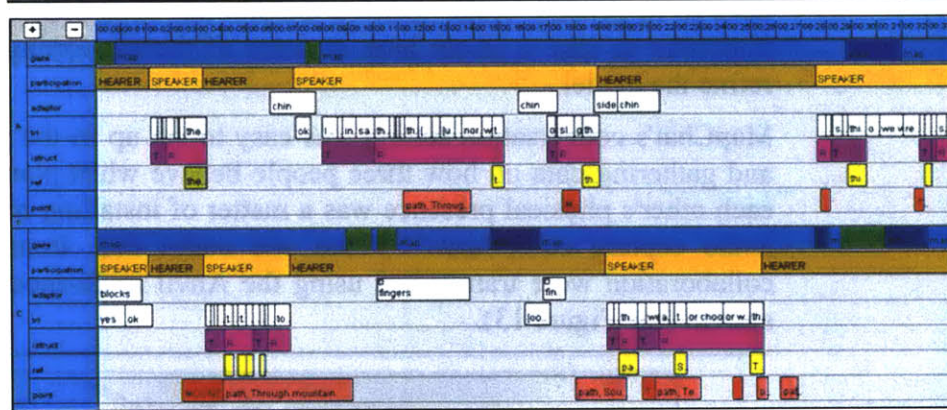
Figure 14: The first 30 seconds of annotations for 2 of the 3 participants. From top to bottom the tracks indicate gaze direction (green=other person/blue=task), role (brown=listener/orange=speaker), idle behavior, speech, information structure (dark purple=theme/light purple=rheme), verbal reference (yellow) and pointing (red and pink).

Another place where the subjects frequently looked was their sheet of notes describing the background information they each had about the landscape in front of them. The action of looking at the notes was very clear as they had to raise the sheet up to a comfortable reading distance.

This gaze behavior got modeled in MapChat by introducing an action frame of type ATTENTION. This frame sets the default resting position for any gaze behavior. The avatar agents make sure they return the gaze to this position after any interruption they may experience. The ATTENTION is first set to TASK when users log on. If a user decides to bring up their notes on their screen (they could do so by pressing the TAB key), an ATTENTION frame with the target of NOTES is sent out. The acting avatar reacts to this frame by bringing out a small notebook and resting its gaze on it instead of the task. Attending to the notes cannot be interrupted (the subjects often seemed to withdraw from the conversation when reading their notes), but while attending to the task, any communicative gaze behavior can override the task gaze, for example when turns are exchanged. The new entries in the action mapping table and the behavior mapping table to reflect these states and associated behavior are summarized in Table 5 and Table 6.

User Action	Action Communicative Function
Pressing TAB	ATTENTION @TARGET='NOTES'
Depressing TAB	ATTENTION @TARGET='TASK'

Table 5: New entries in action mapping table (see Table 2). These bind the key that brings up user notes to a change in attention. Default attention is given to the task.

Action Function	Actor Behavior	Reaction Behavior
ATTENTION @TARGET='NOTES'	ReadNotes	(none)
ATTENTION @TARGET='TASK'	LookAtTask	(none)

Table 6: New entries in the behavior mapping table (see Table 4). These generate avatar behavior reflecting the focus of attention.

6.3.2 Pointing

Mention of landmarks was often accompanied by a pointing gesture. Notice in Figure 14 how the verbal references (indicated as yellow blocks) seem to be generally preceded by pointing gesture (indicated as red blocks). In fact out of all verbal references to specific landmarks (a total of 121 occurrences), 40% occurred within 2 seconds of a pointing gesture towards that same landmark.

It seems that these behaviors relate to the process of information management, as the looking and pointing occurred in close proximity to the reference to discourse entities that are being visually evoked (Prince 1981). When pointing occurs, everyone present would typically also glance toward the pointing target. The rules to provide this behavior in MapChat are summarized in Table 7.

95

Discourse Function	Speaker Behavior	Listener Behavior
REFERENCE	GESTURE_RIGHT	GAZE
@TYPE='VISUAL'	@TYPE='DEICTIC'	@TYPE='GLANCE'
@ID={@*ID*}	@TARGET={@*ID*}	@TARGET={@*ID*}
	GAZE	
	@TYPE='GLANCE'	
	@TARGET={@*ID*}	

Table 7: A new entry in the behavior mapping table (see Table 3) describing the generation of pointing gesture, and associated glances in speaker and listeners, as a result of visual evocation of a discourse entity

6.3.3 Selecting Paths

In the face-to-face situation, the subjects were asked to place small movable markers on top of path segments they wanted to choose as part of the final solution. With the interactive map in the virtual environment, the users click on the paths with their mouse. To provide a visual cue to who is making the selection, the actor's avatar reaches out and points at the segment being highlighted. The rules to add this behavior are summarized in Table 8 and Table 9.

User Action	Action Communicative Function
Selecting part of map	SELECTION
	@TYPE='MAP'
	@TARGET='*TARGET*'

Table 8: A new entry in action mapping table (see Table 2). This binds a mouse click on the interactive map to a selection action shared with the others

Action Function	Actor Behavior	Reaction Behavior
SELECTION	GlanceAt(*TARGET*)	GlanceAt(*TARGET*)
@TYPE='MAP'	PointAt(*TARGET*)	
@TARGET='*TARGET*'		

Table 9: A new entry in the behavior mapping table (see Table 4). This generates a pointing and glancing behavior in both speakers and listeners as a result of path selection

6.3.4 Idle Behaviors

During the course of the face-to-face collaboration, a good amount of time was spent just looking at the map in silence. Even though no-one was saying anything, the bodies were far from motionless. It became clear that it was important to pay attention to the nonverbal behaviors that were not

tied to a communicative event, because otherwise you'd end up with animating short bursts of realistic behavior connected by moments of unrealistic stillness.

These non-communicative behaviors have often been called idle behaviors, because the people are in an idling state while executing them, or self-adaptors, because these behaviors often involve touching your own body or clothing in an apparent attempt to fix the appearance or make oneself more comfortable. From the video of the face-to-face map conversation, 12 distinct idle behaviors were found across all participants. These behaviors got repeated over and over and can therefore be considered a sort of a basic palette of idle behavior in this conversation. Out of the 12 behaviors, 6 were chosen for animation because they did not involve interaction with clothing or objects in the environment that would be hard to model (see Figure 15).

Rubbing nose   Rubbing chin   Rubbing side

Rubbing back   Rubbing fingers   Rubbing neck
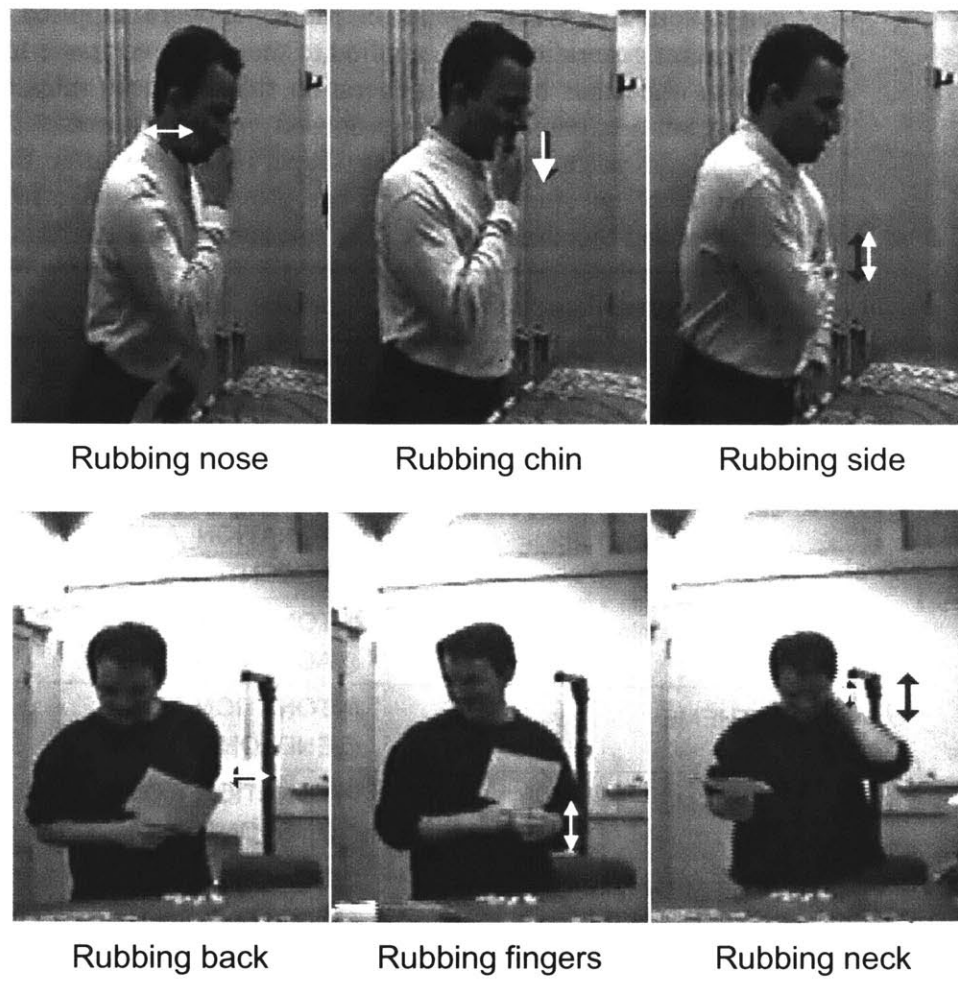
Figure 15: Idle behavior observed in the videotaped data that were then turned into animation sequences. These sequences were then automated by the avatar agents.

Key frame animation was created for these behaviors and they stored with the avatars. The avatar agents were then programmed to automatically

97

execute a random pick from this set whenever they had not been doing anything else for a little while.

6.4 Speech and intonation

MapChat uses the Festival speech synthesizer from the University of Edinburgh to synthesize and play back the messages (when speech is turned on). This speech synthesizer provides word timings that can be gathered before the utterance is played. This allows the system to coordinate the nonverbal behavior as well as the word timings of the text display with the actual spoken words.

Using speech required that proper intonation would be generated so that it would reflect the same communicative intent as was displayed through gesture. Any inconsistencies across modalities would have created confusion. Therefore, instead of relying on Festival's own intonation generation, the avatar agents in MapChat were supplied with a set of intonation generation rules, applied to the delivered frame in just the same way as the other behavior generation rules. These rules implement the intonation assignment rules proposed in (Hiyakumoto, Prevost et al. 1997). In particular, the rules shape intonation after the information structure of an utterance, which has proven to be an effective approach (Prevost and Steedman 1994). The rules are summarized in Table 10.

Discourse Function	Speaker Behavior
CLAUSE	INTONATION_BREAK @DURATION=0.5
NEW within CONTRAST	INTONATION_ACCENT @ACCENT="H*"
NEW within THEME	INTONATION_ACCENT @ACCENT="L+H*"
NEW within RHEME	INTONATION_ACCENT @ACCENT="H*"
THEME	INTONATION_TONE @ENDTONE="L-H%"
RHEME	INTONATION_TONE @ENDTONE="L-L%"

RHEME within
    CLAUSE
    @TYPE='QUESTION'
	INTONATION_TONE
	@ENDTONE="H-H%"

Table 10: New entries in the behavior mapping table (see Table 3) describing the generation of intonation cues for the speech synthesizer when the speaker's message gets produced

98

6.5 Heads-up Display

A heads-up display was created on top of the world display window to support the entering and displaying of messages, as well as system notification messages and the display of the specialized hints for each participant.

Entering messages is done on a single line at the bottom of the screen. When the line is full, it starts scrolling off to the left. The message display is at the top of the screen and displays an entire message in as many lines as are needed on a semi-transparent background in the color of the user that sent the message. The message appears one word at a time to simulate speech and so that the words could be synchronized with the nonverbal behaviors of the avatars. The appearance of each word can be timed exactly by providing an array of word timings. When a word first appears it flashes bright then takes on a slightly lower intensity. After a certain time passes, the word fades out, ultimately wiping out the entire message[6].

6.6 Camera

Three function keys on the keyboard are mapped onto providing views of the virtual environment from three different cameras. The first camera is a first person view, right from the eyes of the avatar representing the user at that client. The second camera provides a view right over the shoulder of that avatar and out into the area in front of it. This view is meant to show yourself along with the people you are meeting. The third camera provides an overhead shot.

---

[6] Thanks to UROP Jae Jang for helping with this

7 Evaluation

7.1 Technical Evaluation

MapChat was built as an instance of the Spark architecture for the purpose of evaluating the architecture itself as well as the theory behind it. It is therefore important to look at how well the architecture survived the MapChat realization from a technical perspective. What follows are some of the issues encountered.

7.1.1 Performance

Time delay

The time from the sign that someone is about to speak, triggered by the first keystroke of a new message, to the actual animated delivery is far longer than would be natural in a face-to-face conversation. Two things contribute to this delay. The first is the time it takes a person to type the entire message. The second is the time it takes the system to process the message and produce the resulting animation.

The first has to do with the messaging medium itself and is shared with any text-based system. While users of such systems accept it, this clashes with the face-to-face paradigm pursued by this thesis. To get rid of this artifact would require speech input, which will be discussed in the section on future work.

The processing time is a problem with MapChat itself. This time can be so long that the users, who are familiar with typical text messaging, have complained and said it reduced the practicality of the system in its current form (see 7.3.5). The time is proportional to the length and complexity of the message and can range from 2 seconds (a short "ok") to 8 seconds (several clauses) in a three-person conversation, where the processor speed of both clients and server is about 1 GHz.

Processing takes a long time mostly because of the huge number of operations performed on the XML tree structure as it is being transformed along the way through the system. One could argue that this is a result of Spark's pipeline type architecture. Efficiency could be increased by parallelizing sets of operations such as behavior generators or re-using operations for multiple purposes such as by combining language-tagging methods. But efficiency would be gained at the cost of reduced flexibility, an essential feature of Spark. The hope is therefore that better XML implementations and faster machines in the near future will alleviate the problem. In the meanwhile, some tweaking of the system without disturbing the architecture is possible. For example, the architecture does not prevent listener agents from adding reactions in parallel.

Another significant factor in the processing delay is the time it takes to synthesize the voice and retrieve word timings. This may again become less of an issue with faster computers, but it is also an artifact of having to deal with text input. Along with the typing delay, this may be addressed by using spoken input instead as will be discussed in future work.

Message length

When MapChat was first put to the test, it quickly became apparent that the system was only able to handle messages under a certain length. Longer messages would lead to structures so large that some of the internal buffers would overflow. For example, the sheer amount of behavior description generated, including all the visemes, could easily produce animation scripts so large that Pantomime's command buffer overflowed. It is of course possible to increase the size of those buffers, within system memory limits, but perhaps there is a more compact way to represent the annotations and behavior commands. Compacting the format may compromise human readability however.

Message queuing

In Spark there is no such thing as simultaneous frames. Even when two frames arrive at the server at the same time, one is placed on the processing queue after the other. While the server processes the first one, the second frame and any other frames arriving at that point have to wait their turn. This adds to the processing delay, but perhaps more importantly, it removes a possible synchrony between events. For example, if two users pressed their positive feedback buttons at the same time, one avatar would start nodding before the other one (though the time difference is hardly noticeable). This may be turned into a feature, however. Utterance frames could be kept in the queue until the previous utterance frame has been completely delivered to prevent overlapping messages. This "feature" is in effect now when the delivery time of the first utterance is shorter than the processing time for the second utterance.

7.1.2 Flexibility

One of the main design goals for Spark was to provide a flexible way to build applications based on it. Implementing support for collaborative route planning in MapChat successfully tested this flexibility. Apart from special user interface features such as a heads-up display and special interactive objects like the map, it proved quite trivial to add new function tags to the frames and related behavior generators.

7.2 Model Evaluation

When a theory of human behavior is implemented, many things can happen that distort or simplify the model, including the very limitations of what is computationally possible today. To get a sense for how well the

MapChat implementation mimics real face-to-face conversation, and to understand better what problems remain, several turns from the videotaped face-to-face route planning exercise (see 6.3) were run through MapChat and the nonverbal behaviors compared.

It was not expected that the behaviors would match exactly, simply because the model represents behavior averaged across many populations and therefore it would not capture certain idiosyncrasies. Furthermore, the model aims to predict all appropriate behavior, which essentially translates into the highest plausible activity, and was therefore expected to produce more behaviors than typically observed.

7.2.1 Data

The first 40 seconds of the videotaped face-to-face route planning exercise were annotated with onsets and durations of the following behaviors: beat gestures, pointing gestures, gaze direction and head movements. An attempt was made to annotate eyebrow movement, but the grainy quality of the video prevented an accurate estimate. The annotations were written down on a grid, termed a *dope sheet*, with the behaviors represented by rows and each column corresponding to a word being spoken (see Table 11).

During the 40 seconds, 8 utterances were exchanged. These utterances were typed into MapChat in the same order they occurred in the face-to-face situation, and the output captured on a dope sheet. The analysis involved comparing the two dope sheets, behavior by behavior. The entire set of dope sheets is provided in Appendix B.

A	Ok	I	recall	my	instructions	saying	that	we
Head		Nod		Nod			Nod	
Gesture							BEAT	
Gaze	Map	C	Map					
Posture								
B								
Head								
Gesture								
Gaze	Map				A		Map	
Posture								
C								
Head								
Gesture								
Gaze	Map				A		Map	A
Posture								

Table 11: A part of the face-to-face conversation represented on a dope sheet. The speaker A nods on a few words, beats on one word and glances at listener C once. The listeners glance at the speaker a couple of times.

7.2.2 Observations

Overall

A couple of things were immediately obvious. The first was that MapChat produced about three times as many emphasis and feedback behaviors as there were in the face-to-face data. As explained above, excess behaviors were expected, but this is an indication that perhaps there should be a way to tone emphasis down for certain people or circumstances. The other observation was that, even during conversation, people rarely looked at each other's faces but were either looking at the map or reading their notes. MapChat produced a lot steadier mutual gaze. Several things may be going on here. Reading the notes was not a behavior generated by MapChat, but was instead an explicit action that a user could take, so this behavior was not expected to match at all. As for gazing at the map, the amount of looking at task during conversation and how that affects typical gaze patterns during turn-taking has not been studied (to my knowledge), but it appears that MapChat's regular group turn-taking rules may not apply. In fact, a close observation of the video reveals that looking at each other's hands on the task surface may indicate turn-taking. For example, as a participant moves her hand into position to talk about something on the surface, the fact that the other participants now rest their gaze on her hand, may ratify her as the next speaker.

Observing the excess of generated behavior only tells part of the story about the success of the model. If the model is really good, it should be able to explain or predict each behavior that occurred in the face-to-face data. When a face-to-face behavior occurs that cannot be explained at all by the model, then that should be looked at more closely. In particular one can think about whether the lack of prediction stems from something that is missing from the model and could be added with little trouble, or something that won't be computed given current or even future technology.

Head nods

Out of 14 head nods in the face-to-face data, 7 were exactly predicted and 5 more occurred within a word or two of the predicted spot. This must be considered a very good fit. Only two head nods happened that did not seem to be explained by the model. The first happened as a speaker briefly looked at a listener while saying "I recall" in sort of an "in character" voice. It has been suggested that head movements can indicate shifts in voice (McClave 2000), but this is not being modeled in MapChat and it is not clear how it could be done (quotation marks could be used as explicit signals though). The second unexplained head nod was a feedback head nod from a listener that was not expected to respond to the speaker's request for feedback. This could have been predicted with a more complete model of listener feedback behavior in groups.

Pointing

Out of the 10 pointing behaviors in the face-to-face data, 3 were exactly predicted, other 3 did not happen exactly at the predicted time and 4 were not predicted at all. All three of those that were not timed right, basically preceded the verbal reference to the same objects by a few words. It seems as if the thought and gesture were completed before the spoken realization was finished. It is not clear how this could be modeled, especially since this only occurs some of the time and is not a fixed offset. In one of these cases there was actually more going on than just the delay. The gesture was showing a path from one landmark to another, and only the second one was mentioned verbally. There was therefore no way that a system could have figured out where to start tracing the finger on the map.

Similarly, three of the remaining four pointing behaviors that were not predicted at all, were missed because the system could not resolve a verbal deictic reference such as "there" or "this one." In the face-to-face situation, participants would use those words because they could actually point. In MapChat users would not use such ambitious reference unless they would accompany it with a mouse click to highlight a path on the map (resulting in a pointing gesture as well), relieving the system of having to figure this out automatically. The last pointing behavior not accounted for is an interesting one because it was actually a listener pointing at the same time as the speaker. It is possible that this may have been an attempt by the listener to take the turn from the speaker. That listener did in fact continue pointing until the turn was hers and she started speaking. Again this may be an indication of an alternate turn-taking model that should take into account the manipulation of a task surface.

7.2.3 Summary

MapChat did well, predictions were good more than half the time, but there are several areas where having more data available and sticking it into the model could have produced even better results, perhaps pushing the accuracy into the 80th or 90th percentile. In particular having data on how turn-taking is accomplished in a task setting, especially with regard to gaze and gesture, would have improved things. However, there were also things that were hard for the model to do, such as inferring what path is being discussed when only parts of it have been mentioned. It may be possible to reason about what is being said; for example, if a participant says "if we go around the mountain and then across the swamp to the bridge" the system could use its knowledge of the map to find a unique path that fits all the criteria mentioned. MapChat does not do this currently, but the problem of reasoning about spatial descriptions is in the realm of doable things, and such algorithms, and representations of paths as relationships between objects, could be added to the model when they become available. Finally, there are things that the model will never be

able to do, and perhaps the verbal deictic (for example "that") serves as a good example. If the listener needs to fill in things that are not at all found in the discourse context available to the system, then there is no hope to making an accurate prediction.

It should also be kept in mind that the domain of conversation could have impact on how well the model is able to predict the behavior. The quality of prediction could range from very low to very high. For example there were no iconic or metaphoric gestures found in the route-planning data. These are perhaps the most complex class of gestures because they directly depict features of objects or actions, or represent purely mental concepts. The model does provide a way to build knowledge about these objects and concepts into the discourse context, but the question is how much work it would take to predict the kind of rich gesturing often found in oral storytelling for example. It remains as future work to test out the model against these other domains.

7.3 User Study

7.3.1 Overview

This chapter reports on a user study conducted with MapChat to evaluate the implementation and test the hypotheses about the model. The first section outlines the study design and procedure. Then section 7.3.3 describes the gathered data and introduces both the subjective and behavioral measures taken. Section 7.3.4 presents the results from analyzing the data. That section is organized according to the claims the results support. Section 7.3.5 rounds up the chapter by presenting results regarding overall clarity and perception of the avatars and messages, as well as summarizing free form comments from the subjects.

7.3.2 Design and Procedure

Groups of three subjects used MapChat to solve a route planning task. The study was a 2x2 design where one of the independent variables was the presence/absence of the avatar. The other independent variable was scrolling text versus synthesized speech for message output (see Table 12). This was a within subject study, where each subject was assigned to two conditions, with a difference in only one of the factors, at random[7].

---

[7] It turned out that this is not a typical "within subject" design and it led to some more involved analysis described later.

	AVATARS = 0	AVATARS = 1
SPEECH = 0		
SPEECH = 1		

Table 12: The four conditions tested in the study. Two factors, speech and avatars, with two levels each. Only the 1st person view was used in the study.

Three computers were located so that no direct visual contact between subjects at each computer was possible. The computers were networked and ran MapChat clients that connected to a server under the supervision of the experiment supervisor.

Each participant was led to separate client stations where they were logged into a version of the client chosen at random for the group (if members had already done one session, then the version chosen had to take their previous condition into account to ensure each member was experiencing a new session that only differed by one factor level). The client first displayed the special briefing intended for each participant separately. Once the participants were done reading the special briefing, they could dismiss the briefing screen and proceed with the task (described in section 6.1). The briefing screen could be recalled at any time by pressing the TAB key. The participants were given as long as they needed to solve the task. When participants were ready to commit to a solution, they each had to press the F12 key on their keyboards, which would then remove their avatar from the room (if they were in an avatar condition) and prompt them to start on their questionnaire. If this was the second time a subject was a participant, using a different version of the client than before, then they got a second questionnaire that asked them to compare the two experiences.

7.3.3 The Data

Subjects, Sessions and Trials

50 subjects were brought in for the study. They were all users of online text-messaging systems (AOL Messenger being the most popular) and native English speakers. There were 29 males and 21 females, ranging in age from 17 to 45. The largest portion of the population, or 30 subjects, consisted of 18-23 years old college students.

All 50 subjects were signed up for 2 MapChat sessions, held on two separate days. A session consisted of three subjects coming together to solve a single route-planning problem, either with map 1 or map 2 (see Figure 12). Each session was randomly assigned to one of the four conditions (AS, AT, NS, NT). A total of 39 sessions were conducted, of which 8 were disqualified because of technical difficulties, leaving the data from 31 sessions to be analyzed. The number of valid sessions per condition is shown in Table 13 and the number of sessions contributing to the AVATAR and SPEECH factors are shown in Table 14.

	AVATAR = 0	AVATAR = 1
SPEECH = 0	9 (NT)	7 (AT)
SPEECH = 1	7 (NS)	8 (AS)

Table 13: Session per condition

		POOLED	GROUPED	
FACTOR	LEVEL	N	MAP 1	MAP 2
AVATAR	0	16	8	8
	1	15	7	8
SPEECH	0	16	9	7
	1	15	6	9

Table 14: Sessions per factor level, as well as per task

Treating each subject's use of the system as a trial, the 31 sessions produced 84 valid trials. The number of trials is less than 93 because 9 times "fill-in" subjects were used (because recruited subjects did not show up), who did not know the task or the nature of the experiment, but were acquainted with the experimenter. Questionnaire data from these 9 subjects were not used to avoid a possible personal bias.

Subjects were assigned randomly to sessions, while making sure they would not solve the same task twice, so the sessions contained a mixture of first-time and second-time subjects. The number of second-timers, essentially the level of system experience of a group, was recorded and was used as a covariate where order-effect could have been expected. 20 subjects ended up doing only one valid trial, because they failed to show up for their second trial, or a trial had to be disqualified (because the session was disqualified or because of scheduling errors). The total of 84 trials therefore contains 32 trial pairs and 20 single trials, balanced across all conditions. The number of trials per condition is shown in Table 15 and how these trials contribute to the two factors AVATAR and SPEECH is shown in Table 16.

	AVATAR = 0	AVATAR = 1
SPEECH = 0	22 (NT)	19 (AT)
SPEECH = 1	19 (NS)	24 (AS)

Table 15: Trials per condition

FACTOR	LEVEL	N
AVATAR	0	41
	1	43
SPEECH	0	41
	1	43

Table 16: Trials per factor level

Each session as a whole provided three sources of behavioral data: The solution chosen by the group, the time it took and the log of all messages and actions taken during the group's discussion. Each trial also provided self-report data in the form of a Trial Questionnaire (see Appendix C) where the subject rated the various aspects of their experience using the system, solving the task and working with others. A Preference Questionnaire was also provided after a trial that completed a trial pair, comparing the two conditions experienced by the subject (see Appendix C).

Measures

The Trial Questionnaire was divided into 6 sections. The first section asked subjects to rate their overall experience using the system. The second section asked the subjects to rate the communication experience along various dimensions and addressed the hypothesis that avatars improve the process of online conversation. The third section asked the

subjects to rate their impressions of the other subjects they worked with and addressed the hypothesis that avatars improve the social outcome of the online conversation. The fourth section asked the subjects to rate their impressions of how well they did on the task and addressed the hypothesis that avatars improve the task outcome of the online collaboration. The fifth section asked condition specific questions, such as how natural the speech sounded. The last section was for free comments.

There were 42 questions in the first four sections of the questionnaire. By logically grouping the questions around the dimensions they measure, the 42 questions were aggregated down to 18 dependent variables. How these contribute to evaluating the process hypothesis and the two outcome hypotheses (task and social outcome) is summarized in Table 17 through Table 19. The self-report measures are marked with an "s" in the type column.

The transcripts of each group's interaction were a rich source for behavioral measures that described the quality of the conversation process for testing the hypothesis that the avatars improve it. The measures that were picked represent standard ways of analyzing conversation and many have been employed in previous studies looking at the quality of video conferencing for example (Whittaker 2002). The focus is on discourse structure management, interaction management, and information management. The process of awareness and engagement management is not tested in MapChat because when participants start interacting they are already committed to the collaborative activity, there is no negotiation involved. However, this process and the advantages of automating this behavior in avatars was extensively studied in the BodyChat experiment (Cassell and Vilhjalmsson 1999). These measures are summarized in Table 17. Two more behavioral measures contribute to testing the hypothesis that avatars improve task outcome and are explained in Table 18.

The Preference Questionnaire, presented to subjects after they had completed two trials, asked them to rate their preference for the systems they tried with respect to six different overall qualities: how useful, how much fun, how personal, how easy to use, how efficient and how easy it was to communicate with the system. Lastly they were asked which system they would use again, with "both" a possible answer. The results from this questionnaire are summarized separately.

Process Measures:

Quality of conversational process	Type	Explanation
Information Management:		
Utterances dedicated to grounding	b	A lot of explicit grounding acts, where subjects double-check to see if everything is being correctly understood, indicates a poor channel.
Interaction Management:		
Number of hints shared	b	Each subject had 5 unique hints to share with everybody for solving the task. The conversation process may be broken if sharing is not taking place.
Equality of participation	b	The difference in number of utterances submitted by the most active and least active participant is an indication of how well the process supports equal access to participation.
Amount of explicit handovers	b	Ending a turn with a direct question, tag questions, or by naming the next speaker, are ways to explicitly ensure a smooth turn transition – something that typically is handled by nonverbal cues in face-to-face conversation.
Overlapping utterances	b	Utterances delivered at the same time cannot be properly read/heard and therefore overlaps should be avoided. Proper turn-taking process should help.
Discourse structure Management:		
Amount of broken adjacency pairs	b	An adjacency pair is a pair of utterances where the first utterance needs a second one as a reply. The process is broken if a request is not paired with a relevant response.
Amount of on-task utterances	b	Staying on-topic is important for solving the task and the quality of the process should contribute to a shared focus of attention.
General:		
Others apparent ability to communicate	s	A subjective measure of how well subjects feel the other participants are able to communicate with them.
Your ability to communicate	s	A subjective measure of how well subjects feel they themselves are able to communicate with the other participants.
Sense of control over conversation	s	A subjective measure of how much the subject feels in control of the conversation. This is to see if the lack of explicit control over the nonverbal behaviors might reduce the sense of a good process.
How close to face-to-face	s	A subjective measure of how close the conversation felt to a face-to-face conversation. The assumption is that the quality of the face-to-face conversation process is higher than a typical online conversation.

Table 17: The measures that describe the quality of the conversation process and a brief explanation of each. Type refers to either a (b)ehavioral measure or a (s)elf-report measure.

Outcome Measures:

Quality of task outcome	Type	Explanation
Quality of solution	b	The task was to find the quickest path. The path that a group chooses can be rated according to how close it is to the optimum path.
Task completion time	b	In conjunction with a good solution, how quickly the group arrives at that solution is a behavioral measure of how efficient the group was.
Feeling of task difficulty	s	A subjective measure of how difficult the participants felt the task was may indicate the presence of problems when solving the task.
Feeling of group efficiency	s	A subjective measure of group efficiency indicates how engaged the group was in solving the task.
Feeling of consensus	s	A subjective measure of consensus indicates how satisfied everyone as a group was with the solution and how well they worked together.
Satisfaction with solution	s	A subjective measure of how satisfied a particular subject was with the solution reflects the confidence in the task outcome.
Comparison with face-to-face	s	Asking subjects how much better they would have solved the task face-to-face provides a comparison with an optimal situation.
Comparison with text chat	s	Asking subjects how much better they would have solved the task using a regular text-chat provides a comparison with the type of system being improved upon.

Table 18: The measures that describe the quality of the task outcome and a brief explanation of each. Two measures are behavioral measures and six are self-report measures.

Outcome Measures:

Quality of social relationship outcome	Type	Explanation
Rating of each person's effort to collaborate	s	A measure of what a subject thinks of each member on the team regarding the effort put in. This may be indication of willingness to work together in the future.
Rating of trust in each collaborator	s	A measure of what a subject thinks of each member on the team regarding trustworthiness. This may be indication of willingness to work together in the future.

Table 19: The two self-report measures that describe the quality of the social outcome.

Analyzing the Trial Questionnaire

The N in Table 16 reflects pooled data, i.e. all trials regardless of which map was being used or whether this was the second or first of the two trials (or the only trial). Both maps and order were balanced across conditions, in the hope that pooling would in fact provide a reasonable data set.

The fact that 62 trials are within-subject and that the within subject condition pairs are not all the same, detracts from the validity of the pooled set. No model for statistical analysis exists for this design, so the data had to be treated as if it came from independent between-subject trials. This is of course an approximation that ignores two possibly critical factors: which map a subject was using (two maps were provided to make within-subjects possible) and whether the subject had used the system before in a previous trial.

The analysis that follows looks at the pooled data, but also examines the data for each map separately and each order-of-use (first or second time MapChat user) separately. Note that dissecting the data by map or order produces sets of data that are completely between subject and thus do not violate the independence of subjects assumed by the statistical methods used.

If a significant result for a particular dependent variable was found in the pooled data, the validity of that finding is supported if the same analysis, applied to the portion of subjects that experienced each map and the portions divided by level of experience, showed an agreeing trend. If the smaller portions do in fact show the same significant results, the support is extremely strong. But the same amount of significance was unlikely since the N is halved for the divided portions (see Table 20). If any portion shows an inverted trend or a significant inverted result, the result from the pooled data can be questioned or dismissed. Also, if the results from the two map groups are found to be significantly different from each other, or the results from the two levels of experience are found to be significantly different from each other, then the pooling of the data for the dependent variable in question is not justified because the data seems to represent significantly different populations. However, if the trend in both populations agrees strongly with the pooled data, this overall trend should be noted.

		Grouping 1		Grouping 2	
FACTOR	LEVEL	MAP 1	MAP 2	ORDER1	ORDER2
AVATAR	0	17	24	21	20
	1	16	27	23	20
SPEECH	0	19	22	24	17
	1	14	29	20	23

Table 20: The Ns for the grouped data by map and experience (order)

No special consideration is needed when analyzing the Preference Questionnaire. That questionnaire was only issued once for each completed pair of trials and because the analysis does not have make use of the pooled set described above, the results are untainted.

Analyzing Behavior Measures

As with the first questionnaire measures, there was no good statistical model that incorporated the fact that 62 of the subjects participated in two sessions. The pooled data (see Table 14) was used as an approximation, where each instance of a subject's participation is treated as being independent. However, when the sessions belonging to each task are analyzed separately, the independence assumption holds since no subjects repeated the same task. In light of this, and also in order to examine possible differences in the tasks themselves, the results from analyzing the behavior data are reported both pooled and divided by task.

7.3.4 Core Results

This study was designed to compare the standard way of having conversations online, i.e. text messaging, with the new Spark-augmented way. The former lacks many of the cues that typically support the crucial processes of conversation as outlined in 2.1, while the latter supplies those cues automatically through the animated avatars. The study was designed in such a way that it would provide behavioral evidence concerning the quality of the conversation, as well as assessments of the experience by the subjects themselves. The expectation being that this quality would improve as a result of the support that the avatars would be giving.

Two indications of conversation quality were considered in the study; the first was how the conversational conduct itself unfolded and the other what the conversation accomplished. These can be considered first and second order effects where a better process would likely lead to a better product. The avatars were expected to improve both effects. While the quality of conversation measures addressed how effective the mediation was, it was also important to evaluate whether the novel avatar interface

introduced any new overhead that would have distracted from the experience. This was addressed through a separate set of questions. It was expected that since the avatars were fully automated, the subjects would not experience any additional overhead.

Ideally the avatar condition would have perfectly generated speech because in a face-to-face paradigm gesture is synchronized with speech not text. However, it was clear that current text-to-speech technology for free-form conversation would not be able to provide completely naturally sounding voices. Therefore speech was treated as a factor in the study design, and the avatars evaluated both with and without synthesized speech. It was expected that if the voices proved very good, then the avatars would be the most effective in the speech condition. But in case the voices proved poor, the effect of the avatars themselves could still be isolated. As it turns out, the speech quality was so poor that the text outperformed the speech in just about every measure, even when combined with the avatars. The effect of the speech itself is not of particular interest here and won't be discussed further unless it interacts with the presence of avatars in some way.

After a section summarizing order and task effects and a section describing the overall user preference, the next four sections will report on the findings of the study with regard to the impact of the animated avatars on the mediated conversation. The first section will focus on the conversation process itself. This section tests the process hypothesis introduced in the model section (3.2.1):

Hypothesis 1: process hypothesis

> Compared to synchronous text-only communication, adding avatars that automatically animate the nonverbal behaviors that in face-to-face conversation support (a) Awareness and Engagement Management, (b) Interaction Management, (c) Discourse Structure Management, and (d) Information Management, will improve the overall process of conversation.

The next two sections will look at the product of the conversation both in terms of how well the participants were able to solve their task and how the conversation contributed to the way they related to each other afterwards. These sections test the first and second part of the outcome hypothesis introduced in model section (3.2.1):

Hypothesis 2: outcome hypothesis

> Compared to synchronous text-only communication, adding avatars that automatically animate the nonverbal behaviors listed in hypothesis 1, improves the (a) task outcome and the (b) social outcome of the online conversation.

115

After the process and outcome hypotheses have been discussed, the next section will then take a look at the more general interface overhead question.

Overall Preference

Of the 31 subjects that completed two trials, 14 subjects experienced two conditions that differed only by whether avatars were present. All of these subjects completed the Preference Questionnaire (see Appendix C), which asks them to rate the strength of their preference for the "no avatars condition" vs. the "avatars condition" along 7 dimensions. The following charts are histograms showing the preference scores given by the 14 subjects. Negative scores, on the horizontal axis, denote preference for "no avatars" and positive scores for "avatars" while 0 indicates no preference. One-tailed t-tests show that the preference for avatars is significant ($p<0.05$) in all questions except for number 5 (it was tested whether the means were greater than 0 = no preference). When asked which system the subjects would want to use again, everyone except one, choose an avatar system or both systems.

Every single subject said that the avatars were more fun and more personal than the text version. The fun factor could well be due to the novelty of the interface. The fact that the avatars were deemed more personal, however, is very interesting considering that the only difference between subjects' avatars was their shirt color and the models themselves looked very simple and somewhat primitive. Since mutual gaze has been found to be a sign of affection, the gaze behavior of the avatars may contribute to making the experience more personal. The behavior here, not just the mere presence of the avatars, seems to be making a difference. This indication is even clearer for question 7, where the majority of subjects say that it is easier to communicate using avatars. There is therefore a strong sense of the avatars adding something significant to the communication beyond what is achieved with text only.

1. WOULD USE AGAIN
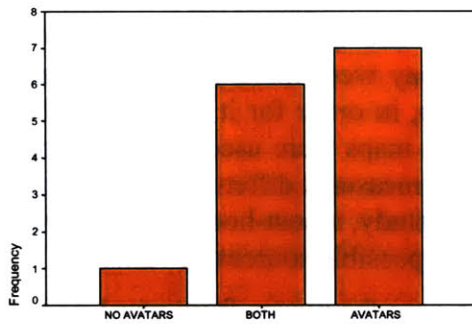
2. MORE USEFUL
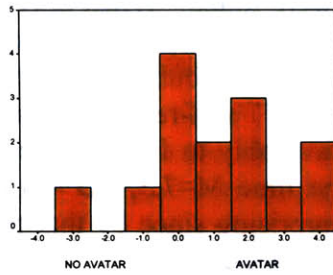
3. MORE FUN
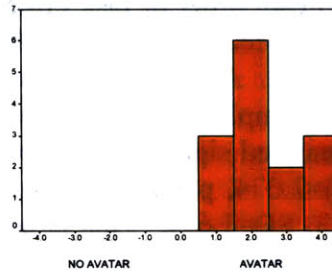
4. MORE PERSONAL
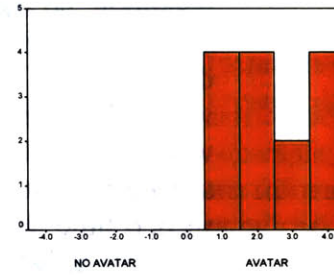
5. EASIER TO USE
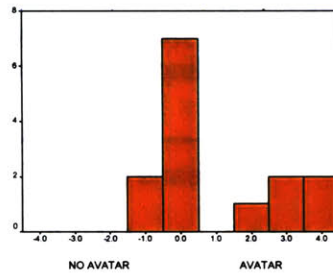
6. MORE EFFICIENT
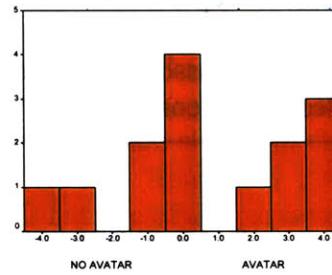
7. COMMUNICATE EASIER
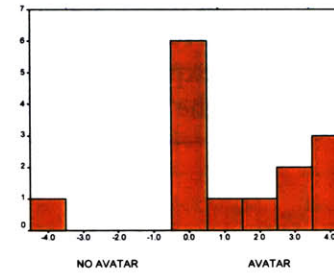
Graph 1: Preference reported by those that used both an avatar system and a no-avatar system. These are histograms showing the number of subjects behind each preference score.

Post-hoc tests

Order and Task effects

Because each subject participated in two trials, it was possible that the second time they used the system they were more experienced and that this would affect the measures. Also, in order for it to be possible to do two different trials, two different task maps were used, making it possible that the maps themselves affected the measures differently. Although both order and maps were balanced in the study, a post-hoc check for order and task effects was conducted to explain possible sources of variance.

Two-tailed t-tests were used to compare the means from map1 and map 2 for all measures and to compare the means from first and second trials for self-report measures. When looking at the data overall, regardless of condition, no significant order or task effects were found. When only the data from the two speech conditions (AS and NS) was examined, a different story emerged. A significant order effect was found for the "task would have been completed better face-to-face" self-report measure, where the trial compared significantly more favorably to face-to-face the second time ($t(41)=2.074$, $p<0.05$, 2-tail, first time M=7.3, SD = 1.42, second time M=6.3, SD=1.69). This is not surprising since by that time the subjects have become more used to the TTS engine.

More strikingly, there was a significant task effect found for 8 of the 21 measures. These are summarized in Table 21. In all cases Map 1 leads to a worse experience than Map 2. Considering that no task effect was found overall (and not in the text condition as seen below), this can only be explained by certain words associated with Map 1 sounding very bad with the TTS, possibly leading to confusion.

MEASURE	2-tailed t-test	Means (SD)	
		MAP 1	MAP 2
tedious	$t(41)=3.196$, $p<0.01$	4.8 (1.13)	3.7(0.99)
entertaining	$t(41)=-3.738$, $p<0.01$	5.1(1.45)	6.5(0.97)
engaging	$t(41)=-2.360$, $p<0.04$	4.6(1.38)	5.6(1.19)
other effort	$t(41)=-2.335$, $p<0.04$	6.1(1.60)	7.0(0.94)
other trust	$t(41)=-2.273$, $p<0.04$	6.3(1.80)	7.3(0.93)
task difficulty	$t(41)=1.970$, $p<0.08$	4.9(1.23)	3.9(1.53)
task consensus	$t(41)=-2.542$, $p<0.02$	5.9(1.69)	7.3(1.60)
brok.adjacency pairs	$t(13)=2.142$, $p<0.08$	0.4,(0.11)	0.2(0.161)

Table 21: Significant task effects in the speech condition. Map 1 provides worse means for 8 measures

When only the data from the two text conditions were examined (AT and NT) an order effect was found for the rating of "others' effort", where the sense for effort significantly increased between the first and second trials ($t(39)$= -2.14, $p<0.04$, 2-tail, first time M = 6.2, SD = 1.56, second time M = 7.2, SD=1.05), and for how much the subjects felt the communication was like face-to-face; the feeling of face-to-face being significantly stronger the second time ($t(39)$=-2.119, $p<0.05$, 2-tail, first time M=3.3, SD=2.07, second time M=4.7, SD=1.93). Both of these are not surprising and are in fact evidence of a good trend. This time only a single significant task effect was found, where the equality of contributions was significantly worse for map 1 than for map 2 ($t(14)$=-4.784, $p<0.01$, 2-tail, map1 M=7.2, SD=3.60, map2 M=20.6, SD=7.37). This is a very strange finding. There is no effect on number of hints shared, so this is not because certain hints in one map were irrelevant, thus reducing someone's opportunity to contribute. It is not clear what contributes to this result.

In general the order effect does not seem like something that greatly affects the data, and for the text condition at least, the task effect is not great either. However, in the speech condition the considerable task effect is an unwanted artifact that introduces a new source of variance that reduces the power of the data in that condition.

Observed Power

In order to understand better what is going on when a measure provides no significant results, the observed power (alpha = 0.05) for each test for an avatar main effect was calculated (using SPSS), both when using the pooled data and the sub-populations. The power is the likelihood that the lack of significant difference is the result of there really not being any difference between the groups being studied.

Unfortunately the power tended to be lower than the generally accepted threshold of 0.8, indicating that the study needed more subjects to compensate for the large variance within many of the measures. The main reason why this turned out to be the case is that the power could not be increased by analyzing the data as coming from a typical within-subject study (as had been expected), which would have isolated the within and between subject variance. In addition, the poor speech and its interaction with the task seems to have added another obfuscating factor.

MEASURE	POOL	MAP 1	MAP 2	ORD. 1	ORD. 2
Behavior (all)	0.3	0.2	0.4		
Process Self-rep.	0.7	0.3	0.4	0.3	0.6
Task Self-rep.	0.9	0.8	0.6	0.8	0.7
Social Self-rep.	0.8	0.2	0.6	0.8	0.2

Table 22: The maximum observed power (alpha=0.05) within each category of measures and within each data population. It is clearly indicated here that a lack of significant results from behavior measures is most likely due to lack of sessions

Table 22 summarizes the results from looking at the power across the measures. The table breaks the measures into four categories: all behavior measures, and then the self-report measures corresponding to each of the three hypotheses (process, task and social outcomes). The table lists the maximum power found for any measure within a category, for each of the tested data populations.

This gives an idea about how much a data population can at best be expected to explain the lack of significant findings. For example, it is clear from the Behavior row that no conclusions can be drawn from a lack of a significant main effect for avatars for any of the behavioral measures other than more sessions would have been needed.

Conversation Process

This section reports on those results that contribute to testing the process hypothesis (see Table 17). 11 different measures of the quality of conversation process were taken. 7 of these were behavioral measures and 4 were self-report measures (see Appendix C for the questionnaires). The means of these measures are shown in Graph 2, where they have all been normalized to fit a scale from 0 to 1 where a higher value represents higher quality. All but one of the means is higher in the avatar condition.

QUALITY OF CONVERSATION PROCESS
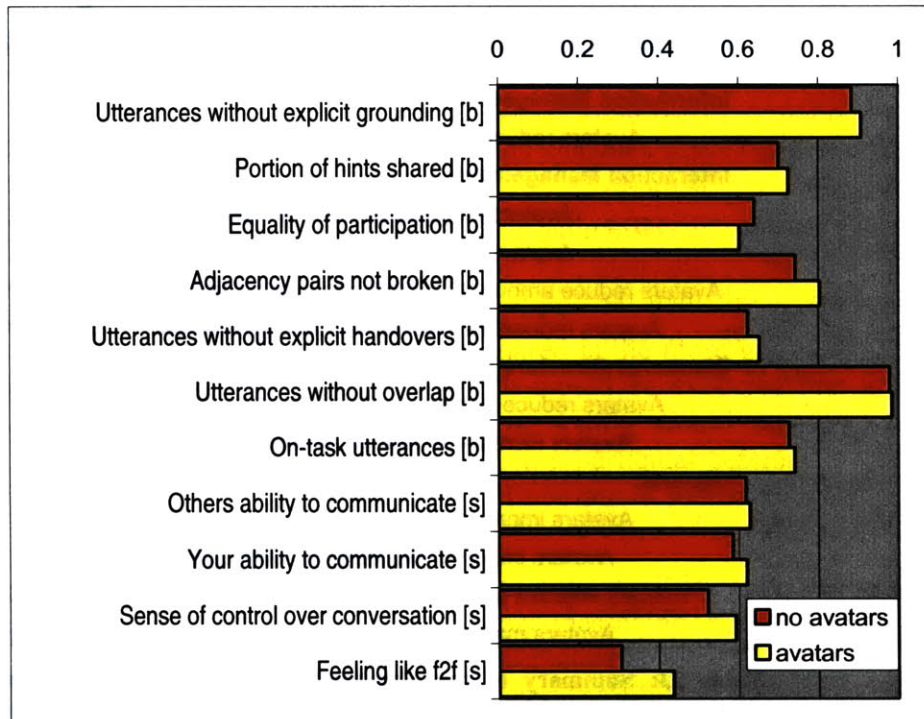
Graph 2: The means of the 11 measures of quality of conversation process in the avatar condition and in the no avatar condition. The means have been normalized as scores from 0 to 1, where higher is better. All but one of the means is higher in the avatar condition.

To test whether the avatars significantly improve the overall quality of the conversation process, a t-test was used to test whether the mean difference between the avatar means and the no avatar means was significantly greater than 0. The result of this test ($t(10)=2.596$, $p=0.014$, 1-tail, $M=0.034$, $SD=0.043$) indicates that this is indeed the case, supporting the process hypothesis. The test is not affected by the independence of trials assumption since the variance taken into account is only the variance between the means and not within each measure. The rest of this section will take a closer look at each of the measures.

What do avatars improve?	Type	Pooled significance	Support
Information Management:			
Avatars reduce portion of grounding utterances	b		weak
Interaction Management:			
Avatars increase number of shared hints	b		none
Avatars increase equality of participation	b		none
Avatars reduce amount of explicit handovers (see text)	b	p<0.05	weak
Avatars reduce the number of overlaps (see text)	b	trend	weak
Discourse Structure Management:			
Avatars reduce portion of adjacency pairs broken	b		none
Avatars increase portion of on-task utterances	b		weak
General:			
Avatars improve others ability to communicate	s		none
Avatars improve your ability to communicate	s		none
Avatars improve sense of control over conversation	s	p<0.08	weak
Avatars make conversation feel more like f2f	s	p<0.05	good

Table 23: Summary of the conversation process measures and the strength of support each provides. Type refers to (b)ehavior and (s)elf-report data. Pooled significance refers to the level of significance when assuming trials are independent and Support reports on the amount of support from both the pooled population and the independent sub-populations

Table 23 summarizes the results from testing each process measure individually. The first column lists the measures (stated as hypotheses with respect to the expected impact of avatars on that measure). The type column indicates whether the measure is a behavior measure or a self-report measure. The letter "s" indicates a self-report measure and the number refers to the associated questionnaire. Pooled significance is the level of confidence that the measure hypothesis is supported by the data if all trials are assumed to be independent (see discussion on data). The Support column indicates the total level of confidence when the map and order sub-populations are taken into account as well (good means that at least 3 groupings provided significant support at p<0.08, partial means at least 2 did and weak means at least one trend at p<0.15).

Grounding

The portion of all utterances during a session that are dedicated to grounding as opposed to contributing new content has often been used as a measure of conversation process quality. A poor channel of communication calls for a lot of explicit grounding acts where participants double-check to see if everything is being correctly understood (see Figure 16).

```
1 GREEN      which one of those is the mine?
2 BLUE       the one with the sign
3 GREEN      oh k
4 BLUE       the tent has the desert map
5 GREEN      oh the TENT has the map
```

Figure 16: An excerpt from an NT session where green performs explicit grounding in response to blue's statements. Utterance 3 is a simple "OK" ("k" is a chat convention for "ok") and utterance 5 is a complete repetition emphasizing the crucial information

In the MAP=2 sub-population avatars significantly reduce the portion of utterances spent on grounding $F(1,12)=3.636$, $p<0.081$, $h2=0.233$.

GROUNDING
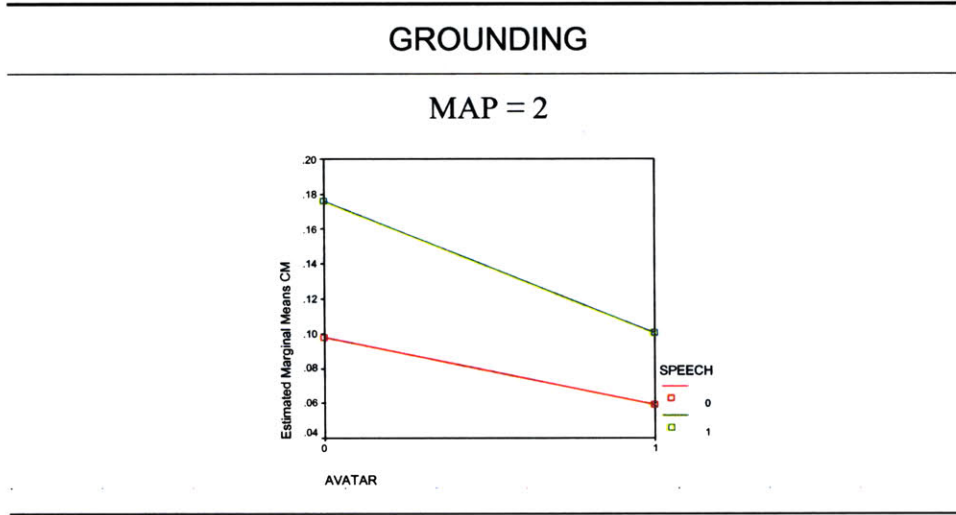
MAP = 2

Graph 3: Avatars significantly reduce the portion of all utterances exchanged during a session that deal with the conversation process itself in the MAP=2 sub-population.

Shared Hints

Each subject received 5 unique hints about the terrain the group had to cross. The hints were constructed so that to have the greatest chance of solving the task well, everyone had to share what they knew with everyone else. This of course relied on everyone being able to contribute equally to the discussion, something that a good conversation process should facilitate. Therefore, the total number of hints shared can be taken as one measure of the quality of conversation process.

Overall, it was very typical for the groups to start sharing all their hints in an orderly fashion (see Figure 17). This was a strategy employed regardless of condition; however, it might have been possible that either the avatar or the speech factor influenced the success of this strategy.

```
1 ORANGE     let's begin by sharing our information
2 GREEN      you go first
3 ORANGE     i'll go first, okay?
4 BLUE       All right…
5 ORANGE     the desert slows you down to 1/3 of normal
            walking speed
```

Figure 17: An excerpt from a session in the AT condition where subjects are about to share their hints in an orderly fashion

No significance was reached for an avatar main effect.

Equality of Participation

Related to how many hints everyone shared is the more general measure of how much everyone participated. It would be expected that a good conversation process in a collaborative setting would facilitate equal participation. Equality of participation in a particular session is taken to be the difference between the number of utterances submitted by the most active and least active participants.

No significant effects were found for the avatars. It is possible that since the subjects were all used to typical text messaging, they were familiar enough with the relatively abrupt participation style (no subtle way to indicate willingness to contribute) to ensure their full participation in both conditions.

Explicit Handovers

When a turn is about to finish, the current speaker can explicitly hand it over to the next speaker by asking a direct question, ending with a tag question (such as "right?") or mentioning the next speaker by name. A turn is defined as the set of utterances contributed by a single participant without interruption from others. An increased portion of all turns that end in explicit handovers indicates that participants are relying more on the verbal channel than the nonverbal channel to manage turn-taking. For example, this is a behavior that has been shown to be more frequent in low quality video conferencing than in face-to-face interaction (Whittaker 2002).

For the purpose of getting a more fine-grained picture of this phenomenon, each transcript was divided into four quarters. Graph 4 shows the portion of turns ending in explicit handover, across the four transcript quarters, for each of the four conditions. There appears to be quite a difference between quarters. The first quarter corresponds roughly to introductions, the second quarter to sharing of information, the third to discussion and the last to decision and farewells. Perhaps the most interesting one is the discussion quarter, where the conditions seem to diverge a lot.

124
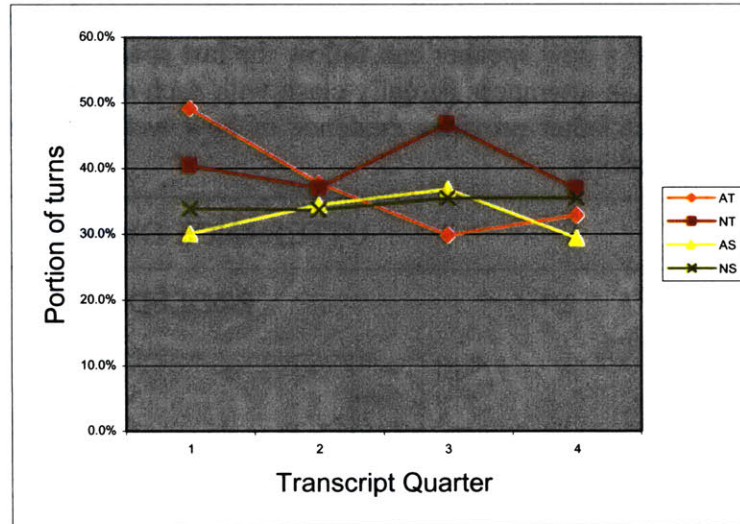
Graph 4: The portion of turns ending in an explicit handover, charted by condition and transcript quarters. This reflects the pooled data.

For the third quarter, no significant main effect was found with the pooled data, though a two-tailed t-test comparing just the AT and NT conditions showed a significant difference in the means (**AT**:M=0.30/SD=0.10, **NT**:M=0.47/SD=0.18, t(14)=-2.212, p<0.022, one-tailed). However, for MAP=1 a significant main effect was in fact found for avatars (F(1,80)=5.772, p<0.035, h2=0.344) but not for MAP=2.

A further examination of the difference between the AT and NT conditions revealed that tag questions such as "right?" and "isn't it?", and heavily emphasized questions (ending with "???"), were almost twice as likely in the NT condition as in the AT, or 3.1% of turns versus 1.8% of turns (see Figure 18). The AS and NS conditions don't show the same drastic difference, perhaps because of problems with the speech, including the fact that intonation for questions was not done correctly.

```
1 ORANGE      does that work?
2 BLUE        is that the route?
3 GREEN       should we get the gold and take the balloon?
4 GREEN       yes
5 BLUE        i think so
6 ORANGE      and we can't take the bridge at all, right?
```

Figure 18: An excerpt from the third quarter of an NT session showing a typical tag question at the end. It is also interesting to notice how it is hard to tell whether utterance 5 is a reply to 1 or 3

Utterance Overlap

When turns are coordinated face-to-face, the turn-taking mechanism helps participants avoid destructive overlapping of utterances while ensuring that a new speaker can follow the last speaker without much delay. How much utterances actually clash with each other and how close they follow each other provides evidence of how well the turn-taking mechanism is working.

UTTERANCE OVERLAP

POOLED
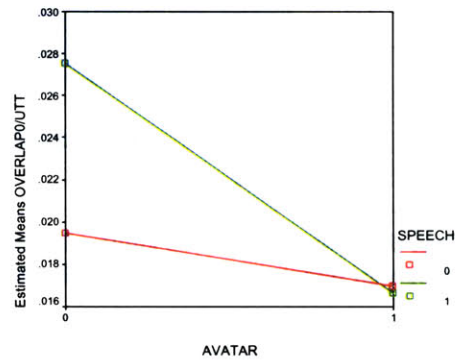
Graph 5: The portion of all utterances exchanged that overlap. Avatars show a trend (p<0.1) for reducing the amount of overlap in the speech condition.

No significance was reached, but a trend was observed where the avatars reduce the amount of overlap in the speech condition. There is a difference in the mean portion of overlapping utterances in the AS condition and the NS condition (**AS:**M=0.017/SD=0.01, **NS:**M=0.028/SD=0.02, t(7)=1.305, p<0.1075, one-tailed).

Broken Adjacency Pairs

An adjacency pair is a pair of utterances where the first utterance demands the second one as a reply. An example would be a question-answer pair. A broken adjacency pair, i.e. where the first in the pair appears in the conversation without the closure of the second one, is a sign of possible failure in the conversation process (see Figure 19). A participant may not have realized that they were being addressed or that a relevant follow-up contribution was called for. The number of broken adjacency pairs is reported here as the portion of all adjacency pairs in a session that were broken.

```
1 ORANGE      what's that tree on the opposite bank from
             the balloon?
2 BLUE        how long for the ship?
3 GREEN       balloon travel is 1/2 normal walking speed
4 BLUE        how long is the ship?
5 GREEN       the route through the mountain seems short so
             it may be worth it
```

Figure 19: An excerpt from the NT condition showing two participants attempting to start a pair (utterances 1 and 2) but both failing to get an answer. Blue even tries a second time in 4 without any luck.

No significant result was reached for this measure.

On-task Utterances

The portion of all utterances during a session that directly relate to solving the task at hand (see Figure 20) has sometimes been used as a measure of how well a communication medium supports focused discussion. Although this measure should only been taken as a part of a bigger picture, it does shed some light on whether something about the conversation process inhibited work focus.

```
1 GREEN       Oh, my, I seem to have an accent…
2 ORANGE      it looks like we're in the fog area
3 GREEN       Yeah, so I think we need a compass.
4 BLUE        mmm
```

Figure 20: An excerpt from a conversation in the AS condition showing two on-task utterances (2 and 3). The first utterance is irrelevant to the task and the last one is a filler.

There is no significant main effect for avatars, but in the speech condition a trend shows a possible difference in the mean portion of on-task utterances for AS and NS (AS:M=0.72/SD=0.1,NS:M=0.64/SD=0.1, $t(7)$=-1.191, $p<0.1365$, one-tailed).

MAP = 2
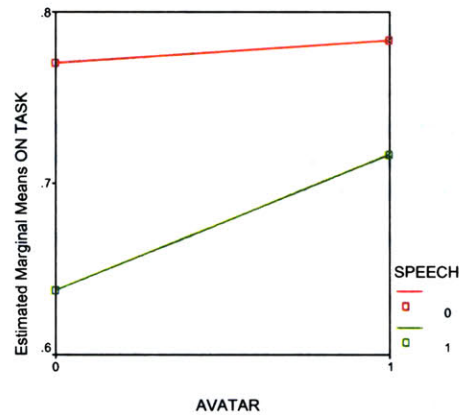
Graph 6: In the speech condition a trend indicates that avatars may be increasing the portion of on-task utterances

Others ability to communicate

The single-trial questionnaire included questions regarding how well the subject understood the others and how well they thought the other participants were able to express themselves. The aggregate result of these questions is a measure of how other participants' ability to communicate was perceived. No significant main effect for avatars was reached.

Your ability to communicate

Subjects were asked how well they could express themselves to others, how well others seemed to understand them and how well the system allowed them to communicate. The aggregate of these questions is a measure of a participant's perceived ability to communicate. No significant main effect for avatars was reached.

Control of conversation

The questionnaire asked the subjects to rate how much control they had over the conversation. This is a question repeated from the earlier BodyChat experiment (Cassell and Vilhjalmsson 1999) and is meant to get at whether subjects felt their avatars were exhibiting irrelevant conversation behavior.

Sub-populations did not back up the significance found in the pooled data and therefore no conclusions can be drawn other than more power may be needed.

Like face-to-face

The questionnaire included a question where subjects were asked to rate how close the communication experience was to a face-to-face experience.

In the overall population, the avatars made the online conversation feel significantly more like face-to-face conversation ($F(1,80)=5.523$, $p<0.021$, $h2=0.065$). Significance was also reached for the MAP=1 sub-population ($F(1,29)=3.744$, $p<0.063$, $h2=0.114$) and the ORDER=2 sub-population ($F(1,36)=3.601$, $p<0.066$, $h2=0.091$).

LIKE FACE-TO-FACE

POOLED
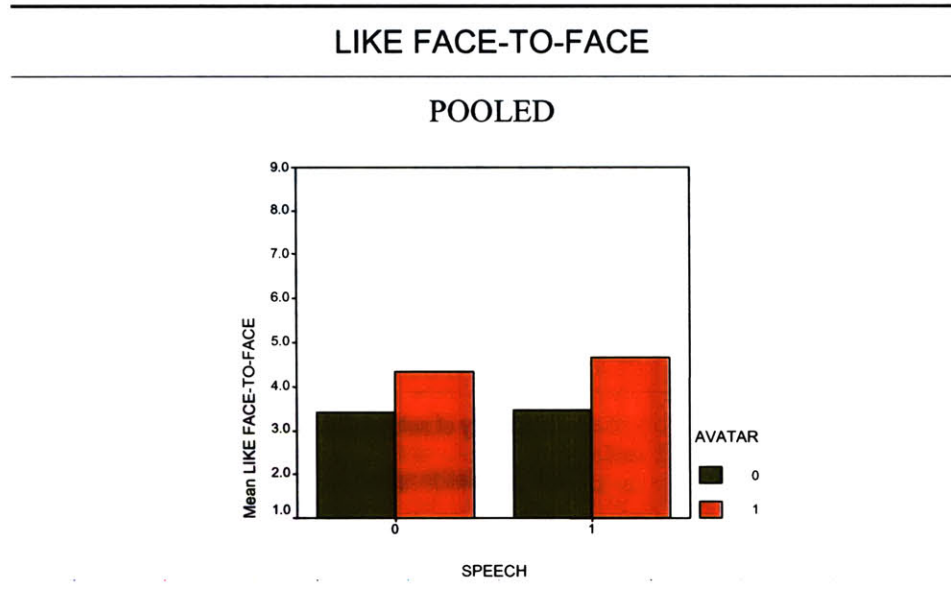
Graph 7: Avatars made the online conversation feel significantly more like face-to-face

Summary

The comparison of the mean difference between the avatar and no avatar means across all 11 measures showed that the avatar condition scored significantly higher, supporting the hypothesis that the avatars improve the overall quality of the conversation process. When looking at individual measures however, there was generally not enough power to produce significant results. Only one subjective measure, namely how close to face-to-face the conversation felt, was a good significant result in favor of the avatars. The avatars do significantly reduce the number of grounding utterances, but only when map 2 was being discussed. This may indicate that there was a difference in the maps themselves and suggests that the effectiveness of the avatars may depend on the context. Trends along what was expected were found in the number of overlaps (avatars reducing them in the speech condition) and the portion of on-task utterances (avatars increasing them in the speech condition). No unexpected effects or trends were observed. While a more definite impact of avatars was expected on each of the measures, the findings are

encouraging and further studies, with greater number of subjects and a more careful design, may be able to show significance where weak evidence was found here.

Task Outcome

This section reports on the results from the study that contribute to testing the part of the outcome hypothesis that has to do with task outcome. The task is the route planning task, and although the chosen path from that task may be the most direct measure of how successfully the group collaborated, a few other measures were also taken to get an overall sense for the quality of collaboration. The quality of the chosen path and the time spent on solving the task constitute behavioral measures and in addition 6 self-report measures were taken (see Table 18). The means of these measures are shown in Graph 8, where they have all been normalized to fit a scale from 0 to 1 where higher is better.

QUALITY OF TASK OUTCOME
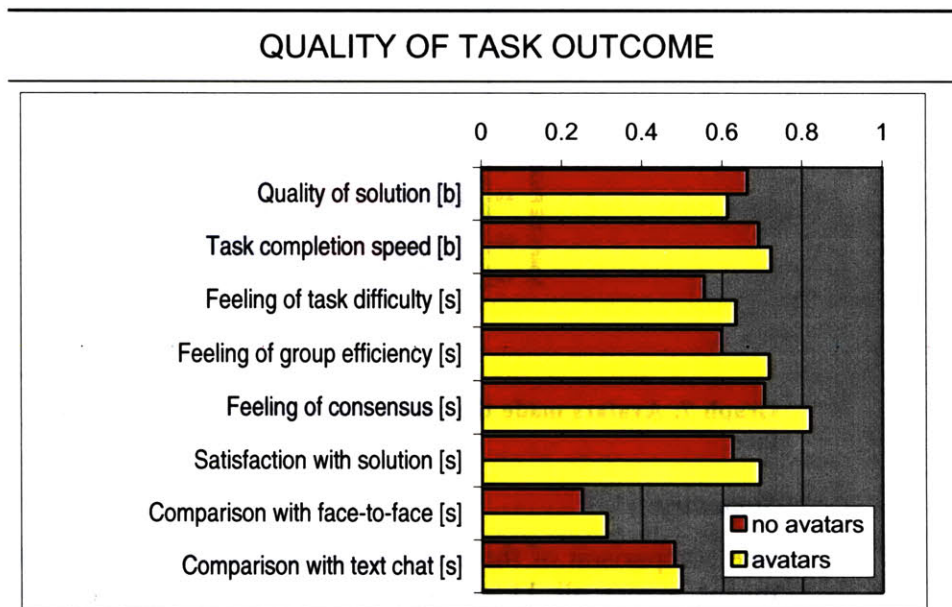
Graph 8: The means of the 8 measures of quality of task outcome in the avatar condition and in the no avatar condition. The means have been normalized as scores from 0 to 1, where higher is better. The mean quality of solution is the only one lower in the avatar condition

To test whether the avatars significantly improve the overall quality of the task outcome a t-test was used to test whether the mean difference between the avatar means and the no avatar means was significantly greater than 0. The result of this test ($t(7)=2.835$, $p=0.013$, 1-tail, $M=0.055$, $SD=0.055$) indicates that this is indeed the case, supporting the task outcome hypothesis. The rest of this section will take a closer look at each of the measures. Table 24 provides a summary of results from testing each task outcome measure separately.

130

Do avatars improve the task outcome?	Type	Pooled significance	Support
Avatars improve the task solution	b		none
Avatars reduce task completion time	b		none
Avatars make the task feel less difficult	s	p<0.08	partial
Avatars makes you feel the group is more efficient	s	p<0.01	good
Avatars improve the feeling of consensus	s	p<0.08	partial
Avatars improve your own satisfaction with solution	s		none
Avatars compare more favorable with face-to-face	s		weak
Avatars compare more favorable with regular text chat	s		none

Table 24: Summary of the task outcome measures and the strength of support they provide for the task outcome hypothesis

Solution

A session was finished when all the participants had agreed on and highlighted a single route from start to finish. By using the information about the terrain and various transport options, the travel time for each route could be calculated. No two routes resulted in the same travel time. The solution from each session was given a score from 0 to 5, where 5 was the score of the fastest route. 0 was given to any solution that was slower than the best 4 routes.

Significant difference between conditions in solution scores was not found. The post-hoc observed power for an avatar main effect in each of the data-populations was less than 0.1, indicating that more sessions would have been needed to draw any conclusions other than that the variance is quite high.

Time to complete

Each group was told that they not only had to come up with a good solution, but that they were also under time pressure. However, they were not given any explicit time limits or shown the passage of time. All groups were allowed to finish; there was never any need for stopping a session.

The median time to complete was 16 minutes. This time does not include initial briefing and time spent filling out questionnaires. A whole session could range from 30 minutes to an hour. The following chart shows the distribution of completion times across all conditions:
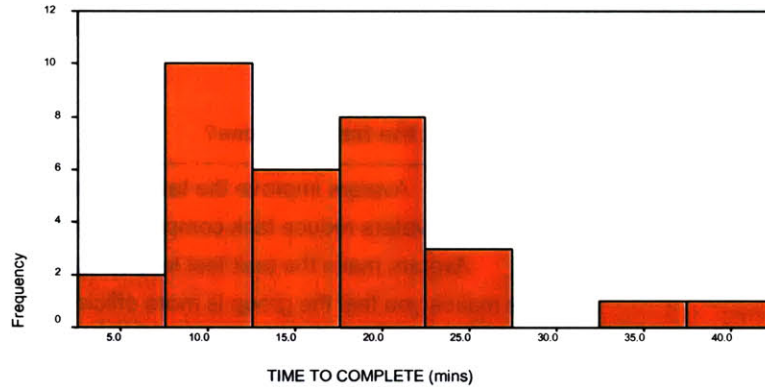
Graph 9: Distribution of total time to complete from all conditions

No significant effect was found for avatars on the total time to complete.

Task difficulty

To collect some subjective measures of task outcome, a few task related questions were included in the questionnaire, the first of which asked the subjects to rate the difficulty of the task.

For the pooled population the avatars made the task feel significantly less difficult ($F(1,80)=3.349$, $p<0.071$, h2=0.040). In the text condition, the avatars significantly reduce the difficulty for the MAP 1 sub-population ($t(20)=-2.377$, $p<0.04$, 2-tailed, AT: M=3.22, SD=1.86, NT: M=4.98, SD=1.50) and for the ORDER 1 sub-population ($t(22)=-2.301$, $p<0.04$, 2-tailed, AT: M=3.73, SD=1.62, NT: M=5.08, SD=1.26).

TASK DIFFICULTY

POOLED
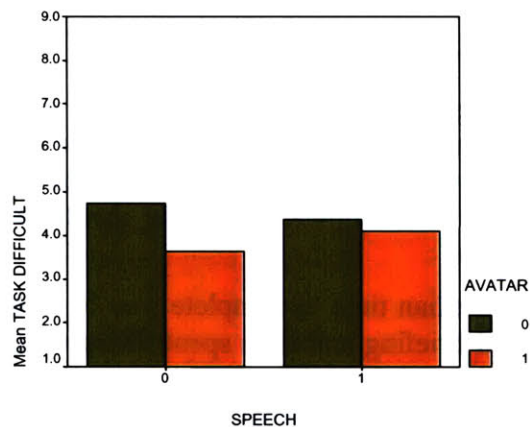
Figure 21: Avatars made the task feel significantly less difficult in the text condition

132

Group efficiency

The questionnaire asked the subjects to rate the group's overall performance and how efficiently the group solved the task. Together these were taken as a subjective measure of group efficiency.

In the overall population, the avatars made the subject feel that the group was solving the task significantly more efficiently ($F(1,80)=11,571$, $p<0.001$, $h2=0.126$). This effect is also significant for the MAP=1 population ($F(1,29)=6.825$, $p<0.014$, $h2=0.191$), MAP=2 population ($F(1,47)=4.723$, $p<0.035$, $h2=0.091$), ORDER=1 ($F(1,40)=5.915$, $p<0.020$, $h2=0.129$) population and the ORDER=2 population ($F(1,36)=6.443$, $p<0.016$, $h2=0.152$).

GROUP EFFICIENCY
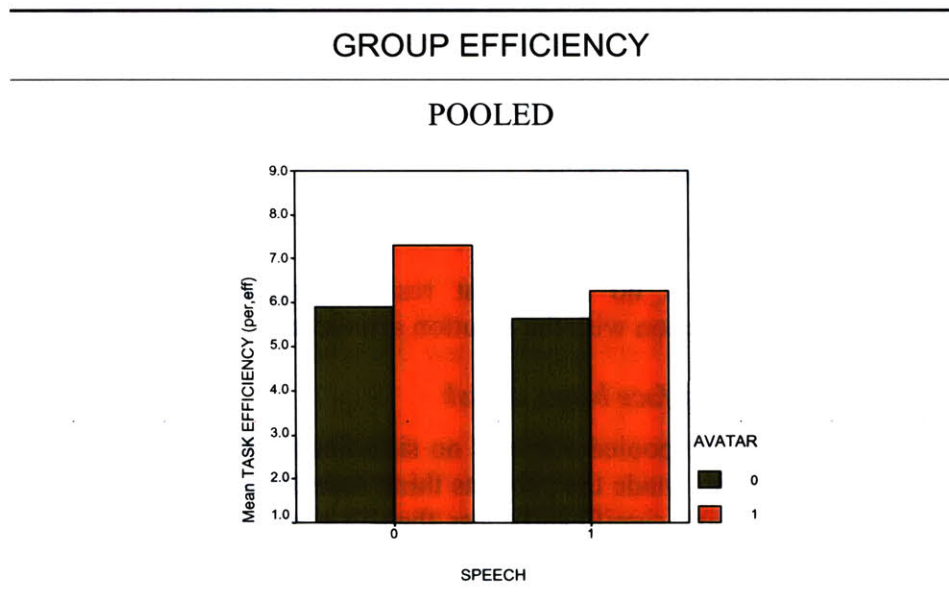
POOLED

Graph 10: Avatars made the subjects feel that the group was solving the task significantly more efficiently

Task consensus

In the pooled population the avatars significantly improved the reported group consensus regarding the solution ($F(1,80)=8,034$, $p<0.006$, $h2=0.091$). This effect is also significant for the MAP=2 sub-population ($F(1,47)=4.196$, $p<0.046$, $h2=0.082$) and ORDER=1 sub-population ($F(1,40)=7.859$, $p<0.008$, $h2=0.164$).

POOLED
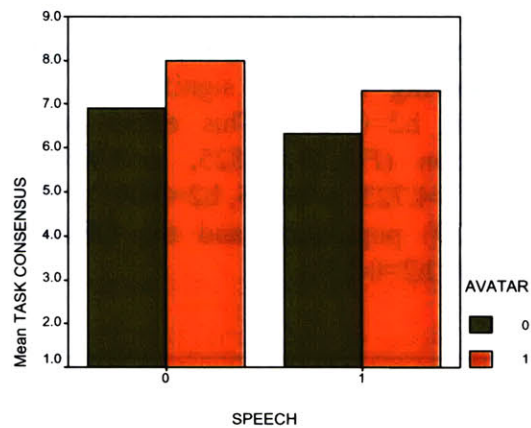
Graph 11: The avatars significantly improved the reported group consensus regarding the solution

Subject's satisfaction

There is no significant result regarding the subject's reported own satisfaction with the solution arrived at.

Face-to-face better at task

For the pooled subjects no significant effects were found. However, the avatars made the subjects think face-to-face would have improved solving the task significantly less than in a non-avatar condition for the MAP=1 sub-population ($F(1,29)=8.589$, $p<0.007$, $h2=0.228$) and an expected trend was observed in the ORDER=2 sub-population ($F(1,36)=3.159$, $p<0.084$, $h2=0.081$).
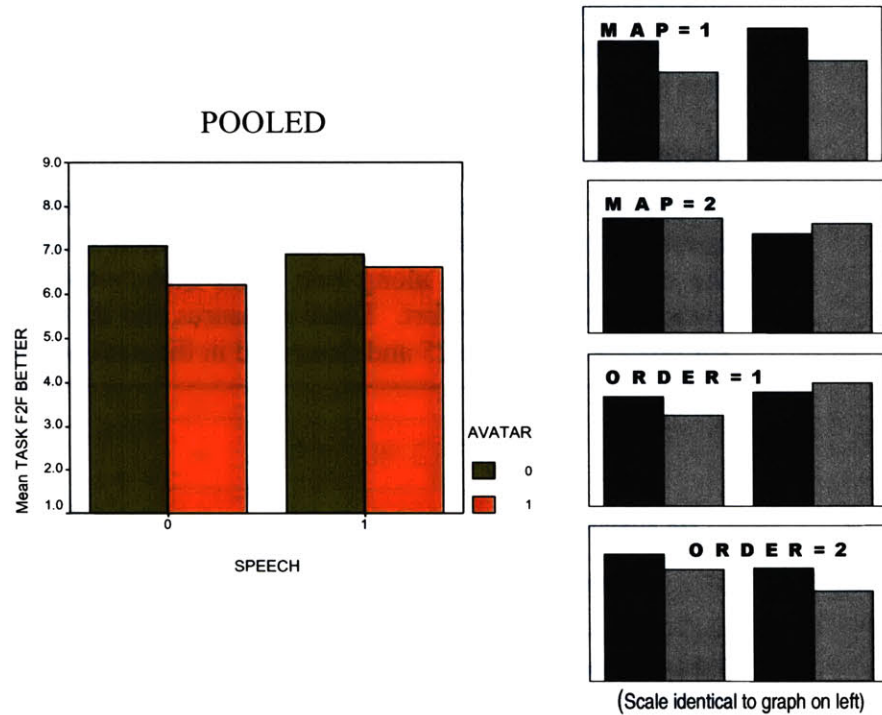
Graph 12: How much better the subjects think face-to-face would have allowed them to solve the task. While the pooled data showed no significant results, those using avatars in the MAP=1 group thought f2f would be a significantly less improvement

Text better at task

There is no significant result regarding how much better the subject thought regular text chat would have allowed them to solve the task.

Summary

The comparison of the mean difference between the avatar and no avatar means across all 8 measures showed that the avatar condition scored significantly higher, supporting the hypothesis that the avatars improve the overall task outcome quality. However, when looking at individual measures, the two behavioral measures fail to provide significant support. Several self-report measures however showed significant support for the avatars. In the text condition only, the avatars made the subjects feel the task they were solving was significantly less difficult. The avatars made the subjects feel the group was being significantly more efficient at solving the task and the feeling of consensus was significantly stronger. The avatars compared significantly more favorably to face-to-face with regard to how well the system allowed the subjects to solve the task, but this significance was only found in the group working on map 1 though a trend was found among those coming for the second time. It is curious

135

that even though the overall collaboration experience seems to have improved with the avatars, the mean task solution got worse, though not significantly. Something about the work process was improved and it is possible that other tasks or settings may see a more direct outcome benefit (see section on proposed follow-up studies below).

Social Outcome

To test the social outcome part of the outcome hypothesis a number of questions on the questionnaire asked how each subject related to each of the other participants along two main dimensions aggregated into two measures: trust and effort. These measures and the significant findings are summarized in Table 25 and described in the next couple of sections.

Do avatars improve the social outcome?	Type	Pooled significance	Support
Avatars improve the sense for other people's effort	s	$p<0.01$	partial
Avatars improve trust in other participants	s	$p<0.04$	weak

Table 25: Summary of the social outcome measures and how strong their support is for the social outcome hypothesis

Other participants effort

The answers to three questions were combined to form a measure of perceived effort. Those three questions asked about the other participant's interest in collaborating, helpfulness and how well they listened to others.

OTHER EFFORT
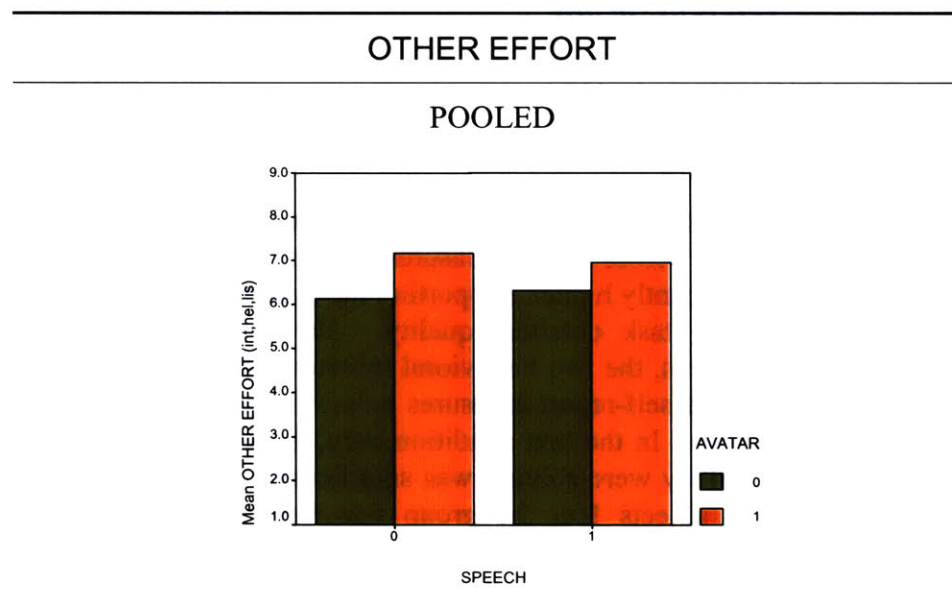
POOLED

Graph 13: Avatars made the subjects feel their partners were putting significantly more effort into the collaboration

136

In the pooled population avatars made subjects feel their partners were putting significantly more effort into the collaboration ($F(1,76)=8.030$, $p<0.006$, $h2=0.096$). This is also significant for the MAP=2 sub-population ($F(1,44)=5,127$, $p<0.029$, $h2=0.104$) and the ORDER=1 sub-population ($F(1,40)=8.055$, $p<0.007$, $h2=0.175$).

Trust in other participants

Again three questions were combined to probe for reported trust in the other participants. These questions asked about trust directly, honesty and how comfortable the collaboration felt.

In the pooled population avatars significantly increased trust in other participants ($F(1,76)=4.256$, $p<0.040$, $h2=0.054$). While other sub-populations showed no significant effects, an expected trend was observed in the MAP=2 sub-population ($F(1,44)=3.775$, $p=0.129$, $h2=0.051$).

OTHER TRUST
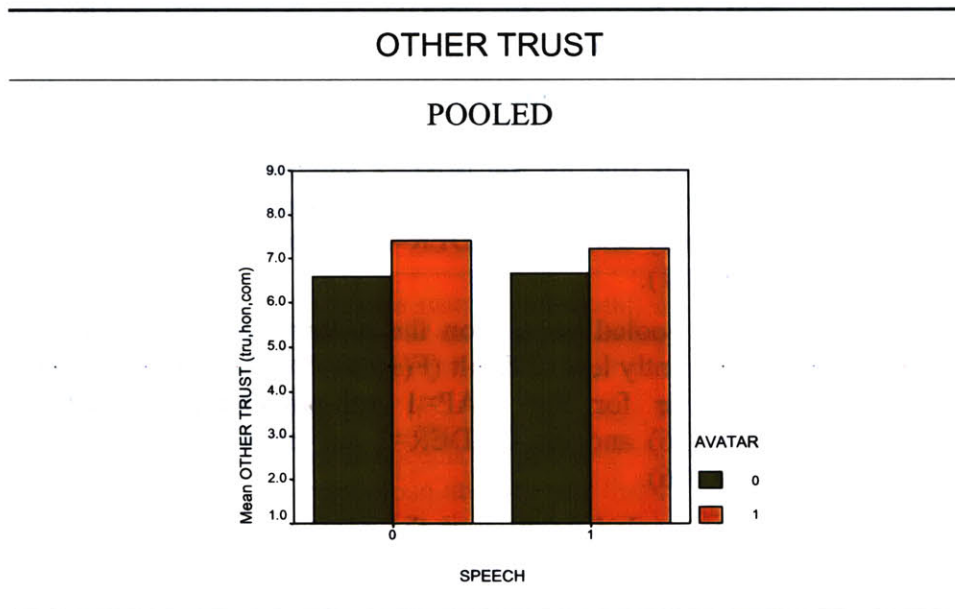
POOLED

Graph 14: Avatars increased reported trust in the other participants

Summary

In the pooled population, the avatars significantly improved the subjects' sense of effort the other participants were putting into the collaboration and the amount of trust they had for them. The former was completely backed up by the significant results from the sub-populations, and can therefore be considered good support, but the latter only had one trend backing it up and is therefore weak. Together these provide some support for the social outcome hypothesis.

Avatar Interface

In order to evaluate whether the novel avatar interface introduced any new overhead that would have distracted from the experience, a few questions

137

addressed the overall experience of using the system. These and the results of testing for the impact of the avatar interface are summarized in Table 26.

Overall experience of using the avatar system	Type	Pooled significance	Support
Avatar interface made experience less tedious	s	p<0.01	good
Avatar interface made the experience less difficult	s	p<0.04	good
Avatar interface made the experience more engaging	s	p<0.01	good
Avatar interface made the experience more comfortable	s	p<0.01	good
Avatar interface provided more control over conversation	s	p<0.08	weak
System felt easier to use than a non-avatar system	s[8]	p<0.04	good

Table 26: Summary of questionnaire results that addressed the overall experience of using the system

In the pooled population the avatars made the overall experience feel significantly less tedious ($F(1,80)= 14.167$, $p<0.000$, $h2=0.150$). The same was true for the MAP=2 sub-population ($F(1,47)=13.649$, $p<0.001$, $h2=0.225$), the ORDER=1 sub-population ($F(1,40)=7,745$, $p<0.008$, $h2=0.162$) and the ORDER=2 sub-population ($F(1,36)=4.517$, $p<0.040$, $h2=0.111$).

In the pooled population the avatars made the overall experience feel significantly less difficult ($F(1,80)=5.647$, $p<0.020$, $h2=0.066$). The same was true for the MAP=1 sub-population ($F(1,29)=4.560$, $p<0.041$, $h2=0.136$) and the ORDER=1 sub-population ($F(1,40)=3.763$, $p<0.059$, $h2=0.086$).

In the pooled population the avatars made the overall experience feel significantly more engaging ($F(1,80)=8,686$, $p<0.004$, $h2=0.098$). The same was true for the ORDER=1 sub-population ($F(1,40)=3.591$, $p<0.065$, $h2=0.082$) and the ORDER=2 sub-population ($F(1,36)=5.961$, $p<0.020$, $h2=0.142$).

In the pooled population the avatars made the overall experience feel significantly more comfortable ($F(1,80)=13,620$, $p<0.000$, $h2=0.145$). The same was true for the MAP=1 sub-population ($F(1,29)=12.121$, $p<0.002$, $h2=0.295$), the MAP=2 sub-population ($F(1,47)=3.816$, $p<0.057$, $h2=0.075$) and the ORDER=2 sub-population ($F(1,36)=13.135$, $p<0.001$, $h2=0.267$). Furthermore an expected trend was found in the ORDER=1 sub-population ($F(1,40)=2.659$, $p<0.111$, $h2=0.062$).

In the pooled population the avatars made the users feel they were in greater control of the conversation ($F(1,80)=3,174$, $p<0.079$, $h2=0.038$).

---

[8] This measure is repeated here from the Preference Questionnaire

As reported earlier in the Preference section above, after subjects completed two trials, the ones that experienced an avatar condition and a no-avatar condition were asked to rate the strength of their preference for the "avatars system" versus the "no avatars condition" according to the "ease of use".

Significantly more people leaned towards the avatar-based system as the mean preference score of 1.0 (positive sides with avatars) was significantly higher than 0 (no preference) ($t(13)=2.082$, $p<0.03$, one-tailed, M=1.00/SD=1.80).

EASE OF USE

PREFERENCE
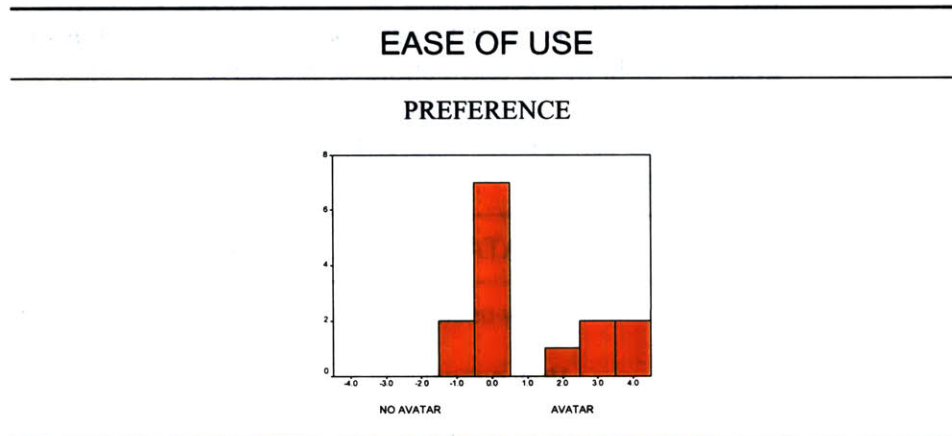
Graph 15: Significantly more subjects felt the avatar based system was easier to use than the system without avatars

Summary

From the single trial questionnaires, the avatar system felt significantly less difficult and tedious and the experience felt significantly more engaging and comfortable. The sub-populations back this up completely. In the pooled population the subjects using the avatars felt they were in significantly greater control of the conversation. This particular result is not backed up by the sub-populations and is therefore weakened. It is, however, in line with the results from the earlier BodyChat experiment. When users picked a system they would prefer for ease of use, significantly more picked the avatar system. These results give strong evidence to the claim that the avatar interface did not introduce additional complexities or overhead for the users; in fact, the overall experience only improved.

7.3.5 Other Results

Modalities

To assess whether the subjects were in fact paying any attention to the avatars on the screen, the questionnaire included three questions unique to the avatar condition. These questions were "How useful to the interaction do you think the avatars were?", "How natural did the avatar behavior

139

seem?" and "Were you paying attention to the avatars?". The answers to these questions turned out to differ between the text and speech conditions. On the question about usefulness, only 31% said the avatars were very useful in the text condition, but 50% said they were very useful in the speech condition (see Table 27). 31% said they were not that useful in the text condition, but 20% in the speech condition. A similar outcome was found for the perceived avatar naturalness. 25% said the avatar behavior was very natural in the text condition, but that number rises to 42% in the speech condition. 38% said the behavior was not natural in the text condition, but 25% said so in the speech condition. However, both groups of subjects seem to have paid a similar amount of attention to the avatars. 69% in the text condition and 63% in the speech condition were paying attention to the avatars most of the time. 13% in the text condition and 17% in the speech condition said they were paying little attention to the avatars.

AVATARS	Text	Speech
Very useful	31%	50%
Very natural	25%	42%
Attended most of the time	69%	63%

Table 27: How many subjects rate avatars very useful, very natural and something they are paying attention to most of the time, depending on whether they are in a text condition or a speech condition

The conclusion from this is that while avatars seem more useful and natural when coupled with speech (supporting the modality hypothesis), the subjects generally paid close attention to them. It is therefore safe to assume that the avatars were in fact a well-noticed feature in the avatar condition and that any significant differences between a non-avatar and an avatar condition can be attributed to their presence.

To assess whether the verbal communication modality, speech or text, was effectively delivering the typed messages, questions were included about how well the subjects understood what was said. When replying to the question "How much of the text messages were you able to read?", about 10% said they were only able to read little (scores 1 to 3 out of 9). When replying to the question "How easy was it to read the text messages?", about 20% said readability was low (scores 1 to 3 out of 9). The biggest complaint, as seen from the freeform comments, was that text messages from multiple subjects overlapped each other, making it impossible to finish reading a message after a new message was submitted. It is interesting to note that the median score on these two questions was higher, though not significantly so, when avatars were being used (first question went from a median score of 6.5 to 7.0 and second question from 5.0 to 6.0).

The TTS scored very low on naturalness. When answering the question "How natural were the voices?" 65% said they were very unnatural in the no-avatar condition, but 50% in the avatar condition. In both conditions about 25% seem to have had a hard time understanding the speech (scores 1 to 3) according to their answers to "How well did you understand the voices?". In the comments, many users complained about a funny accent (the speech synthesizer is British) and many pointed out that intonation did not properly differentiate between statements and questions (the intonation rule for questions had not been added at the time of the study).

In summary, the readability of the text was generally rather low, though most of it got read. The speech was very unnatural and not very clear, but most of it was understood. The verbal delivery was adequate for the purpose of this evaluation, but was not something one would implement in a practical application. Having avatars seems to improve how these modalities were rated, though not significantly.

User Comments

At the end of each questionnaire a few blank lines were provided and subjects encouraged to write down any suggestions or comments they felt like sharing with the designers of the system they just used. Collecting these comments and looking at what got repeated mention provided some valuable insight.

Across all conditions, a few subjects mentioned that the system was too slow overall to be practical for collaboration (this is due to long text processing times) but that once the messaging speed matched the speed of current messaging systems, it would be a whole new game. Some subjects said that some of their messages never got transmitted for some reason. This would be because malformed strings that would crash the parser are removed instead of being allowed to wreak havoc. This is a seldom occurrence though.

By far the greatest number of comments in the speech condition reported on the poor quality of the voices. They would mention funny accents, the fact that questions didn't sound like questions and general difficulty in understanding what was said. Some mentioned that text captions along with the speech would be an improvement and that some way of browsing or repeating previous utterances would also help.

In the text condition, the greatest complaint was that messages overlapped, making them difficult to read when multiple subjects submitted simultaneously. Many subjects really missed the history-browsing feature of regular text chat systems and commented that such a feature would help with solving the task.

A number of subjects in the avatar condition wrote that they had a lot of fun. More specific avatar comments included that they were excellent, engaging, very cool and fun to watch. Some subjects mentioned that the

eye contact was nice and that having the avatars definitely helped with turn-taking. For some subjects the avatars gave them a greater "group-like" sense. But it was also pointed out that while the avatars are great and get the job done, they are not perfect yet. A couple of subjects said it was not always clear what the avatars were pointing at and one subject said that the movements did not look natural at all. A few subjects really wanted the ability to design their own avatar.

8 Discussion

8.1 Possible follow-up studies

While subjectively an improved task outcome was supported in the study, the avatars did not improve the objective task solution as had been expected. The reason is probably a combination of the following:

1. *Task:* Solving this particular task did not rely heavily on the interaction between subjects. Even if everyone was briefed differently about the terrain, the subjects quickly shared this information and could possibly proceed with solving the puzzle on their own.

2. *Motivation:* There was little motivation for scoring well, so many subjects didn't really try hard. This was clear from looking at the transcripts and seeing subjects say things like "let's pick a random path and get out of here!"

3. *Study Design:* The experimental design was too complex, leaving a weak statistical model for analysis.

4. *Implementation:* The output animation and speech in the implemented system, together with overall slowness, did not do a good enough job of representing the behaviors that were generated, and so the face-to-face-like effects were not as strong as expected.

Incorporating the lessons learned, possible follow-up studies could be proposed to explore the issue of task outcome further[9]. Such studies should address the shortcomings mentioned above.

Task

Not all tasks depend the same way on interaction. Some tasks call for more interdependence of participants than others. A classification of tasks according to the level of interdependence has been proposed by (McGrath 1984) and further elaborated on by (Cugini, Damianos et al. 1999). Research has shown that as the level of interdependence increases, the benefits of being face-to-face on productivity increase (Straus 1997). Specifically, that research compared text chat with face-to-face for solving three kinds of tasks. Lowest on the interdependence scale, and the one showing the lowest performance gain was the "idea generation" task. In this kind of task the participants are essentially engaged in a brainstorming session where everything goes. The task in the middle, showing some more performance gain, was an "intellective task" or a puzzle-like task.

---

[9] Thanks to Deepa Iyengar for insightful discussions about this

The route-planning task is of this type. Even though participants were given different pieces of information, once they all shared that information (often done right up front), the task essentially became a puzzle that each participant could in effect solve without much help. The third task, the one that showed the greatest performance gain when face-to-face cues were present, was a "judgment task" or a "decision-making task." This is a task where participants are asked to develop consensus on issues that do not have correct answers.

A typical judgment task involves having the subjects order items on a list according to a subjective metric such as perceived importance. In order to make use of the shared visual space provided by Spark, these items should be represented visually as props. A good set of items, that would have an interesting visual representation are classic inventions. The task could involve having the subjects, as a group, place 5 inventions in an order, from the most significant to the least significant.

Because there is no single correct solution, the task outcome would involve measuring other task related characteristics. Three existing measures could be used for a task of this nature: time used to reach consensus, strength of consensus and amount of persuasion.

Time: The time at which subjects have achieved an ordering that is maintained until they decide they are done or are asked to quit.

Consensus: All subjects would be asked to order the same objects according to their own judgment in a post-test questionnaire. The distance of a group solution from a total consensus is calculated as the sum of the rank that everyone's individual post-test two top choices received in the group ranking (top choice has a rank value of 1 and bottom choice a value of 5).

Persuasion: As well as the post-test mentioned above, all subjects would also order the objects in a pre-test questionnaire. The extent to which the conversation made each participant change their opinion would be calculated as the difference in ranking between the two top choices on the pre-test to the ranking of those same items on the post-test (greater difference meaning greater the persuasion effect of the conversation).

The first study showed subjective positive impact of the avatars on the feeling of consensus, so having a follow-up task rely even more on consensus and measuring this effect objectively is likely to provide strong results.

Motivation

A real-world reward is one of the best ways to motivate subjects to work hard on the task. Most subjects tend to participate because monetary reward is involved. If they are promised a bonus for good performance they are more likely to put in some extra effort. In order to add competition to the task of ranking inventions, the instructions could

144

actually be "please rank the inventions in the order you think most MIT students would rank them if asked to order them from the most to the least important to their current quality of life." Then the subjects could be told that a certain random portion of those that get it "right" will get a bonus. The most popular ranking to come out of the experiment itself would be deemed to be the "right" answer for the purpose of handing out the bonus.

Design

While the first study addresses the question how seeing the animated avatar bodies affected the communication, follow-up studies could ask other but related questions. Here three different study designs are suggested, all using the task and outcome measures described above.

Study I

> *Hypothesis*: "Groups using the new face-to-face paradigm do better on a judgment task over those groups that use the state-of-the art in online collaboration."

> *Goal*: Compare the face-to-face avatar paradigm with a shared workspace paradigm currently representing the state-of-the art in online collaboration. This comparison has the potential to demonstrate the power of a new paradigm and uses a system most people are familiar with as a reference.

> *Method*: Two sets of groups solve the judgment task, one using a popular collaboration system such as NetMeeting that integrates a text chat with a shared whiteboard and another using a Spark based system. In NetMeeting the inventions to be ordered would be images on the whiteboard. In the 3D avatar environment, they would be objects on the table in front of the avatars.

Study II

> *Hypothesis*: "Groups that use avatars modeling conversational behavior in avatars do better on a judgment task than groups that use minimally behaving avatars"

> *Goal*: To show the impact of modeling appropriate behavior by demonstrating that the effects found with the new animated avatars so far are not due to their mere presence but to their carefully crafted behavior.

> *Method*: Two sets of groups solve the judgment task, both using a Spark based system, but for one group all behaviors are turned off except for lip movement when speaking and random idle movement.

Hypothesis: "Groups that use the animated avatars and groups that interact face-to-face show improvement in judgment task performance over groups that use text only chat. Groups interacting face-to-face show the greatest improvement."

Goal: To show that the avatars that animate typical face-to-face behavior actually move the performance of online collaboration closer to that of actual face-to-face.

Method: Three sets of groups solve the judgment task, one face-to-face, one using the Spark based system with the avatars visible and one with no avatars visible.

The follow-up studies should be between-subject studies to get around a possible learning effect. A between-subject study would also alleviate problems associated with scheduling groups of subjects for return visits and the possibly shifting group membership.

Implementation

For the study that has been conducted, the implemented system and animation was deemed "good enough" by experts to represent the theoretical model. As mentioned in the MapChat technical evaluation, there were still a few issues, especially with time lag. Furthermore, the animations themselves felt a little "stick-figure-like" because Pantomime is currently only capable of rendering joint rotations of stiff segments with no natural deformation of the body.

All of these technical issues are under constant improvement. Lag times improve as computers get faster, and several parts of the MapChat implementation are being fixed and optimized as a result of running the user study. Using Pantomime to control a skinned character animation rendering engine instead of Open Inventor has been successfully tested, so future animations in Pantomime may see drastic improvement in naturalness.

8.2 Applications and special considerations

Spark was meant to support a variety of CMC applications. Some different kinds of applications and what needs to be considered when employing Spark in those new situations are discussed next.

Regular chatting and messaging

Chat rooms are popular places to hang out and socialize. Graphical chat rooms, sporting avatars, are already widely used. Applying Spark to a chat rooms is relatively straightforward though a few things need to be considered.

First, it may be hard to anticipate the topics that are going to be covered during such free form discussion. It is therefore not clear how the knowledge base could be prepared so that many interesting gestures (i.e. anything other than beat gestures) would emerge. Some chat rooms are organized around particular subjects, so a few key items may be set up beforehand. Most chat rooms also contain a considerable amount of introductions, farewells and small talk, all of which could be represented to some degree in the knowledge base. Given the emphasis on relationship building and maintenance in chat rooms, it may be worth adding relational behaviors to the model (see 9.3).

Second, many graphical chat rooms provide ways for users to customize their avatars. In fact, systems like the Palace thrive on the idea that everyone can supply their own pictures and animations for their avatars. This helps people build an online identity. Spark does not place any constraints on what the avatars look like, other than that they should be able to exhibit a certain range of nonverbal behaviors.

Third, chat rooms and especially instant messaging systems, are usually lightweight applications that are quick to launch, don't drain a lot of system resources and are easily resizable. Because these are meant to support casual interactions, these systems are typically run in multi-tasking mode with other applications. Pantomime as a character animation engine is not appropriate for this use, but Spark's animation output could be compiled for any kind of animation system, including a Flash style plug-in.

Fourth, people that are already familiar with text chat employ all sorts of chat conventions for expressing themselves effectively. These conventions need to be taken into account when analyzing text messages, both because they may confuse algorithms that expect regular language but also because additional information about communicative intent may be extracted from known conventions.

Collaborative work

People in settings less casual than chat, such as business meetings, tend to ask for more realistic looking avatars, mainly so that they can easily verify the identity of those they are meeting. Spark does not prevent this at all, but as the avatar's visual quality approaches that of the real person, one starts to also expect higher quality of motion. If appearance and motion quality do not go hand-in-hand, the effect can be quite jarring. Therefore sticking to carefully drawn animated portraits rather than animated photographs may be the safest way to go.

People doing "serious" business may want to be able to control the amount of automated behavior allowed to ensure people can't completely fool their collaborators into thinking that they're working really hard.

Alternatively, the automation can be fed with information from sensors about the actual attentive state of participants, for example using a camera.

There needs to be a way to manipulate the environment and in particular any objects that are being discussed or worked on. A complex scenario such as a training session may require the avatars to operate simulated equipment. This would call for a whole new range of hand and arm motions in addition to the standard conversational gesturing. The system could incorporate plan recognition to automatically initiate equipment manipulation as the instructor describes each step in a procedure for example. However, it should also be possible to manipulate the environment directly similar to how MapChat provided a way to point at paths explicitly using the mouse. At all times it should be clear from the avatar limb movements who is manipulating what and what they are doing, with and without the exchange of words.

Multiplayer Games

There are many different styles of multiplayer games that call for different styles of conversation. Many games essentially incorporate a typical chat room and so the free-form discussion and chat conventions concerns from the chat category above are applicable. Even if anything goes in the discussions, the game worlds themselves provide a very rich context for generating behavior. Places, beings and artifacts are all known and can be incorporated into nonverbal references either through deictic gestures or iconic representations. For example when telling someone you just came back from an encounter with a group of leprechauns your avatar could generate a "low sweeping gesture" representing "short folks".

Other games, especially games that emphasize tactical cooperation, have much more constrained conversations going on and even provide shortcut keys with the most commonly exchanged utterances. In a limited domain like that the knowledge base and the behavior generators can be fine tuned to fit the scenario. It is important to point out that even though the utterances are pre-canned, the associated behavior can still depend on the context (for example "drive back to the base" could either have an associated pointing gesture towards a nearby vehicle on "drive" or pointing towards the base on "the base", all depending on what the previous command was). The approach of dynamically augmenting messages is therefore still applicable when message content is pre-determined.

In persistent game worlds, it is important that a certain amount of individual context be kept with each character between sessions so that behavior is consistent. For example, the avatar agent should always be able to access who are currently your friends and whom you have developed hostility towards so that it doesn't accidentally invite a band of bugbears to a friendly chat. Like with chat rooms, customization is important and beyond customizing appearance, the augmented avatars

would allow their users to set up and tweak reactive behaviors. For example, a user could define a "really friendly invitation to chat" sequence of behaviors reserved for their closest friends.

Game world also share a lot with general collaborative work applications, including the possibility of manipulating the environment directly. The game worlds could contain a variety of activities that would have a well-defined local discourse context and special interactive objects that become a part of the conversation. For instance, a place for constructing magical items would involve all sorts of ingredients that can be combined in particular ways. If these ways are known by the system, the gestures associated with describing how to make a healing potion for example could be very descriptive.

8.3 Interesting issues

8.3.1 Appropriate behavior

Wrong behavior

There is no absolute guarantee that an automatically picked behavior represents the actual communicative intent of a user – it is simply an approximation based on the available data. In some cases this may have a serious impact on the conversation that is taking place and in other cases this may slip by relatively unnoticed in the face of other stronger cues. It is important to try to avoid the former from happening.

One approach would be to give all communicative functions an *impact rating* that roughly corresponds to how large an effect executing that function would have on an ongoing conversation or how critical it is to get it right. In addition, whenever a function markup is added, a *certainty parameter* could be included. This parameter would reflect the strength of the evidence behind this particular tagging. For example when identifying a discourse entity the parameter could represent how good the match is to the corresponding entry in the knowledge base and the relative strength of other contenders. The product of the impact rating of a marked up communicative function and the complement of the certainty parameter would be the risk factor of portraying the intent. Depending on the situation, a risk threshold can be set that would simply block any behaviors that have a certain likelihood of having a negative impact.

Finding the right values for impact ratings and certainty parameters is a difficult task. This task could be aided by the users themselves by allowing them to give feedback back to the system when they notice something wrong. The users would essentially train the system by providing negative re-enforcement. A similar approach has been demonstrated with such an extension to BodyChat (Gorniak 2000).

Pre-emptive listener behavior

What is the point in automating listening responses before the actual listener has had a chance to hear what is being said and form a "real" reaction?

Although it is possible to program arbitrarily complex listener behavior, the main reason for providing this functionality is to support minimal channel maintenance. Behavior such as slight head nodding and eyebrow movement serve to assert the speaker that they are being noticed and listened to, but they don't have to mean that the listener is fully understanding or agreeing with the speaker. This would correspond to behaviors carrying out the two lowest levels of grounding according to the "four layers of grounding theory" (Clark 1996). It is then not until after the speaker has completed the utterance that the listener will then provide explicit evidence of understanding or agreement, taking the grounding behavior to the next level. Therefore there does not have to be any conflict between the automatically generated low level listener responses and the higher-level transmitted responses.

The importance of automatically generating signal and channel grounding behaviors is clear when one considers that the absence of these behaviors can be taken as evidence of a failed signal transmission, e.g. that the speech is not heard, or a broken channel of conversation, e.g. lack of proper attention. Not generating these behaviors could be disruptive for the speaker who is expecting a certain level of participation.

Deceptive behavior

Users can program their avatars to show interest and alertness. How does it contribute to a better conversation or collaboration if this invites users to send a deceptive message about their status? A user may not be following the conversation at all, while their avatar convinces everyone else that they are processing all the information being shared.

There is always room for abuse. The inconsistency would quickly be discovered when the active participation of the users in question is required and their contributions have to fit into the ongoing conversation. The goal of the avatar automation is to help participants to participate more fully by giving them an expanded range of behaviors, not to take over any of the participant's responsibilities. There may be situations though where that is called for (see 9.5).

8.3.2 Appropriate technology

Smart recipients

It is a computer that has to infer a speaker's intent from a narrow information stream and then augment it to make it easier for the recipient to understand what the speaker meant to communicate. Is that assuming

that the computer is smarter at making the right inference than the recipient?

In some communication environments, e.g. graphical learning environments and online games, the speaker is already represented by an avatar and it is important that the avatar behave in a manner consistent with the communicative intent. The system has no choice but to infer this intent and animate the avatar accordingly because he lack of appropriate behavior or random inappropriate behavior is likely to make it harder for the listeners to arrive at their own judgment about what is going on.

Seemingly redundant speaker behaviors added by the system, such as deictic gestures along with full textual reference, can act as a focusing device for the listener, essentially underlining the important context for the listener's interpretation efforts.

Furthermore, the system can draw from resources that are not immediately available to the listener to generate non-redundant behaviors. These resources are represented by the various knowledge bases contained within the discourse context. For example, a feature of a newly introduced object can be depicted through gesture without being mentioned in the message itself, simply by tracing a referent to a rich entry in a domain knowledge base.

Balance of control

How do you know how much of the avatar behavior in general should be left up to automation? The short answer is that it depends entirely on the context of use. But for each context there are several factors that need to be considered. Perhaps the most important thing to have in mind is that ultimately the users should feel in absolute control of the situation they are dealing with, which possibly may be achieved through greater automation at the behavioral level. For example, being able to tell your avatar that you wish to avoid certain people may free you from having to worry about accidentally inviting them to chat by making an unexpected eye contact.

There are other factors to consider as well. First of all, the avatar may have access to more resources than the user to base its behavior on. These resources basically represent the remote environment in which the avatar resides. Beyond what is immediately visible, the avatar may even be able to use senses not available to the human user. In the example above, the avatar would be able to know whether the person you are trying to avoid is standing behind you and therefore would not make the mistake of turning around to face them. Time is also a resource, and sometimes it is crucial that an avatar reacts quickly to a situation. A time delay from the user to the avatar could force control over the situation out of the user's hands.

Related to the resource of time, the avatar can maintain consistent continuous control of the remote situation even if the link from the user is a discrete one. The discreteness may be the result of a physical link that

can only support control commands in short bursts, or it could be that high cognitive load requires the user to multi-task. In either case, delegating control to the avatar may ensure that the remote operation is not interspersed with abrupt standstills.

Although an avatar is meant to be a representation of a user, it does not necessarily mean that the avatar can only mimic what the user would be able to do. In fact, the avatar is an opportunity to extend the capabilities of the user, even beyond the capability of being in a remote place. For example tele-operated robots, which in a sense are physical avatars, may be able to perform operations such as changing a valve at super-human speeds. The user, or operator, may therefore want to leave the execution up to the robot after making sure it has been maneuvered into the right spot. Similarly, in a social setting, an avatar could have certain nonverbal behavior coordination skills programmed that are beyond what the user would be able to orchestrate. A user could for example choose an avatar that knew how to produce the gestural language of a riveting speaker, leaving the exact control of that skill up to the avatar itself.

On the other hand, some users may want the opportunity to interface closely with any new skill sets offered by the avatar and in a way learn to wield them as their own. This idea of learning and then refining your control over new expressive capabilities of a device is what underlies the research on musical hyper-instruments (Machover 1991). Training and practice to use a communication interface is not something people are commonly ready to do, but being able to deepen the level of control to fit increased human expertise is something to keep in mind.

9 Future Work

9.1 Overview

The goal of the work presented in this thesis is to augment online conversation by employing avatars that model face-to-face behavior. The goal can be divided into 3 parts: understand what a person means to communicate (input), define a set of processes crucial for successful interaction and the set of behaviors that support them (model), and finally coordinate those behaviors in a real-time performance (output). Future work can expand on each of these parts.

9.2 Input and interpretation

Speech

While text is will most likely continue to be the most popular messaging and chat medium, voice-over-IP technology is providing increasingly higher quality voice conferencing for applications ranging from shared whiteboards to games. There are certainly situations where voice is the best option, such as when hands are not free to type. It is therefore important to consider what it would take to augment a speech stream using the approach presented here.

Speech recognition is not good enough yet to provide a precise text transcript of any conversational chat. This means that all words and word boundaries of a message can't be known using today's state-of-the art. Certain keywords could be spotted, however, and those may be enough to keep track of some relevant discourse entities.

However, the speech signal carries a lot of useful information that directly contributes to identifying the important units of discourse. The intonation contour and pitch accents carry out functions of information structure, feedback elicitation and turn taking, all of which could then be tagged directly from an intonation analysis.

Until speech recognition gets better, the intonation units could be used instead of the words as the basic units being processed and annotated by the Spark pipeline. The behaviors in the end would then be synchronized to these units, preserving proper co-ordination of verbal and nonverbal modalities.

Observed behavior

In a similar way that linguistic cues can be found in the written messages or intonation cues in the spoken channel, other behaviors may also hold cues to a person's communicative intent. Even though the person is not engaged in a true face-to-face interaction, the behavior observed while

using the communication system is likely to reflect what is going on in the mediated interaction. For example, it is not uncommon to see violent bursts of laughter from people engaged in lively text chat or people lean into their screens when attempting a difficult task during a game.

It is possible to capture many of these behaviors passively through cameras or carefully placed sensors. Once captured, they have to be interpreted and their communicative function described in a frame to be transmitted either alone or as along with verbal content. It may sound odd that the idea is not to send the observed behavior directly since it has already been captured. But as mentioned earlier, the captured behavior in the physical environment may not map correctly onto the avatar in the virtual environment due to different visual configurations. Therefore some translation may be necessary and a functional description will ensure the translations will maintain the original intent. It is also interesting to think of the functional representation as a very compressed representation of behavior, using only a few bits to send information about a large set of observations. The functional representation will then be decompressed on the receiving end, producing a full range of behavior again, adjusted to fit the new environment.

Plans and artifacts

Business meetings often involve agendas and training sessions revolve around the procedures being taught. These are examples of explicit plans that could contribute to the generation of behaviors that help structure the conversation. By applying plan-recognition techniques on the message exchange, the discourse module could not only mark when topics shifts occur, but also what the topic is and where it is embedded in the overall topic structure. This would allow more fine tuned topic shift behaviors, for example making a distinction between a major shift and a minor shift.

When artifacts are involved, for example shared documents or simulated equipment, knowing what part is being discussed becomes very important because the direction of visual attention needs be appropriately generated. Plan-recognition can be helpful here again, but beyond that the artifacts themselves could contain information about where the important visual features are located and how they can be brought into view or manipulated. For example, a complex piece of equipment may require the avatars to flip it over and open a hatch before being able to point out the feature being discussed.

Direct control and input devices

Analyzing the text or speech signal, observing the user or attempting to recognize progression of plans, are all example of passively learning about what people are communicating to each other. Passive methods can capture spontaneous and involuntary cues, and don't distract the participants from being engaged in conversation. However, that does not

mean that the participants should not have the option of explicitly stating their intent through directly manipulating the communication interface. For example, BodyChat allowed participants to set a toggle switch on the screen to indicate whether they were available for a chat or not. While chatting, the participants could also enter a special control code into the text stream to indicate their intent to leave.

Explicit control can either add to what already has been automatically gleaned from the conversation (for example, an emoticon could add a facial expression), or it can override the system's passive interpretation (for example, surrounding a word with stars would emphasize that word, even if it was *given*). The keyboard and mouse can gather explicit commands through keywords, button presses and cursor movements, and may be the most convenient and accessible devices, especially for text-based messaging. New kinds of devices can be explored as well. For example a foot pedal could allow someone typing a message to indicate voice levels from whispering to shouting. If spoken input is being used, then the hands are free to grasp or wear other kinds of devices such as wands, 6D space balls, gloves, game pads, joysticks, or even rigged puppetry controls.

It is very important to map control functions to control degrees of freedom at the appropriate level. A balance has to be struck between expressive power and not burdening users with too many control details. Signaling high-level intent with a single button, such as agreement, may be preferred over several separate low-level control buttons, such as one for a head nod and another for a smile.

9.3 Modeling

This thesis has focused on modeling the nonverbal behaviors that support the processes of conversation, but there is a whole lot more to human expression. One can think of expression as the output from a composite of many different layers, each contributing or modifying behaviors. Some layers have to do with permanent traits such as personality or physical constraints, and others with more transient phenomena such as mood and attitudes.

The communicative layer, central to this thesis, provides the fundamental mechanisms for humans to open, maintain and use a channel of communication with other human beings. Yet, it is a layer that has often been overlooked when social or human-like behavior is modeled in animated characters and avatars. Many of the other layers are well represented in the research literature, however.

Extending Spark to encompass a wider range of human behavior and behavior quality might involve adding modules to the pipeline that implement other existing models. These are some interesting candidates:

- Models of personality

- Models of physical constraints

- Models of emotion

- Relational models

- Sociological models of roles

9.4 Output and behavior realization

Human articulation

Pantomime does a decent job of representing human articulation, but it is far from being mistaken for the real thing. Realistic procedural human motion is still a holy grail in computer graphics. Motion that has to do with conversational behavior, gesturing in particular, has proven to be difficult to model. Spontaneous conversational gesture moves effortlessly from relaxed forms to precise representations of ideas and objects, all in perfect synchrony with speech. Similar to the production of phones in speech, gesture can also coarticulate, adding even more variation to an already idiosyncratic process.

Some interesting work exists on how the quality of gesture motion can be controlled, for instance reflecting different moods (for example the EMOTE (Chi, Costa et al. 2000)), but less work has been done on parametrizing the exact form of conversational gesture. A ripe area for future work lies in finding a useful set of gesture primitives (shapes, trajectories, etc.), finding the most expressive quality control parameters, finding ways to snap gesture peaks and intervals to a timeline and finding methods to naturally blend from one gesture to another or add one gesture on top of another.

Stylized characters

Avatars do not need to replicate human appearance completely. In fact, there may be several reasons why photo-realistic avatars don't always make sense. One reason is that in many online environments, especially game and educational worlds, the users are taking on personas or characters that reflect imaginary inhabitants of those worlds and the avatars are simply not expected to resemble the users themselves. Another reason is that when people see something that looks like a human in minute detail they also expect it to move completely naturally. Since the quality of avatar movement and behavior does not yet match that of real humans, the mismatch can at best look a bit odd and at worst signal the wrong intent or appear pathological. Visual appearance should therefore not raise expectations that can't be met by the behavior. A third reason is that avatars are often displayed on small screens that can't render them life-sized. Minute size and low resolution can make it hard to recognize certain behavior. It can therefore be useful to exaggerate some of the

human features, such as the size of face and hands, as well as make some of the movements bigger, such as the raising of the eyebrows.

Stretching the boundaries of stylizing human appearance and behavior while still retaining their familiarity and readability is an interesting area of research. This thesis provides a good starting point by outlining the behaviors that significantly contribute to a conversation. Particular attention should be paid to the rendering of these behaviors so that their communicative function does not get lost. In fact, this may be the set of behaviors that should be made the most prominent through whatever techniques are appropriate for the chosen style of rendering. Traditional animation for example provides many techniques for conveying strong larger-than-life expression that can also be applied to computer animation (Lasseter 1987). Behaviors that lie outside this basic set, such as direct actions, idling or transitions, can be approached with more flexibility.

Robots

There is nothing that fundamentally prevents the articulated avatar from being embedded in the physical world as a robot. The same script that describes the animated avatar performance could manipulate the joint angles on a humanoid robot. One of the main challenges here would be to make the robot aware of its environment so that it could correctly target surrounding people or objects, for example when generating pointing and looking behaviors. In a virtual environment, the entire world is already represented in a format accessible to the avatar.

A robotic avatar has an advantage over virtual avatars in that the real world becomes its playing field, so to speak. It can move between physical locations, bringing the "communication interface" with it wherever it goes, it can manipulate objects and operate equipment (assuming a skillful robot) and interact either directly with humans or with other robotic avatars. When dealing with this sort of robotic "rendering," the concept of an avatar agent is very relevant. Some level minimal of autonomy is already needed for the robot to maintain balance and maneuver without getting stuck, but it could also use peripheral vision or sensing not available to its user to spot events of interest and automate reactive attention (bringing those events to the users attention as well), or it could provide conversation cues, such as attentive listening feedback to co-present participants.

Abstract visualization

As mentioned in the section on related work, there is interest in creating visual interfaces to online chat that don't employ articulated avatars at all. Instead, these interfaces provide abstract visualizations of the chat process as well as ways to browse the chat product, often in the form of histories. The goal of such systems is to make the interface both intuitive and informative. This thesis pursues the same goal, but addresses the intuitive

issue by modeling human nonverbal behavior, whereas abstract visualization relies on techniques from graphic design, illustration and visual arts. The informative part is where both approaches can well draw from the same source. The processes of conversation presented in this thesis and their automated analysis could easily drive displays other than animated avatars. A system, such as ChatCircles (Viegas and Donath 1999) could render its visualization from functionally annotated frames, and thus be able to do things like highlight emphasized words, stretch or move circles to show associations with referents, color the circles according to topic or show reference and topic information in the history display. By making many of the underlying discourse processes explicit in the functional representation of a message, this thesis provides a layer of information ripe for the picking.

9.5 Other mechanisms

Programmed behaviors

One feature of avatar agents hinted at but not fully exploited in this thesis is how they can be programmed to respond automatically to events in the world on behalf of their users. These programs can be a lot more complex than giving reactive listener feedback in response to feedback requests from speakers. They can make use of the avatar agent's ability to sense the entire environment and manipulate it. For example, a program could be written to simulate paranoid behavior where the avatar will turn and attend to anything happening in the periphery, or a special friends program could initiate a hugging sequence whenever another user from a special close friends buddy list approaches.

In the same way that today's avatar-based systems allow users to customize appearance to build unique identities, future systems can allow users to further refine those identities by customizing avatar behavior. Creating accessible identity programming tools for users is an interesting problem for future research. In some shared textual environments, such as MOOSE Crossing (Bruckman 1998), users already use an object-oriented scripting language to create interactive places, artifacts and creatures. A popular strategy is to derive behaviors from existing objects, but then add a personal twist. Similarly, programmable avatar agents could be built from shared extendable components that all fit together to form a unique skill set[10]. The challenge lies in coming up with data and operation primitives at the right level, balancing ease of use, flexible composition and expressive power.

---

[10] An example of a custom avatar skill is the "sword fighting skill" provided by Hiro's avatar in the novel *Snowcrash* (Stephenson 1992)

Complete autonomy

The assumption so far has been that behind each avatar there is a person communicating and supplying the actual content of what is being said. Increasing the amount of autonomy given to an avatar can challenge this assumption, blurring the boundaries between an avatar and an embodied conversational agent. For certain scenarios it can be helpful to use avatars capable of producing responses by themselves. For example, in an online customer service center, an avatar representing the support staff could greet each customer. Actual staff does not need to be present behind the avatar to begin with. The avatar could start with some automated questions to gather basic information, even attempt to answer some of the questions the customer may have. At a point where automated responses are not enough, the avatar can call in the staff to take over. From the perspective of the customer, they are interacting with the same person the whole time. Here the programmed avatar agent is helping to create a consistent conversational interface, when the person behind it is unable to provide consistent continuous input. This idea should be explored further. In particular it raises questions of how the avatar agent can deal with and recover from breakdowns in communication and how they can best inform the human who is taking over about what has transpired so far and what the expected next step is.

10 Conclusion

10.1 Supported Claims

Automatically animating avatars augments online conversation by greatly improving the subjective experience and by bringing the communication process closer to that of face-to-face. Studies of actual human conversational behavior can be used as a model and a resource to accomplish this.

The approach, represented by the Spark architecture, enables precise coordination of verbal messages and supporting nonverbal communicative behavior. Analyzing the ongoing conversation and marking the important units of discourse, based on the current discourse context, is important in achieving this. The discourse units then become the basic units of behavior.

The flexible architecture invites extension through new input devices, processing modules, knowledge sources, behavior rules and output media. The flexibility stems from a pipeline structure, the use of a common XML frames representation format, and the abstraction of function from behaviors.

10.2 Contributions

This thesis makes contributions to several different fields of study:

To the field of computer mediated communication, the thesis presents a theory of how textual real-time communication can be augmented by carefully simulating visual face-to-face behavior in animated avatars. The thesis demonstrates the theory in an implemented architecture and evaluates it in a controlled study.

To the field of human modeling and simulation, the thesis presents a set of behaviors that are essential to the modeling of conversation. It is shown how these behaviors can be automatically generated from an analysis of the text to be spoken and the discourse context.

To the field of HCI, the thesis presents a novel approach to augmenting an online communication interface through real-time discourse processing and automated avatar control. For avatar-based systems, it provides an alternative to manual control and performance control of avatars. For any communication system it introduces the idea of a communication proxy in the form of a personal conversation agent remotely representing participants.

To the field of systems engineering, the thesis presents a powerful way to represent, transmit and transform messages in an online real-time messaging application. The power lies in how an XML frame

representation of a message is first annotated with communicative function markup and then transformed through a set of simple rules to produce behavior markup that describes a complex but well coordinated visual presentation of the message. Also, by embedding the transformation in avatar agents on the receiving end, there is an opportunity for different representations on different clients, co-ordination between multiple avatar agents or objects in the simulated environment, and a sustained remote activity on behalf of the user in the absence of user input.

To the field of computational linguistics and discourse analysis, the thesis presents a unique platform for experimenting with the relationship between language and behavior in the context of multi-party conversation. The ease with which new linguistic markup and behavior rules can be added supports rapid-prototyping of theoretical models and encourages exploration.

10.3 Theory limitations and challenges

Reading thought is hard

The approach to augmenting communication presented in this thesis relies on a system's ability to understand what a person means to communicate and to predict what behavior would best further that intent. To do this perfectly is an AI-complete problem. The idealistic vision of a communications device being able to read our minds and transmit our thoughts is as unlikely to be achieved, if not more unlikely, than being able to transmit our thoughts directly from our mind to the recipients mind. Therefore any solution based on this approach will have to make certain trade-offs, for example between careful planning and properly bootstrapping the system on the one hand and precision of results on the other.

Representing the world is hard

The interpretation process makes use of discourse context, which essentially is the common knowledge participants draw from and refer to when communicating with each other. Representing this discourse context and making the same kinds of knowledge inferences as humans would is a very challenging task. No single knowledge reference format has been developed, that would for example cover everything from describing the current visual scene to describing past events experienced by the participants to describing what the last speaker just said. Yet all these factors contribute to a discourse context that could determine the exact shape of a gesture. Without a consistent way to refer to this context, it is hard to write computational rules to extract the relevant relationships. It will therefore be necessary to restrict the knowledge domain, leading to a richer set of behaviors when the communication is in line with the chosen

domain but then becoming sparser as the discussion veers off into a more general territory.

Not possible in true real-time

Humans tend to know what they are about to communicate before turning that intent into a stream of words and gesture. That intent however, may not be clear to those being communicated to until the message is completed. Similarly, a system may have to wait to see the whole message before being able to suggest appropriate gestures because they rely on the intent, not just the words themselves. This is fine when the communication is conducted in a messaging fashion, where an entire message is composed before being transmitted. However, this causes a problem when augmenting a continuous stream, such as a telephone conversation. It may be necessary to buffer the speech to provide a window wide enough for analysis. It is not clear how large the window needs to be, but it is certainly a limitation of the approach that completely real-time augmentation is impossible.

10.4 Fundamental Issues Addressed

Beyond improving upon current synchronous computer mediated communication technology and providing better avatars, this thesis addresses some fundamental issues of human communication and expression that won't change as higher transmission bandwidth and better fidelity communication systems become available. These issues are the mapping problem, expressive animation and human augmentation.

The Mapping Problem

When an attempt is made to directly project participants from remote locations into a shared communication environment, for example by means of video, it is not possible to strip away the limitation introduced by the fact that the locations are physically separate. At best participants can see into each other's location as if watching through windows. Going beyond this point requires mapping people into each other's spaces or into a new common space, which inherently can't involve an exact one-to-one mapping of behavior from each original location because of the new spatial and social configuration. This is what I have termed the mapping problem.

The thesis addresses this problem by treating a communication channel as a transformation of communicative behavior from one place to another. Key to this transformation is abstracting communicative intent from the behavior representation. By understanding what a person means to communicate, the system can adjust the behaviors on the remote end of the communication channel to fit that intent. For example, if a person wants to request feedback from someone, a representation of that intent will allow the channel to ensure that mutual eye contact is experienced on

163

both sides of the channel, even though the two locations were spatially incompatible.

Expressive Animation

The thesis contributes to computer-mediated communication, but in doing so it has also proposed a new way to animate conversational behavior in animated characters. These characters can be the avatars of people engaged in some sort of role-playing where the point is not to communicate with one another directly but actually to communicate through the avatar personas. This sort of avatar puppetry requires controls to begin with. This thesis starts to address the fundamental issue of control and points the way in the promising direction of greater autonomy of the avatar.

The animated characters that benefit from this work do not even have to be avatars in a communication environment, but any characters that we need to have interact socially with other people or deliver lines from a script. By giving the characters the ability to understand what they are meant to say and to read into the context for producing appropriate nonverbal behavior, the thesis suggests how to reduce the amount of tedious work required by human animators directing them, and it suggests new possibilities for interactive characters that don't have the luxury of any human direction.

Human Augmentation

No matter how good technology gets at transmitting a person's gesture and voice, there will always be people for whom fluently gesturing and speaking in the first place is a challenge. This thesis introduces a way to take a possibly narrow channel of communication, here in the form of text, and expand it into a full range of human communicative behavior, including gesture and speech. For people living with paralysis, for example, avatars or even robots that can augment conversation could become a new way to interface with the social world around them.

10.5 Only the beginning

This thesis has presented a new theoretical framework for augmenting mediated conversation. Automation lies at the center of this framework, where it both interprets what is being communicated and generates supporting behaviors based on a model of face-to-face conversation. In essence, it proposes a "smart" communication technology where an autonomous agent is acting on behalf of the participants, to the best of its ability, to help overcome deficiencies in the conversation channel.

While the thesis demonstrates an effective implementation, it has really just started to explore the new paradigm of mediation through a "conversation agent." Different communication scenarios, different

communication devices, different ways of accessing information about the context of communication, and different ways to represent participants present a variety of exciting new opportunities and challenges for further exploration.

References

Allen, J. (1995). *Natural Language Understanding*. Redwood City, CA, The Benjamin/Cummings Publishing Company, Inc.

Andre, E., T. Rist, et al. (1998). "Integrating reactive and scripted behaviors in a life-like presentation agent." *Proceedings of AGENTS'98*: 261-268.

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

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

Ball, G. and J. Breese (2000). "Emotion and Personality in a Conversational Agent." *Embodied Conversational Agents*. J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Cambridge, MA, MIT Press: 189-219.

Barnett, J. (2001). "Where have you been? -- A case study of successful implementation of undergraduate online learning communities." The 12th International Conference on College Teaching and Learning.

Barrientos, F. (2000). "Continuous Control of Avatar Gesture." Bridging the Gap: Bringing Together New Media Artists and Multimedia Technologists, First International Workshop, Marina del Rey, CA, ACM.

Bavelas, J. B., N. Chovil, et al. (1995). "Gestures Specialized for Dialogue." *Personality and Social Psychology* 21(4): 394-405.

Blumberg, B. M. and T. A. Galyean (1995). "Multi-Level Direction of Autonomous Creatures for Real-Time Virtual Environments." SIGGRAPH, ACM.

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

Breazeal, C. and B. Scassellati (1998). "Infant-like Social Interactions Between a Robot and a Human Caretaker." *Special Issue of Adaptive Behavior on Simulation Models of Social Agents*. K. Dautenhahn.

Brown, G. and G. Yule (1983). *Discourse Analysis*. Cambridge, UK, Cambridge University Press.

Bruckman, A. (1998). "Community Support for Constructionist Learning." *Computer Supported Cooperative Work* 7: 47-86.

Bruckman, A. (2000). "Situated Support for Learning: Storm's Weekend with Rachael." *Journal of the Learning Sciences* 9(3): 329-372.

Cary, M. S. (1978). "The Role of Gaze in the Initiation of Conversation." *Social Psychology* 41(3): 269-271.

Cassell, J. (1999). "Embodied Conversation: Integrating Face and Gesture into Automatic Spoken Dialogue systems." *Spoken Dialogue Systems.* Luperfoy. Cambridge, MA, MIT Press.

Cassell, J., H. Vilhjalmsson, et al. (1999). "Requirements for an Architecture for Embodied Conversational Characters." *Computer Animation and Simulation '99.* N. Magnenat-Thalmann and D. Thalmann. Vienna, Austria, Springer Verlag.

Cassell, J. and H. Vilhjalmsson (1999). "Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous." *Autonomouse Agents and Multi-Agent Systems* 2(1): 45-64.

Cassell, J., H. Vilhjalmsson, et al. (2001). "BEAT: the Behavior Expression Animation Toolkit." SIGGRAPH01, Los Angeles, CA, ACM.

Chang, J. (1998). *Action Scheduling in Humanoid Conversational Agents.* Master of Engineering. MIT. Cambridge, MA.

Cherny, L. (1995). *The MUD Register: Conversational Modes of Action in a Text-Based Virtual Reality.* PhD. Stanford University.

Cherny, L. (1999). *Conversation and community : chat in a virtual world.* Stanford, Calif., CSLI Publications.

Chi, D., M. Costa, et al. (2000). "The EMOTE model for Effort and Shape." SIGGRAPH, New Orleans, ACM.

Chovil, N. (1991). "Discourse-Oriented Facial Displays in Conversation." *Research on Language and Social Interaction* 25(1991/1992): 163-194.

Clark, H. H. (1996). *Using Language,* Cambridge, UK, Cambridge University Press.

Clark, H. H. (2001). "Pointing and Placing." *Pointing: Where language, culture, and cognition meet.* K. Sotaro. Mahwah, NJ, Lawrence Erlbaum Associates: 2-25.

Clark, H. H. and S. E. Brennan (1991). "Grounding in Communication." *Perspectives on Socially Shared Cognition.* L. B. Resnick, J. M. Levine and S. D. Teasley. Washington, American Psychological Association: 127-149.

Colburn, A. R., M. F. Cohen, et al. (2000). "The Role of Eye Gaze in Avatar Mediated Conversational Interfaces." MSR-TR-2000-81. Technical Report. Microsoft Research. Seattle.

Colburn, A. R., M. F. Cohen, et al. (2001). "Graphical Enhancements for Voice Only Conference Calls." MSR-TR-2001-95. Technical Report. Microsoft Corporation. Redmond, WA.

Cugini, J., L. Damianos, et al. (1999). "Methodology for Evaluation of Collaborative Systems." MITRE. Bedford, MA

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

Damer, B. (1997). *Avatars!: Exploring and Building Virtual Worlds on the Internet*, Peachpit Press.

Damer, B. (1997). "Putting a Human Face on Cyberspace: Designing Avatars and the Virtual Worlds They Live In." *SIGGRAPH 97*. S. DiPaola, J. Paniaras, K. Parsons, B. Roel and M. Ma. Los Angeles, CA.

Damer, B. (1998). "Avatars! : exploring and building virtual worlds on the Internet." Berkeley, CA, Peachpit Press.

Damianos, L., J. Drury, et al. (2000). "Evaluating Multi-party Multi-modal Systems." Technical. MITRE

Darrell, T., S. Basu, et al. (1997). "Perceptually-driven Avatars and Interfaces: active methods for direct control." 416. Perceptual Computing Section Techincal Report. MIT Media Laboratory. Cambridge, MA.

Dickey, M. D. (1999). *3D Virtual Worlds and Learning: An analysis of the impact of design affordances and limitations in Active Worlds, Blaxxun Interactive and Onlive! Traveler; and a study of the implementation of Active Worlds for formal and informal education*. Dissertation. The Ohio State University. Columbus, OH

Dodge, M. (1998). "Avatars, Identity and Meta-Place: The Geography of a 3-D Virtual World on the Internet." Place and Identity in Age of Technologically Regulated Movement, Santa Barbara.

Doherty-Sneddon, G., A. H. Anderson, et al. (1997). "Face-to-Face and Video-Mediated Communication: A Comparison of Dialogue Structure and Task Performance." *Journal of Experimental Psychology: Applied* 3(2): 105-125.

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

Donath, J. (2001). "Mediated Faces." Cognitive Technology: Instruments of Mind, Warwick, UK, Springer-Verlag, Berlin.

Donath, J. (2002). "A Semantic Approach To Visualizing Online Conversations." *Communications of the ACM* 45(4): 45-49.

Dourish, P. and S. Bly (1992). "Portholes: Supporting Awareness in a Distributed Work Group." Conference on Human Factors in Computer Systems CHI'92, Monterey, California.

Duncan, S. (1974). "On the structure of speaker-auditor interaction during speaking turns." *Language in Society* 3: 161-180.

Garau, M., M. Slater, et al. (2001). "The Impact of Eye Gaze on Communication using Humanoid Avatars." CHI 2001, Seattle, WA, ACM.

Garcia, A. and J. B. Jacobs (1998). "The Interactional Organization of Computer Mediated Communication in the College Classroom." *Qualitative Sociology* 21(3): 299-317.

Gibson, W. (1994). *Neuromancer*. New York, Ace Books.

Goffman, E. (1963). *Behavior in public places; notes on the social organization of gatherings*. [New York], Free Press of Glencoe.

Goffman, E. (1983). *Forms of Talk*. Philadelphia, PA, University of Pennsylvania Publications.

Goodwin, C. (1981). *Conversational Organization: Interaction between speakers and hearers*. New York, Academic Press.

Gorniak, P. (2000). "Classpaper: Learning Avatars." Classpaper. MIT. Cambridge, MA.

Grosz, B. and C. Sidner (1986). "Attention, Intentions, and the Structure of Discourse." *Computational Linguistics* 12(3): 175-204.

Grosz, B. J. (1981). "Focusing and description in natural language dialogues." *Elements of Discourse Understanding*. A. K. Joshi, B. L. Webber and I. A. Sag. Cambridge, UK, Cambridge University Press: 84-105.

Halliday, M. A. K. and R. Hasan (1976). *Cohesion in English*, Longman.

H-Anim (2001). "Specification for a Standard Humanoid." Web Document. (http://h-anim.org/Specifications/H-Anim1.1/). VRML Humanoid Animation Working Group

Herbsled, J. D., D. L. Atkins, et al. (2002). "Introducing Instant Messaging and Chat in the Workplace." CHI, Minneapolis, MN, ACM.

Herring, S. C. (1996). *Computer-mediated communication : linguistic, social, and cross-cultural perspectives*. Amsterdam; Philadelphia, J. Benjamins.

Hirschberg, J. (1990). "Accent and Discourse Context: Assigning Pitch Accent in Synthetic Speech." AAAI 90.

Hirschberg, J. and D. Litman (1993). "Empirical Studies on the Disambiguation of Cue Phrases." *Computational Linguistics* 19(3): 501-530.

Hiyakumoto, L., S. Prevost, et al. (1997). "Semantic and Discourse Information for Text-to-Speech Intonation." ACL Workshop on Concept-to-Speech Technology.

Hughes, C. E. and J. M. Moshell (1997). "Shared Virtual Worlds for Education: The ExploreNet Experiment." *ACM Multimedia*.

Inoue, T., k.-i. Okada, et al. (1997). "Integration of Face-to-Face and Video-Mediated Meetings: HERMES." *Proceedings of ACM SIGGROUP'97*: 385-394.

Isaacs, E. A. and J. C. Tang (1994). "What video can and cannot do for collaboration: a case study." *Multimedia Systems* 2: 63-73.

Isaacs, E. A. and J. C. Tang (1997). "Studying video-based collaboration in context: From small workgroups to large organization." *Video-Mediated Communication*. K. Finn, A. Sellen and S. Wilbur, Lawrence Erlbaum Associates, Inc.: 173-197.

Johnson, M. P., A. Wilson, et al. (1999). "Sympathetic Interfaces: Using a Plush Toy to Direct Synthetic Characters." *Proceedings of CHI'99*: 152-158.

Kendon, A. (1987). "On Gesture: Its Complementary Relationship With Speech." *Nonverbal Behavior and Communication*. A. W. Siegman and S. Feldstein. Hillsdale, Lawrence Erlbaum Associates, Inc.: 65-97.

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

Kendon, A. (1990). "The negotiation of context in face-to-face interaction." *Rethinking context: language as interactive phenomenon*. A. Duranti and C. Goodwin. New York, Cambridge University Press: 323-334.

Kendon, A. (1996). "Cues of Context." *Semiotica* 109(3/4): 349-356.

Kobayashi, M. and H. Ishii (1993). "ClearBoard: A Novel Shared Drawing Medium that Supports Gaze Awareness in Remote Collaboration." *IEICE Transactions on Communications* E76-B(6): 609-617.

Koda, T. and P. Maes (1996). "Agents with faces: The effect of personification." Fifth IEEE International Workshop on Robot and Human Communication.

Krauss, R. M. and S. R. Fussell (1991). "Constructing Shared Communicative Environments." *Perspectives on Socially Shared Cognition*. L. B. Resnick, J. M. Levine and S. D. Teasley. Washington, American Psychological Association: 172-200.

Kurlander, D., T. Skelly, et al. (1996). "Comic Chat." *Proceedings of SIGGRAPH'96*: 225-236.

Lasseter, J. (1987). "Principles of traditional animation applied to 3D computer animation." SIGGRAPH, ACM.

Lee, A., A. Girgensohn, et al. (1997). "NYNEX Portholes: Initial user reactions and redisign implications." *Proceedings of ACM SIGGROUP'97*: 385-349.

Lee, C., S. Ghyme, et al. (1998). "The Control of Avatar Motion Using Hand Gesture." VRST '98, Tapei, Taiwan, ACM.

Lehtinen, E., K. Kakkarainen, et al. (1998). "Computer supported collaborative learning: A Review." CL-Net Project

Lester, J. C., J. L. Voerman, et al. (1999). "Deictic Believability: Coordinated Gesture, Locomotion, and Speech in Lifelike Pedagogical Agents." *Applied Artificial Intelligence* 13(4-5): 383-414.

Lindeman, B., T. Kent, et al. (1995). "Exploring Cases On-line with Virtual Environments." Computer Support for Collaborative Learning '95, Bloomington, IN.

Machover, T. (1991). "Hyperinstruments: A Composer's Approach to the Evolution of Intelligent Musical Instruments." *Cyberarts*. W. Freeman. San Francisco, CA.

McCarthy, J., Miles, V., Monk, A., Harrison, M., Dix, A., Wright, P. (1993). "Text-based on-line conferencing: a conceptual and empirical analysis using a minimal prototype." *Human-Computer Interaction* 8(2): 147-183.

McClave, E. Z. (2000). "Linguistic function of head movements in the context of speech." *Journal of Pragmatics* 32: 855-878.

McClean, P., B. Saini-Eidukat, et al. (2001). "Virtual Worlds In Large Enrollment Science Classes Significantly Improve Authentic Learning." The 12th International Conference on College Teaching and Learning.

McGrath, J. E. (1984). *Groups: Interaction and performance*. Englewood Cliffs, NJ, Prentice-Hall.

McNeill, D. (1992). *Hand and Mind*. Chicago and London, The University of Chicago Press.

Moreno, R., R. E. Mayer, et al. (2000). "Life-Like Pedagogical Agents in Constructivist Multimedia Environments: Cognitive Consequences of Their Interaction." World Conference on Educational Multimedia, Hypermedia, and Telecommunications, Montreal, Canada.

Morningstar, C. and F. R. Farmer (1990). "The Lessons of Lucasfilm's Habitat." The First Annual International Conference on Cyberspace.

Nakanishi, H., C. Yoshida, et al. (1996). "Free Walk: Supporting casual meetings in a network." *Proceedings of ACM CSCW'96*: 308-314.

Nardi, B. A. and S. Whittaker (2002). "The Place of Face-to-Face Communication in Distributed Work." *Distributed work: New ways of working across distance using technology*. P. Hinds and S. Kiesler. Cambridge, MA, MIT Press.

172

Nardi, B. A., S. Whittaker, et al. (2000). "Interaction and Outeraction: Instant Messaging in Action." Computer Supported Cooperative Work, ACM.

Neale, D. C. and M. K. McGee (1998). "Making Media Spaces Useful: Video Suppor and Telepresence." HCIL-98-02. Hypermedia Technical Report. Human-Computer Interaction Laboratory, Virginia Tech

O'Conaill, B. and S. Whittaker (1997). "Characterizing, Predicting, and Measuring Video-Mediated Communication: A Conversational Approach." *Video-Mediated Communication*. K. Finn, A. Sellen and S. Wilbur, Lawrence Erlbaum Associates, Inc.: 23-49.

Oviatt, S. L. and P. R. Cohen (1991). "Discourse structure and performance efficiency in interactive and noninteractive spoken modalities." *Computer Speech and Language* 5(4): 297-326.

Paulos, E. and J. Canny (1998). "PRoP: Personal Roving Presence." SIGCHI, ACM.

Polanyi, L. (1988). "A Formal Model of the Structure of Discourse." *Journal of Pragmatics* 12: 601-638.

Prevost, S. and M. Steedman (1994). "Specifying intonation from context for speech synthesis." *Speech Communication* 15: 139-153.

Prince, E. P. (1981). "Toward a Taxonomy of Given-New Information." *Radical Pragmatics*. Cole, Academic Press: 223-255.

Reeves, B. and C. Nass (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* Cambridge, UK, Cambridge University Press.

Rickel, J. and W. L. Johnson (1998). "Task-Oriented Dialogs with Animated Agents in Virtual Reality." *Proceedings of the 1st Workshop on Embodied Conversational Characters*: 39-46.

Rosenfeld, H. M. (1987). "Conversational Control Functions of Nonverbal Behavior." *Nonverbal Behavior and Communication*. A. W. Siegman and S. Feldstein. Hillsdale, Lawrence Erlbaum Associates, Inc.: 563-601.

Roussos, M., A. Johnson, et al. (1998). "Learning and Building Together in an Immersive Virtual World." *Presence* 8(3): 247-263.

Russell, M. C. and C. G. Halcomb (2002). "Bringing the Chat Room to the Classroom." *Usability News* 4(2).

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

Schiffrin, D. (1987). *Discourse markers.* Cambridge, UK, Cambridge University Press.

Schourup, L. (1999). "Discourse Markers." *Lingua* 1999(107): 227-265.

Smith, M., J. J. Cadiz, et al. (2000). "Conversation trees and threaded chats." Computer Supported Cooperative Work, ACM.

Spears, L. (2001). "The Pedagogy of On-Line Instruction in Laboratory Science." The 12th International Conference on College Teaching and Learning.

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

Straus, S. G. (1997). "Technology, Group Process, and Group Outcomes: Testing the Connections in Computer-Mediated and Face-to-Face Groups." *Human-Computer Interaction* 12: 227-266.

Sugawara, S., G. Suzuki, et al. (1994). "InterSpace: Networked Virtual World for Visual Communication." *IEICE Transactions on Information and Systems* E77-D(12): 1344-1349.

Suler, J. (1996). "Life at the Palace: A Cyberpsychology Case Study." Website. (http://www.rider.edu/users/suler/psycyber/palacestudy.html). Department of Psychology, Rider University

Suthers, D. D. (2001). "Collaborative Representations: Supporting Face to Face and Online Knowledge-building Discourse." 34th Hawai`i International Conference on the System Sciences, Maui, Hawai`i, IEEE.

Takeuchi, A. and T. Naito (1995). "Situated facial displays: Towards social interaction." CHI'95, ACM.

Taylor, M. J. and S. M. Rowe (2000). "Gaze Communication using Semantically Consistent Spaces." CHI 2000, The Hague, The Netherlands, ACM.

Thorisson, K. (1996). *Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills*. Ph.D. Dissertation. Massachusetts Institute of Technology. Cambridge, MA.

Torres, O. E., J. Cassell, et al. (1997). "Modeling Gaze Behavior as a Function of Discourse Structure." First International Workshop on Human-Computer Conversation.

Turkle, S. (1995). *Life on the screen : identity in the age of the Internet*. New York, Simon & Schuster.

Umaschi Bers, M. (1999). "Zora: a Graphical Multi-user Environment to Share Stories about the Self." Computer Support for Collaborative Learning, Stanford University, Palo Alto, CA, Lawrence Erlbaum Associates.

Vertegaal, R. (1999). "The GAZE Groupware System: Mediating Joint Attention in Multiplarty Communcation and Collaboration." CHI'99, Pittsburgh, PA, ACM.

Vertegaal, R. and Y. Ding (2002). "Explaining Effects of Eye Gaze on Mediated Group Conversations: Amount or Synchronization." CSCW 2002, New Orleans, LA, ACM.

Viegas, F. and J. Donath (1999). "Chat Circles." *Proceedings of CHI'99*: 9-16.

Vilhjalmsson, H. (1997). *Autonomous Communicative Behaviors in Avatars*. MS Thesis. Massachusetts Insitute of Technology. Cambridge, MA.

Vronay, D., M. Smith, et al. (1999). "Alternative Interfaces for Chat." UIST'99, Asheville, NC, ACM.

Waters, R. C. and J. W. Barrus (1997). "The rise of shared virtual environments." *IEEE Spectrum*(March 1997): 20-25.

Werry, C. C. (1996). "Linguistic and Interactional Features of Internet Relay Chat." *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspective*. S. C. Herring. Amsterdam, John Benjamins: 47-63.

Whittaker, S. (2002). "Evaluating and explaining the quality of computer mediated communication using communication affordances and discourse tagging." ISLE Workshop on Dialogue Tagging for Human Computer Interaction, Edinburgh, Scotland.

Whittaker, S. and B. O'Conaill (1997). "The Role of Vision in Face-to-Face and Mediated Communication." *Video-Mediated Communication*. K. Finn, A. Sellen and S. Wilbur, Lawrence Erlbaum Associates, Inc.: 23-49.

Appendix A:
Overview of Function and Behavior tags

Discourse Function

Tag	Attributes	Explanation
UTTERANCE		The entire text message being sent
	SPEAKER	The sender of the message
	HEARER	The addressee of the message
	SCENE	Identifies where the conversation is taking place
W		A single word token
	LEM	The neutral form of the word (*lemma*)
	POS	The part-of-speech class in the standard Penn Treebank code
	SYN	Light syntax code
NEW		An open classed (nouns, verbs, adjectives) lexical item that is seen for the first time in the discourse
ACTION		Verb phrase
	ID	The name of a related action class described in the Knowledge Base
OBJECT		Discourse entity
	ID	The unique ID of the discourse entity
REFERENCE		Identifies how a discourse entity is evoked using Prince's taxonomy (Prince 1981)
	TYPE	Either VISUAL or TEXTUAL depending on whether the entity is already part of visual or textual context
	ID	The unique ID of the discourse entity
	SOURCE	If textually evoked, this is the ID of the person who last referred to the entity
CONTRAST		A word that contrasts with a word occurring previously in the discourse
	ID	If the contrast is with another word in the same utterance, both words get the same ID
EMPHASIS		Particular attention is drawn to this part of the utterance
	TYPE	The unit being emphasized, either a WORD or an entire PHRASE
CLAUSE		A clause or proposition
	TYPE	The general communicative purpose of the clause, currently only QUESTION or EXCLAMATION
THEME		The part of a clause that ties it to preceding discourse
RHEME		The part of a clause that contains a new contribution to the discourse
TOPICSHIFT		Movement within the discourse structure
	TYPE	The type of movement can be NEXT, PUSH, POP, DIGRESS and RETURN
TURN		Negotiation of the floor
	TYPE	The floor negotiation action taken, can be TAKE, KEEP, REQUEST or GIVE
GROUNDING		Maintaining the communication channel or verify understanding
	TYPE	REQUEST or GIVE back channel feedback, or AFFIRM understanding

Communicative Behavior		
Tag	**Attributes**	**Explanation**
EYEBROWS		Raising the eyebrows
HEADNOD		Nodding the head
GAZE		Move eyes and head to look in a certain direction
	TYPE	Can either be LOOK, for tracking a target, or GLANCE, for a brief glance. Can also be AWAY for a "up and away" thinking glance
	TARGET	The ID of the object being looked at
POSTURESHIFT		Changing the posture of the body
	BODYPART	Can either be just the UPPER part, LOWER part or BOTH
	ENERGY	How much overall energy or effort is put into the movement
GESTURE_BOTH		Gesture requiring both hands
	LEFT_HANDSHAPE	The identifier for a handshape
	LEFT_TRAJECTORY	The identifer for a gesture trajectory
	RIGHT_HANDSHAPE	The identifier for a handshape
	RIGHT_TRAJECTORY	The identifer for a gesture trajectory
	TYPE	Can be BEAT, ICONIC, or DEICTIC
GESTURE_LEFT		Gesture for the left hand
	LEFT_HANDSHAPE	The identifier for a handshape
	LEFT_TRAJECTORY	The identifer for a gesture trajectory
	TYPE	Can be BEAT, ICONIC, or DEICTIC
GESTURE_RIGHT		Gesture for the right hand
	RIGHT_HANDSHAPE	The identifier for a handshape
	RIGHT_TRAJECTORY	The identifer for a gesture trajectory
	TYPE	Can be BEAT, ICONIC, or DEICTIC
INTONATION_ACCENT		A pitch accent
	ACCENT	The type of accent (e.g., "H*")
INTONATION_BREAK		A brief pause
INTONATION_TONE		The tone of an entire intonation phrase
	ENDTONE	The type of endtone (e.g., "L-L%")

Appendix B:
Comparing MapChat output to face-to-face data

FACE TO FACE		Allright	so	we	have	a	month	to	get	over	there
	A										
	Head						NOD				
	Gesture										Windmill
	Gaze	Map									
	Posture										
	B										
	Head										
	Gesture										
	Gaze	Map									
	Posture										
	C										
	Head										
	Gesture										
	Gaze	Map									
	Posture										

MAP-CHAT		Allright	so	we	have	a	month	to	get	over	there
	A										
	Head	NOD					NOD		NOD		
	Gesture	BEAT					BEAT		BEAT		
	Gaze	Away			C						
	Posture										
	B										
	Head										
	Gesture										
	Gaze	A				C					
	Posture										
	C										
	Head						NOD				
	Gesture										
	Gaze	A									
	Posture										

183

	Who	do	you	think	of	this	way	or	this	way	this	way	which	way	you	want	to	take
A Head																		
Gesture																		
Gaze	Map																	
Posture																		
B Head																		
Gesture																		
Gaze	Map														Info			
Posture																		
C	Who	do	you	think	of	this	way	or	this	way	this	way	which	way	you	want	to	take
Head																		
Gesture	Mountain						Mountain path											
Gaze	Map																	
Posture																		

	Who	do	you	think	of	this	way	or	this	way	this	way	which	way	you	want	to	take
A Head			NOD															
Gesture																		
Gaze	C		C															
Posture																		
B Head																		
Gesture																		
Gaze	C		A															
Posture																		
C	Who	do	you	think	of	this	way	or	this	way	this	way	which	way	you	want	to	take
Head				NOD			NOD									NOD		NOD
Gesture				BEAT			BEAT									ICONIC		BEAT
Gaze	Away		A															
Posture																		

184

A Ok I recall my instructions saying that we can only go third of normal walkin pace when you go through the mountains
Head NOD NOD NOD NOD
Gest BEAT Mountains
Gaze Map C Map
Postu

B
Head
Gest
Gaze Map A Map Info
Postu

C
Head
Gest
Gaze Map A Map A Map Info
Postu

A Ok I recall my instructions saying that we can only go third of normal walkin pace when you go through the mountains
Head NOD NOD NOD NOD Mountains
Gest BEAT BEAT BEAT BEAT NOD
Gaze C C C Mountains
Postu

B
Head
Gest
Gaze A C Mountains
Postu

C
Head NOD NOD NOD
Gest
Gaze A Mountains
Postu

The utterance across the top: **Ok · so · it · will · slow · us · down · going · through · the · mountains**

First segment (yellow)

A
Modality	Annotation
Head Gesture	NOD (at "us/down"); BEAT + Mountains (at "down/going/through")
Gaze	Map (spanning most of the utterance)
Posture	

B
Modality	Annotation
Head Gesture	
Gaze	Info (start) → Map (from "us" onward)
Posture	

C
Modality	Annotation
Head Gesture	Mountain (path) (from "going/through" onward)
Gaze	Map (spanning most of the utterance)
Posture	

Second segment (green)

A
Modality	Annotation
Head Gesture	NOD (at "will"); NOD (at "slow"); BEAT (at "will"); Mountains (at "mountains")
Gaze	C — Away — C ... Mountains (at end)
Posture	

B
Modality	Annotation
Head Gesture	
Gaze	A — C ... Mountains (at end)
Posture	

C
Modality	Annotation
Head Gesture	NOD (at "Ok"); NOD (at "it")
Gaze	A ... Mountains (at end)
Posture	

A		
Head		
Gesture		
Gaze	Map	
Posture		

B		
Head		
Gesture		
Gaze	Map	
Posture		

C	So if we go this way then we will have a choic throu the swa or choo it throu or we can stay in the tent
Head	NOD ... NOD ... NOD ... Shake
Gesture	MOUNTAIN (path) ... TENT to SWAMP (path) ... Tent BEA Tent BEAT
Gaze	Map
Posture	

A		
Head	NOD ... NOD ... NOD ... NOD	
Gesture		
Gaze	C ... Swamp C ... Tent	
Posture		

B		
Head		
Gesture		
Gaze	A ... Swamp C A ... Tent	
Posture		

C	So if we go this way then we will have a choic throu the swamp or choo it throu or we can stay in the tent
Head	NOD NOD NOD NOD
Gesture	BEA Swa mp BEA Tent
Gaze	A Away A Swamp A Away Tent
Posture	HIGH

A | Head | Gesture | Gaze: Map | Posture

B | What is does anybody know anything about this tent | Head | Gesture: Tent | Gaze: Map | Posture

C | Head | Gesture | Gaze: Map | Posture

A | Head | Gesture | Gaze: B | C | Tent | Posture

B | What is does anybody know anything about this tent | Head: NOD | Gesture: BEAT | Tent | Gaze: Away | Away | C | Posture

C | Head | Gesture: NOD | Gaze: B | Tent | Posture

Yeah I know somethin about this tent ok we will see if I can recall about this tent

A
- Head
- Gesture: Tent | NOD | Tent
- Gaze: Map | Away | Map
- Posture

B
- Head
- Gesture: NOD | NOD
- Gaze: Map | Info
- Posture

C
- Head
- Gesture: NOD
- Gaze: Info | Map | A | Info | Map
- Posture

Yeah I know somethin about this tent ok we will see if I can recall about this tent

A
- Head
- Gesture: Tent | NOD / BEAT | Tent
- Gaze: Away | B | B
- Posture

B
- Head
- Gesture: NOD | NOD
- Gaze: A | Tent | A | Tent
- Posture

C
- Head
- Gesture
- Gaze: A | Tent | A | B | Tent
- Posture

A — There is some guy some dude over there who can who can give us a map a detail map of the desert yeah
Head
Gesture — NOD ... NOD
Gaze — Map ... Info ... Map ... C
Posture

B
Head
Gesture
Gaze — Info ... Map
Posture

C — Oh is it desert
Head
Gesture
Gaze — Map ... A ... Map ... Away ... Map ... A
Posture

A — There is some guy some dude over there who can who can give us a map a detail map of the desert yeah
Head
Gesture — NOD / BEAT ... NOD / BEAT ... Desert / NOD
Gaze — Away ... B ... Away ... B ... Desert
Posture

B
Head
Gesture — NOD ... NOD
Gaze — A ... Desert
Posture

C
Head
Gesture
Gaze — A ... B ... Desert
Posture

Appendix C:
Questionnaires from user study

SPARK: QUESTIONNAIRE A (TRIAL QUESTIONNAIRE)

Thank you for using the Spark system. Following are questions related to your experience you just had. Your participation is voluntary and you are not required to answer all or in fact any of these questions, but we would appreciate it if you would answer them to the best of your ability. Your answers are confidential and your anonymity will be assured.

I. GENERAL "FEELING"

To what extent do the following words describe **your experience** while using Spark? (Circle one dot for each word)

	Not at all								Extremely
1. Boring	●	●	●	●	●	●	●	●	●
2. Difficult	●	●	●	●	●	●	●	●	●
3. Easy	●	●	●	●	●	●	●	●	●
4. Engaging	●	●	●	●	●	●	●	●	●
5. Enjoyable	●	●	●	●	●	●	●	●	●
6. Confusing	●	●	●	●	●	●	●	●	●
7. Exciting	●	●	●	●	●	●	●	●	●
8. Friendly	●	●	●	●	●	●	●	●	●
9. Immersive	●	●	●	●	●	●	●	●	●
10. Frustrating	●	●	●	●	●	●	●	●	●
11. Fun	●	●	●	●	●	●	●	●	●
12. Intuitive	●	●	●	●	●	●	●	●	●
13. Alive	●	●	●	●	●	●	●	●	●
14. Entertaining	●	●	●	●	●	●	●	●	●
15. Tedious	●	●	●	●	●	●	●	●	●
16. Warm	●	●	●	●	●	●	●	●	●

II. GENERAL INTERACTION (circle one dot for each question)

1. How well do you feel you were able to understand what the other participants were saying?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   very well

2. How well do you feel you were able to **express yourself** with the other participants?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   very well

3. How well do you think the others understood you and understood what you meant to communicate?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   very well

4. How well do you feel the other participants were able to **express themselves with you**?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   very well

5. In general, how well do you think this system allowed you to communicate what you need to say?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   very well

6. How much **control did you have over the conversation**?

No control   ●   ●   ●   ●   ●   ●   ●   ●   ●   total control

7. How much control do you think the other participants had over the conversation?

Not control   ●   ●   ●   ●   ●   ●   ●   ●   ●   total control

8. How much did the interaction feel like a face to face conversation?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   very much

9. Ho strong was the feeling of you being in the tower?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   very well

The avatar [only in avatar condition]

1. How useful to the interaction do you think the avatars were?

Not at all   •   •   •   •   •   •   •   •   •   very

2. How **natural** did the avatar gesture seem?

Not at all natural   •   •   •   •   •   •   •   •   •   very natural

3. Were you paying attention to other people's avatars?

Not at all   •   •   •   •   •   •   •   •   •   all the time

The voice [only in voice condition]

1. How natural were the voices?

Not at all natural   •   •   •   •   •   •   •   •   •   very natural

2. How well did you **understand** the voices?

Not at all   •   •   •   •   •   •   •   •   •   very well

The text [only in text condition]

1. How much of the text messages were you able to read?

None   •   •   •   •   •   •   •   •   •   all

2. How **easy** was it to read the text messages?

Very hard   •   •   •   •   •   •   •   •   •   very easy

III. THE TASK (circle one dot for each question)

1. How difficult was the task?

Very easy   •   •   •   •   •   •   •   •   •   very difficult

2 How well do you think the **group performed** on the task you were given?

Not at all well   •   •   •   •   •   •   •   •   •   very well

3. How much do you think **you contributed** to the final solution?

Nothing   •   •   •   •   •   •   •   •   •   a lot

4. How **satisfied** are you with the final solution?

Very unsatisfied   •   •   •   •   •   •   •   •   •   very satisfied

5. How strong do you think the group's **consensus** is about the final solution?

None at all   •   •   •   •   •   •   •   •   •   very strong

6. How **efficiently** did the group solve the task?

Very inefficient   •   •   •   •   •   •   •   •   •   very efficient

7. How certain are you that the final solution you came up with is the **best solution**?

Not at all   •   •   •   •   •   •   •   •   •   very

8. If you had been solving the task **face-to-face**, do you think the solution would have been?

Much worse   •   •   •   •   •   •   •   •   •   much better

9. If you had been solving the task using **regular text messaging**, do you think the solution would have been?

Much worse   •   •   •   •   •   •   •   •   •   much better

IV. THE OTHER PARTICIPANTS

You worked with two other people on solving the task. For each of these other participants, please answer the following questions. Circle one dot for each question.

PARTICIPANT COLOR : GREEN

1. How interested do you think the participant was in collaborating?

Not at all • • • • • • • • • Very

2. How comfortable were you collaborating with this participant?

Not at all • • • • • • • • • Very

3. How rich was the interaction with this participant?

Very poor • • • • • • • • • Very rich

4. How helpful was this participant in solving the task?

Not at all • • • • • • • • • Very

5. How honest do you think this participant is?

Not at all • • • • • • • • • Very

6. How well did you trust this participant?

Not at all • • • • • • • • • A lot

7. How well did this participant listen to others?

Not at all • • • • • • • • • Very well

8. Would you want to meet this person in real life?

Not at all!! • • • • • • • • • Absolutely!!

PARTICIPANT COLOR : **BLUE**

1. How interested do you think the participant was in collaborating?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   Very

2. How **comfortable** were you collaborating with this participant?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   Very

3. How rich was the interaction with this participant?

Very poor   ●   ●   ●   ●   ●   ●   ●   ●   ●   Very rich

4. How **helpful** was this participant in solving the task?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   Very

5. How honest do you think this participant is?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   Very

6. How well did you **trust** this participant?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   A lot

7. How well did this participant listen to others?

Not at all   ●   ●   ●   ●   ●   ●   ●   ●   ●   Very well

8. Would you want to **meet this person** in real life?

Not at all!!   ●   ●   ●   ●   ●   ●   ●   ●   ●   Absolutely!!

SPARK: QUESTIONNAIRE B (PREFERENCE QUESTIONNAIRE)

Thank you for using two versions of the Spark system. Following are questions related to your experience of using two different systems. We also include some questions regarding demographics and background. Again, your participation is voluntary and you are not required to answer all or in fact any of these questions, but we would appreciate it if you would answer them to the best of your ability. As before, your answers are confidential and your anonymity will be assured.

I. COMPARISON

A.   ID of first system you used   _____ (filled in by experimenter)

B.   ID of second system you used   _____ (filled in by experimenter)

For each of the criteria listed below, please circle one dot representing the strength of your preference for the first (A) or second (B) system (middle denotes no preference).

1. More useful

System A   •   •   •   •   •   •   •   •   •   System B

2. More fun

System A   •   •   •   •   •   •   •   •   •   System B

3. More personal

System A   •   •   •   •   •   •   •   •   •   System B

4. Easier to use

System A   •   •   •   •   •   •   •   •   •   System B

5. More efficient

System A   •   •   •   •   •   •   •   •   •   System B

6. Easier to communicate

System A   •   •   •   •   •   •   •   •   •   System B

7. Which one would you use again? (circle one)

System A   System B   Both

II. BACKGROUND

1.   Gender: _____

2.   Age: _____

3.   Occupation: _____

4.   How many hours a day do you use a computer?

        a.   Less than 1
        b.   1-4
        c.   5 or more

5.   What do you primarily use a computer for? (circle any that apply)

        a.   Email
        b.   Word processing
        c.   Web
        d.   Games
        e.   Other: _____

6.   Have you used a **text** chat program (AOL or MSN Instant Messenger) before?

    Yes
        If yes, which one _____

    No

7.   Have you used a **graphical** chat program (similar to the one you just tried) before?

        Yes
        If yes, which one _____

        No

Appendix D:
Summary of Means and ANOVA tests

| | Overall Means | | | | ANOVA Main Effects for Avatars | | | | | | | | | | | | | | | | | | |
| | Avatar | | No Avatar | | Pooled | | | | Map 1 | | | | Map 2 | | | | Order 1 | | | | Order 2 | | | |
	M	SD	M	SD	F	Sig.	h2	Pow.	F	Sig.	h2	Pow.	F	Sig.	h2	Pow.	F	Sig.	h2	Pow.	F	Sig.	h2	Pow.
BEHAVIORAL DATA	N=15		N=16		DF=27				DF=11				DF=12											
Portion of grounding utterances	0.09	0.06	0.12	0.06	2.21	0.15	0.08	0.30	0.00	0.97	0.00	0.05	3.64	**0.08**	0.23	0.42								
Number of shared hints	10.87	3.58	10.50	4.31	0.27	0.61	0.01	0.08	0.98	0.34	0.08	0.15	0.32	0.58	0.03	0.08								
Max-Min number of contributions	12.33	7.78	11.19	7.61	0.24	0.63	0.01	0.08	0.01	0.91	0.00	0.05	1.77	0.21	0.13	0.23								
Portion of turns with explicit handovers	0.35	0.07	0.38	0.11	0.54	0.47	0.02	0.11	1.02	0.33	0.08	0.15	0.14	0.72	0.01	0.06								
Portion of overlapping utterances	0.02	0.01	0.02	0.02	1.20	0.28	0.04	0.18	0.54	0.48	0.05	0.10	0.70	0.42	0.05	0.12								
Portion of adjacency pairs broken	0.20	0.12	0.26	0.18	1.95	0.17	0.07	0.27	0.98	0.34	0.08	0.15	0.62	0.45	0.05	0.11								
Portion of on-task utterances	0.74	0.10	0.73	0.10	0.43	0.52	0.02	0.10	0.03	0.86	0.00	0.05	0.97	0.34	0.07	0.15								
Quality of task solution	3.07	2.02	3.31	1.85	0.16	0.69	0.01	0.07	0.21	0.66	0.02	0.07	0.31	0.59	0.02	0.08								
Task completion time	15.87	5.44	17.06	9.07	0.07	0.79	0.00	0.06	0.00	1.00	0.00	0.05	0.43	0.52	0.03	0.09								
SELF-REPORT DATA*	N=43		N=41		DF=80				DF=29				DF=47				DF=40				DF=36			
Others ability to communicate (II1,II4)	6.0	1.8	5.9	1.7	0.36	0.55	0.00	0.09	0.04	0.85	0.00	0.05	0.71	0.40	0.01	0.13	0.74	0.40	0.02	0.13	0.08	0.79	0.00	0.09
Your ability to communicate (II2,II3,II5)	6.0	1.8	5.7	1.9	0.99	0.32	0.01	0.17	0.66	0.42	0.02	0.12	0.42	0.52	0.01	0.10	0.98	0.33	0.02	0.16	0.10	0.75	0.00	0.10
Sense of control over conversation (II6)	5.8	1.5	5.1	1.8	3.17	**0.08**	0.04	0.42	2.35	0.14	0.07	0.32	2.10	0.15	0.04	0.29	1.00	0.32	0.02	0.16	2.03	0.16	0.05	0.36
Likeness of conversation to f2f (II8)	4.5	2.1	3.4	1.8	5.52	**0.02**	0.06	0.64	3.74	**0.06**	0.11	0.46	3.18	**0.08**	0.06	0.42	1.97	0.17	0.05	0.28	3.60	**0.07**	0.09	0.54
Difficulty of task (III1)	3.9	1.4	4.6	1.8	3.35	**0.07**	0.04	0.44	2.05	0.16	0.07	0.28	1.78	0.19	0.04	0.26	1.59	0.21	0.04	0.23	1.23	0.27	0.03	0.26
Efficiency of group (III2,III6)	6.7	1.3	5.8	1.4	11.6	**0.00**	0.13	0.92	6.83	**0.01**	0.19	0.71	4.72	**0.03**	0.09	0.57	5.92	**0.02**	0.13	0.66	6.44	**0.02**	0.15	0.77
Feeling of consensus (III5)	7.6	1.4	6.6	1.8	8.03	**0.01**	0.09	0.80	2.19	0.15	0.07	0.30	4.20	**0.05**	0.08	0.52	7.86	**0.01**	0.16	0.78	1.23	0.28	0.03	0.26
Satisfaction with solution (III4,III7)	6.5	1.8	6.0	1.7	2.44	0.12	0.03	0.34	0.05	0.82	0.00	0.06	1.75	0.19	0.04	0.25	0.47	0.50	0.01	0.10	2.43	0.13	0.06	0.41
Face-to-face would make task easier (III8)	6.5	1.7	7.0	1.5	1.91	0.17	0.02	0.28	8.59	**0.01**	0.23	0.81	0.44	0.51	0.01	0.10	0.07	0.79	0.00	0.06	3.16	**0.08**	0.08	0.50
Text-chat would make task easier (III9)	5.0	2.0	5.2	2.2	0.14	0.71	0.00	0.07	0.00	0.99	0.00	0.05	0.09	0.77	0.00	0.06	0.28	0.60	0.01	0.08	0.00	0.98	0.00	0.08
Other participant's effort (IV1,IV4,IV7)	7.0	1.1	6.2	1.5	8.03	**0.01**	0.10	0.80	1.30	0.26	0.04	0.20	5.13	**0.03**	0.10	0.60	8.05	**0.01**	0.17	0.79	0.87	0.36	0.02	0.20
Trust in other participants (IV2,IV5,IV6)	7.2	1.1	6.6	1.5	4.36	**0.04**	0.05	0.54	0.85	0.36	0.03	0.14	2.39	0.13	0.05	0.33	2.45	0.13	0.06	0.33	1.00	0.32	0.03	0.22
Tedious (I1,^I7,I15)	3.8	1.3	4.9	1.4	14.2	**0.00**	0.15	0.96	1.05	0.31	0.04	0.17	13.6	**0.00**	0.225	0.951	7.75	**0.01**	0.162	0.775	4.52	**0.04**	0.111	0.543
Difficult (I2,^I3,^I12)	3.6	1.4	4.3	1.6	5.65	**0.02**	0.07	0.65	4.56	**0.04**	0.14	0.54	2.46	0.12	0.05	0.336	3.76	**0.06**	0.086	0.473	1.63	0.21	0.043	0.238
Confusing (I6,I10)	3.6	1.8	4.2	1.7	3.21	**0.08**	0.04	0.42	0.95	0.34	0.03	0.16	1.72	0.20	0.035	0.251	1.39	0.24	0.034	0.211	1.4	0.24	0.037	0.21
Engaging (I4,I9,I13,II9)	5.6	1.4	4.7	1.4	8.69	**0.00**	0.10	0.83	3.47	**0.07**	0.11	0.44	2.57	0.12	0.05	0.35	3.59	**0.07**	0.08	0.46	5.96	**0.02**	0.142	0.661
Comfortable (I8,I16)	5.8	1.3	4.7	1.3	13.6	**0.00**	0.15	0.95	12.1	**0.00**	0.29	0.92	3.82	**0.06**	0.08	0.48	2.66	0.11	0.06	0.36	13.1	**0.00**	0.267	0.941
Entertaining (I5,I11,I14)	6.3	1.4	5.4	1.7	7.14	**0.01**	0.08	0.75	2.51	0.12	0.08	0.33	2.68	0.11	0.054	0.361	3.51	**0.07**	0.081	0.448	3.3	**0.08**	0.084	0.424

* Numbers in parenthesis refer to the corresponding question on the Trial Questionnaire. More than one number indicates an aggregate.