Generation of Non-Verbal Behavior for an Embodied Conversational Character

by

Sola Grantham

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of

Bachelor of Science in Electrical Engineering and Computer Science

and

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1998

© Massachusetts Institute of Technology 1998. All rights reserved.

Author .....
Department of Electrical Engineering and Computer Science
May 8, 1998

Certified by .....
Professor Justine Cassell
AT&T Career Development Professor of Media Arts & Sciences
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

ARCHIVES

Generation of Non-Verbal Behavior for an Embodied Conversational Character

by

Sola Grantham

Submitted to the Department of Electrical Engineering and Computer Science
on May 8, 1998, in partial fulfillment of the
requirements for the degrees of
Bachelor of Science in Electrical Engineering and Computer Science
and
Master of Engineering in Electrical Engineering and Computer Science

Abstract

This thesis discusses the generation of multi-modal conversational cues in animated computer characters. Drawing from psycholinguistic literature of behaviors observed in human-human interactions, a study of turn-taking cues in a human-computer interaction, and previous work in creating animated conversational agents, I attempt to formulate a set of behaviors and appropriate times to produce these behaviors for a new conversational character, Rea.

Thesis Supervisor: Professor Justine Cassell
Title: AT&T Career Development Professor of Media Arts & Sciences

Acknowledgments

I always know when I am talking to someone about something super important to me because I invariably get emotional, yet come out of the conversation the better for having had it. To those people who have dealt with me in tears, each in your own way, Justine (ignoring the tears and continuing the talk), mom (working the energy), and Cotton(saying it'll be ok and holding me), Thanks, you each did exactly the right thing and I needed you all. Also thanks to Erin who has provided the invaluable support that no one other than someone in the same group on the same project in a similar position, with similar life experience, and sense of humor, could provide. I would also like to thank the GNL+ group, particularly, Aalok (for helping me by excellent proofreading at the last moment when no one had time), Deepa (for a fun day of statistics and just being around), Glen (for a cheerful smile and help with all the confusing administrative stuff), Hannes (for always being easy going, but also always being there at three in the morning), Joey (for never seeming quite as stressed as everyone else), Julie (for bringing me into the group), Kimiko (for always cheering for me at racketball and being my friend), Lee (for data, and great prompt coding in a pinch), Nick (for playing with words), Schuyler (for amazing moral support and helping hands anytime he comes to the lab) and Stefan (for chocolate).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Statement of Purpose

As we move towards enabling people to interact with their computers the way they interact with other humans by providing faces, bodies and voices for computer characters, we must also provide mechanisms by which computers can use these newfound qualities. In this thesis, I discuss the generation of multi-modal conversational behavior for interactive computer characters. I claim that the way to know which behaviors to generate in a conversation and when to generate them should grow out of the study of human-human interactions, particularly since humans hold face-to-face conversations everyday using multimodal cues, often unconsciously, and thus may expect appropriate conversational cues when interacting with embodied human-like characters[1]. In addition, I believe that the key to learning which behaviors are important to implement is to look at the discourse functions served by the various behaviors. In this thesis, I draw on psycholinguistic literature, a human-computer interaction study and previous work in creating characters to formulate a set of behaviors for a new conversational character, Rea.

[1] A failure to provide appropriate behaviors to go along with a face, body and voice will often produce a very "broken", "wooden" or "mechanical" character, even if users are unsure why the character is not "alive" enough.

1.2 Human-Computer Conversation

Human-Computer conversation is an area of research that could be considered to fall within the field of human-computer interaction, yet means something much more specific. There is a wide range of imaginable interfaces between humans and computers. Today, we primarily interact via keyboard and mouse, but we can imagine an interface as fully human as the android Data of Star Trek. Somewhere between these interfaces lies the conversational computer character, an animated humanoid which resides on a screen, but can communicate with humans the same way we communicate with each other.

Human-human conversations occur on various channels ranging from exchanging grammatically correct text, to typed casual language in a chat system, to telephone conversation, to face-to-face exchange. Each of these media suggests difficult problems for computer researchers who want to make computers interact with humans in ways similar to humans, such as natural language understanding and generation, speech recognition and generation, planning, inference, and linguistics. It is in the arena of face-to-face interaction between human and computer that I wish to situate this thesis.

1.3 Multi-Modal Behavior

Consider for a moment a computer which can pass a Turing test; that is, it can convince humans through an exchange of text messages that it is human. Now consider how much more the computer must be able to do to pass a face-to-face Turing test. One of the key factors in considering face-to-face interactions is the multiple modalities with which we can send signals to each other, including speech, intonation, gesture and gaze. Not only can we use these channels as ways to get our message across, we can also use them to structure our interactions such that we can get feedback from our listeners and so we can smoothly transition from one speaker to another. As we work toward creating a conversational computer character, one of the things we will

need to learn is how to allow computers to understand and utilize these channels. This task requires the integration of a great many technologies, including computer sensing and animation (or robotics) along with speech and language processing. My goal in this thesis is to focus on one portion of this task, namely the generation of meaningful multimodal actions for the computer humanoid.

1.4 Goals of This Research

There are four important goals of the research presented in this thesis, and the larger research program of creating a conversational humanoid. The first is to create a platform for studying the relationship between verbal and non-verbal behavior. This aspect includes a strong influence from and contributions to work in psycholinguistics and discourse. The second is the study of computer understanding and generation of speech and associated behaviors. This touches on areas of computational linguistics such as natural language processing and generation. Third is design of architectures for embodied conversational systems. Emphasis in this aspect is on creating interactive and real-time systems. Lastly, we are examining the conversational character as an interface agent that takes advantage of social and communicative expectations. This perspective is based in the field of human-computer interaction.

1.5 What This Thesis is Not

In discussions with others unfamiliar with our research in multimodal communication, I recently heard a comment along these lines:

> Oh, so if I gave you a videotape of a person sitting alone in a room, maybe with some props, then you could tell me the personality or emotional state of the person, just from their non-verbal behavior?

This statement surprised me because although we had been talking to this person for quite awhile, there had clearly been a number of misunderstandings. I would like

to use this quote as a starting place to mention what is not being addressed in this thesis. To begin with, our work is on conversational behaviors. Thus a person alone, not speaking, will exhibit none of the phenomena introduced in section 2. Secondly, in this thesis, I will be discussing the generation of behaviors, rather than the detection of these behaviors[2] although the much of the background is the same. A third issue is that of real time interactivity versus off-line processing. In many of the functions I will be discussing, part of the importance lies in the ability of a system to show human-like reactivity even when faced with significantly longer speech processing times. Lastly, although I do believe that a person's emotions and personality can affect how they express themselves, the work in presented in this thesis is concentrated on developing a basic set of behaviors which are used in everyday conversations which does not encompass emotional variation. It focuses on similarities among English speakers rather than differences.

1.6 Layout of Thesis

In chapter 2, I introduce the non-verbal modalities gesture, intonation, gaze and facial expression and the discourse functions served by these modalities. In addition, I review previous psycholinguistic research about multimodal conversational cues observed in human-human interactions.

Chapter 3 surveys other related computer character research, and discusses how they differ in approach.

In chapter 4, I describe an experiment in which human subjects interact with a computer character. The study attempts to examine the effects of non-verbal behaviors on the subjects' impressions of the interaction and character.

Chapter 5 discusses three previous projects which use human-based multimodal behaviors in conversational characters.

In section 6, I describe Rea, a conversational character currently being developed.

[2]This was not actually a misunderstanding of the person, as he was asking about the MIT Gesture and Narrative Language Group's research program in general which does include trying to detect and understand these behaviors in humans as well.

I give particular attention to the Generation Module, the begininegs of which I have designed and implemented.

In section 7, I present future directions for work on that character and drawbacks in the design.

Chapter 2

Background - Communicative Behaviors

2.1 Modalities

In face to face conversations, people usually do not get information only from the words. Many other modalities also allow the conveying of information. These include intonation, gesture, gaze and facial expression. It has been argued that we do not need to study gesture and gaze because people are perfectly capable of speaking on the phone. An extension of this argument would be that we do not even need intonation, as we are in general also able to communicate by just exchanging text (this real time exchange of text to communicate is becomingly increasingly popular with the ability of people to use computers to contact distant parties). However, I believe that the question this example addresses is whether or not we can compensate for the lack of these non-verbal modalities. We can. However, the question of whether we use these modalities is better addressed by the example of trying to read a transcription of a phone call, or listen to a audio recording of a meeting. In both cases, it soon becomes clear that there is information missing. People can compensate for the lack of gesture, gaze, and intonation, and facial expression, but without reason, they don't; they use the information communicated in these channels. I believe that this is sufficient reason to study these behaviors. In this section, I will introduce the major

non-verbal modalities on which I will concentrate throughout this thesis. In the next section I will discuss the types of discourse functions served by information in these channels.

2.1.1 Intonation

Intonation is composed of factors present in speech other than the words themselves. Pierrehumbert and Hirschberg[33] present a system of intonational descriptions which distinguishes stress(accent placement), tune, phrasing, and pitch range. Stress indicates the relative prominence of syllables in an utterance. Important here is phrasal stress, which is dependent on the context of the word and utterance, as opposed to lexical stress, which operates at the word level, and is associated with the proper pronunciation of the word. Tune is the contour of fundamental frequency(f0) pitches over an intonational phrase. Phrasing is a way of breaking up an utterance into intonationally meaningful groups of words. Lastly, pitch range is the extent over which the f0 varies.

2.1.2 Gesture

Gesture, as I use it in this thesis, refers to those movements made by the hands in the context of speech. I am specifically not referring to the highly culturally specific emblematic gestures such as a thumbs-up gesture or codified gesture such as sign language. Rather, I am speaking of the relatively more amorphous gestures people make when speaking, often unconsciously. Gesture of this type has been argued to be an integral part of the process of language production, working in conjunction with speech to convey ideas[26][22].

McNeill[26] categorizes gestures during speech into four basic categories: iconics, metaphorics, beats, and deictics. Morphologically, iconics and metaphorics are typically gestures which represent concrete or abstract features of objects, actions or ideas through the shape or motion of the gesture. Deictics represent information spatially, referencing entities by location (real or established in the course of speech). Accord-

ing to McNeill, beat gestures differ from other types of gestures in that they retain the same form independent of content. They consist of a generally small up and down motion of the hand, rather than the preparation / stroke / relaxation phases of other gesture. I believe that because they lack the semantic information present in other gestures, timing infomation is particularly crucial.

The timing of gesture and intonational stress have been shown to be related[22][26], with the stroke of a gesture frequently occurring during or just previous to the primary intonational stress. Kendon also finds a phrasal level correspondence. Tuite[40] suggests that the timing of both intonation and gesture match an internal pulse, while McClave[25] demonstrates that the relationship between beats and pitch accents is not always in exact synchrony.

2.1.3 Head Movements

Hand gestures are not the only body movements important in conversation. Duncan[14] reports also transcribing head movements, shoulder movements, foot movements, and posture shifts in the course of dyadic, face-to-face conversations. Of these, I will focus primarily on head movements such as nodding, turning, and shaking. In addition, I will often refer to gaze, which is an indication of where the eyes, or eyes and head, are pointing.

2.1.4 Facial Movements

Movements of the face can be very expressive, and can include eyebrow raising and lowering, squinting or widening the eyes, blinking, wrinkling the nose, and shaping the mouth. Research on human facial movements has primarily been divided into three areas. The first is the association of particular sets of movements, facial expressions, with emotional states[15][34]. The second is the association of facial movements with discourse functions[13]. The last is the study of how to make the lips form appropriately in synchrony with speech. One system incorporating all three of these areas is described in "Generating Facial Expressions for Speech"[30]. In this thesis, I

will primarily discuss the second set of functions of facial movements.

2.2 Discourse Functions

The above introduction to the various important non-verbal modalities is relatively cursory. This is because I wish to emphasis not the modalities, but the discourse functions that can be served by behaviors in these modalities. In many cases the same discourse function could be served in more than one way. In fact, because there is frequently a way to accomplish the discourse function just by using more words, I will often give a verbal gloss of the non-verbal behaviors.

2.2.1 Interactional vs Propositional

When speaking of verbal and non-verbal communicative behavior it is often helpful to make a distinction between those words and actions which are primarily used for conveying the ideas from one person to the other, and those which are important to help control the flow of the interaction. In this thesis I will refer to the first set as those that serve propositional functions and the latter, interactional functions. I may also abbreviate this and say that a particular behavior is interactional[1]. However, what I mean is not that the morphology of the behavior is inherently interactional, but that the discourse function it is being used to accomplish is one of regulating the conversation.

An example of this distinction can be made in speech with the words, "I see". These words can be used to convey propositional content such as in "I see two sea gulls out over the ocean," where part of the content being transmitted is that the speaker can see the birds. However, in everyday conversation, a listener can also use the words "I see" to indicate that he is following the speaker's line of reasoning or even just that he is paying attention. This function is much more interactional than in the previous

[1]This use of the word interactional is similar to Shiffrin's "interactional sociolinguistics"[38], which focussed on the various functions speech can perform when difference levels of context are taking into account.

example. Should the listener fail to indicate understanding or attendance, the speaker may try to repeat himself, possibly re-explaining something until comprehension is signaled or even check explicitly with, "Are you paying attention?" or "You see?". On the other hand, if the user has uttered, "I see" or done some other functionally equivalent behavior, the speaker may then move fluidly on to the next idea.

Bavelas'[3][2] research in face-to-face interactions presents a very similar framework for gesture, in which gestures are categorized as 'topic' or 'interactive'. A topic gesture is one in which the semantic content of the gesture overlaps with or complements the speakers topic, whereas an interactive gesture is one which plays a role in regulating the discourse. These interactive gestures are further subdivided into citing another's contribution, seeking response, delivery of new information, and turn management. In relation to the categorization of gestures in section 2.1.2, interactive gestures are proposed to subsume the category of beats as well as include some metaphoric and deictic gestures.

In the next few sections, I will present a range of primarily interactional functions and describe ways in which various non-verbal behavior is used in human-human interactions to perform these functions.

2.2.2 Turn-taking

One of the most basic and important aspects of a conversation is the frequent exchange of speaker and listener roles. With the exception of occasional overlaps in speech, it is almost always the case that only one person at a time is speaking. This occurs effortlessly in normal conversation, yet is much more difficult to maintain in real time text conversations online, or in cb radio communication, both of which mediums have developed extra turn-taking cues to compensate for lack of the normal multimodal cues.

In face to face communication there are many ways to indicate whose turn it is, when that turn is over and who will speak next. Duncan[14] lists at least 6 behaviors which when used in conjunction with the end of phonemic clauses, tend to occur frequently at the end of a turn. Here, I will discuss four categories of cues, a) verbal,

b) gaze, c) intonation d) gesture.

At the most basic level, the speaker role is defined by who is speaking. If two people speak at once, chances are one or both of them will stop and possibly even resort to explicitly indicating who should speak next. Conversely, in the absence of other non-verbal cues, if both participants are silent for a while, either one may then speak next and take the turn. Explicitly verbal cues at the end of a turn may include the use of stereotyped expressions such as "but uh", "or something" or "you know"[14].

However, most conversation is not filled with overlapping speech and pauses during the exchange of speaking turns can be quite short. One of the ways in which people can indicate the initiation, continuation or completion of a turn is through gaze. Initiation of a turn is often accompanied by the new speaker gazing somewhere other than directly at the other person. This behavioral cue is also strongly associated the continuation of a speaker-turn[14]. Thus when yielding the turn, the speaking will almost always look at the listeners (or one of the listeners). Gaze of the listener(s) is usually focused primarily on the speaker, or in the case of an exchange of turns, the next speaker.

Intonation also plays a role in turn-taking, particularly in the anticipation of the end of a speaking turn. Duncan notes that drops in pitch and intensity, paralinguistic drawls on the final stressed syllable of a clause, and pitch tunes that indicate terminal junctions, are often displayed at the end of a turn. Pierrehumbert and Hirschberg[33] also mention a compression of pitch range which often anticipates the end of an utterance, and discuss a particular phrase tune (LH%) which conveys a forward-looking reference or "more to come".

Gesture also plays a role in turn-taking. Duncan states that when the speaker is gesturing, it "virtually eliminates the claims to the turn by the auditor". He uses the initiation of a gesture as a signal, similar to looking away from the listener, as a cue indicating the continuation or initiation of a turn.

2.2.3 Inclusion and Backchannel Feedback

Interaction between speaker and listener is not just a matter of exchanging the speaking turn. There is a lot of interaction during a speaking turn as well. The most important type of this interaction is backchannel feedback. Backchannel feedback is a term for things a listener does to indicate he is paying attention or understands what the speaker is saying. Duncan counts as backchannel behavior the asking of clarifying questions, finishing sentences, nodding, and saying "mm hmm" or equivalent one or two syllable utterances. However, in his data, nods and "mm hmms" far outweighed the other types in occurrence. Chovil[13] found that half of the facial displays she categorized in the listener were backchannel displays and consisted of brow raises, mouth corners turned down, eyes closed, or lips pressed.

Along with backchannel feedback, there are behaviors which the speaker can perform to include the listener in the dialog, without giving up the turn. Some of these are related directly to the backchannel feedback. Duncan defines within-turn signals as looking at the user or finishing a syntactical utterance, and continuation signals as looking away or having hands in gesture space. The three types of behaviors, within-turn signal, backchannel feedback, and continuation signal, are related in that if the speaker makes a within-turn cue, than the listener will be more likely to make a backchannel signal. At this point, if the speaker intends to keep the turn, he will exhibit a continuation signal.

Also mentioned as interactive behaviors which could elicit feedback are small utterances inserted into the speech such as "y'know", "am I right", or "you see". In addition these could be accompanied with gesture[3] (eg: an upturned hand slightly pointed towards the listener), intonation (eg: a pitch rise on a particular word the speaker in unsure the listener is familiar with), facial display[13] (eg: eyebrow raises) or some combination of these behaviors.

Brenner and Hulteen[7] discuss feedback in depth in a spoken language system, emphasizing the importance of positive and negative feedback. They lay out a framework with different levels of feedback for 0) not attending 1) attending 2) hearing 3)

parsing 4) interpreting 5) intending 6) acting and 7) reporting. Even though they discuss only explicitly verbal feedback, some of these levels could also be useful in determining facial expression, particularly in terms of showing the user that the system has heard an utterance, but is still processing it and forming a response.

Chovil[13] finds that in human dyadic conversation, people will frequently form a thinking or remembering facial display either in a pause or accompanying speech. These facial displays include eyebrow raising or lowering, mouth twisted to the side, or one corner of the mouth pulled back.

2.2.4 Theme/Rheme, Emphasis, and Contrast

Another area in which non-verbal behaviors play a large role is in highlighting important propositional information and its role in the utterance. One of the ways in which human convey this is by using intonation. In fact, it is noticeably more difficult to follow the monotone and unaccented utterances produced by text-to-speech systems than it is to follow a similar utterence which has been annotated with and intonational information. Pierrehumbert and Hirschberg[33] suggest that different intonational patterns are used depending on how the utterance relates to the previous discourse. For example consider the phrase: [2]

George withdrew fifty dollars.

This sentence can be accompanied with different intonational tunes depending on whether it was the response to

Who withdrew fifty dollars?

or

How much did George withdraw?

In the first case, the word "George" would be accompanied by a rise in pitch to indicate that it is the new information, whereas in the second case, "fifty" would be the accented word.

[2]This example was modified from [9]

In this thesis, I will be be following the work of Steedman, Prevost, Cassell, and Hiyakumoto[35][36][19] in the field of producing appropriate intonation for text-to-speech and concept-to-speech systems. In these systems intonation was generated automatically from the information structure. Each utterance was annotated with theme and rheme indications. Roughly speaking theme is that part of an utterance which provides a tie back to the previous discourse, while rheme contains the new bit of information. In addition, certain key words were marked as focused. Then, from these markings, associated intonational tunes were chosen to go with the different parts of the speech to convey the meaning to a listener.

Intonation is not the only modality which could relate to theme/rheme distinctions. Cassell[11] suggests that content adding gestures may also occur more frequently during rhematic portions of speech. In addition, beat gestures are another way to emphasis important words or sections of words.

Chovil found that in a study of nearly 1200 coded facial displays[3], one fourth of them served syntactic functions. Of the specific syntactic functions, emphasizers, underliners, and question markers were the most abundant. Emphasizers occured on a intonationally stress word, underliners across a whole phrase and question markers across a whole question. The first two were primarily marked with raised eyebrows. Whereas the questions were marked with raised or lowered eyebrows.

Contrast

In many cases it is important to distinguish to the listener which of a set of items to which the speaker is referring. This in actually more of a propositional function that an interactional one, but I mention it here as it is directly related to the above discussion on intonation. This task can be done verbally, as in "Please pick up the red block". However, with additional intonational stress, this could be contrasted with the green block or the red crayon without using additional words, "Please pick up the RED block" or "the red BLOCK". The above algorithms take this into account

[3]Chovil indicates that since smiles were overwhelmingly abundant in her sample, smiles without other facial changes were excluded due to economic consideration.

when determining which words are focused.

Another way of contrasting items that humans use frequently is through gesture. Deictic gestures can be used to point out concrete objects or to establish and refer to in a spacial manner entities which are not physically present[27]. An example of this would be if a speaker were discussing his opinion of two co-workers.

> Shelly <gesture to left> always writes in the margins, but <gesture to right> Polly uses extra paper. I wish <gesture to left> she wouldn't do that.

In the absence of gesture, the referent for "she" is ambiguous but would more likely be taken to be Polly. However, the gesture indicates that it is actually Shelly's habits the speaker does not like.

This particular use of deixis is also more propositional than interactive. However, both Bavelas's citing function of gesture and an example in [27] in which the speaker uses a deictic gesture at the paranarrative level[4] to ratify shared information are cases of interactive functions of deictic gesture.

2.2.5 Other Discourse Functions

There are other discourse functions which could be accomplished utilizing non-verbal behavior, though most extensive research in these functions has tended to be linguistic analysis. Areas include cohesion[18], footing[17], and discourse markers[37]. McNeill[26] discusses a type of gesture called a cohesive gesture which repeats an earlier gesture in form, tying the current utterance back to the previous theme. A shrug, whether done with the face, shoulders or hands can show a speaker's stance with respect to what they are saying, as can intonation. Beat gestures[26] and eyebrow raises[13] can be used to signal jumps among topic levels or narrative levels

[4]McNeill et al.[27] discussql deixis in the domain of storytelling, in which three narrative levels are distinguished. The example mentioned is in Figure 2 line 6.

2.2.6 Emotion and Personality

One function that people often associate with non-verbal behaviors, facial movements in particular, is the expression of emotion. Because this is so often the case, I feel that it is important to discuss emotion briefly here.

Much writing about emotion refers to six basic emotions: happiness, surprise, sadness, anger, fear, and disgust. Imagery studies have tested the extent to which various evoked responses are similar in different people[34], and the recognizability of different emotions from facial expressions[15]. Picard[32] also discusses the recognition and expression of affect in facial displays and in intonation

However, it is still unclear to what extent a computer needs to be able to understand and generate emotional expression in order to have a conversation. Chovil rated facial expressions in conversation by function and found that only about one sixth conveyed personal reaction. However, this data is hard to judge for three reasons. First, she did not include smiles unaccompanied by other facial expression. Second, personal reaction need not mean personal emotional reaction. And third, the subjects were asked to talk about a meal of food they disliked, a conflict or argument and a close-call experience, all of which might be accompanied by larger than usual number of emotional facial displays.

Emotions and emotional expression[5] may add to the depth, personality, entertainment value or even believability of a character, but these are not the goals being addressed in this thesis. The goal here is to enable the computer character the multimodal tools to hold a fluid face-to-face conversation.

Personality is another variable that people generally ask about. While it is true that personality will affect a persons' conversational expression, and for any single multimodal behavior, there are people who may rarely use that behavior, there are a wide range of behaviors that almost everyone within a culture does understand and use. If asked to hold a conversation with a stranger in a supermarket line, one would

[5]I am using emotions and emotional expression here as very distinct phenomenon. A person may show an emotional facial display such as disgust when talking about an item he (or someone else) dislikes, while not actually feeling the emotion of disgust at the time.

be far more likely to worry about what to say, not how they use their hands, face and voice to say it. Yet computers today do not yet even have the basics of how to use these channels, much less how to modify those behaviors depending on personality and emotion..

Chapter 3

Related Work

Recently there has been much interest in the creation of interactive animated computer characters. Although many of these projects are addressing many similar issues and facing many of the same problems, each of them seems to have a unique approach and individual spin on how to proceed. In this section I will analyze some projects which seem to be related to our goals of creating a conversational character. In particular, I will focus on major differences in architectural approach and to the methods used to determine the behavior set of the characters. I will begin with a discussion of an architecture design presented by Blumberg and Galyean[4]. Next I will examine an attempt to integrate the speech system, Glinda[20][21], into a behavior-based character system developed by researchers in the Oz group at CMU[23]. Following that I talk about Improv[31] and lastly on to the Microsoft Persona project[1].

3.1 Directable Characters

There has been a lot of work recently in creating characters which have some life of their own and can choose among behaviors and interact with users in a real-time fashion. All of the projects mentioned in this chapter, as well as the Rea project described in chapter 6 share this characteristic. However, after that there tends to be a wide range of goals and approaches. One of the key differences between the projects mentioned below and the Rea project, is that a key goal in the works below

is that the characters be entertaining. This goal leads naturally to an approach based in classical animation or theater, and then working towards giving the characters autonomy and "directability". In contrast to this approach, my research begins with the goal not of making the character entertaining, but rather someone who a user can talk to face-to-face. The focus is on the conversational interaction between human and character, rather than the internal life and believability of the character.

3.1.1 Autonomous Creatures

One group which is working on behavior-based autonomous creatures is the Synthetic Characters group at the MIT Media Lab. Their goals differ from ours significantly in that they are working with characters that are not necessarily human based, and the primary mode of action by the character is not speech. In addition, they are concerned with creating interesting autonomous creatures with a wide range of time-varying behaviors dependent on various internal goals, perception, and user direction at various levels[4]. Towards this end, they have developed architectures that allow for the creation of behaviors which then compete at run-time to determine the movements of the creature.

What is most interesting about this work in relation to Rea is the similarities in architecture with respect to how they produce low level actions. Both systems have a module responsible for the high level decision of what the agent will do/say next depending on perception and internal state[1]. Then this is passed on to another module[2] which translates the high level directives into appropriate low-level actions appropriate to be further manipulated into actual animation and in the case of Rea, speech.

The mapping of high level to low level actions means slightly different things in the two systems. An example in the Autonomous Creatures architecture would be the mapping of the goal directed FORWARD command into a creature specific WALK implementation. Whereas an example in Rea might be turning a high level

[1]This module is termed Behavior in the Autonomous creatures, and the Reaction Module in Rea
[2]Controller or Generation Module

TAKE-TURN into LOOK-AWAY and SHOW-THINKING-FACE. Yet the similarity is enough to see that another aspect of the creatures architecture may also be integrated into the Rea architecture. This feature is the ability of the high level module to make suggestions to the translation module about how to perform certain translations. An example in [4] is if a portion of the higher module determines the dog is sad, it may suggest that if the dog should walk forward in the near future, it should walk slowly, but not actually command the dog to walk forward. It is possible that some similar mechanism may be possible for modifying the surface realization of Rea's behaviors as well, without actually modifying the underlying need to perform discourse functions.

3.1.2 Believable Agents

Another group working on the creation of engaging characters is the Oz group at Carnegie Mellon University. They also have developed a behavior-based architecture for constructing non-linguistic believable agents called Hap[23]. Characters in Hap can interact in a virtual environment and may be either completely autonomous or directed by a human user who can change variables such as emotional state or what high level action to attempt next. Recently, they have been including the use of natural language by incorporating the Glinda text generation system[20][21] into Hap, giving the characters the ability to generate text along with other behaviors[23].

They report some very compelling results in their work, namely, the integration of linguistic and non-linguistic behaviors and the generation of language in a real-time incremental fashion which can be be interrupted or modified depending on intervening circumstances, and the ability to have emotion and personality to affect the generation. Particularly this second feature, the ability to modify speech and react to the environment from moment to moment even while in the middle of a speaking, is in line with our goals as well.

However, there are some major difference. The first is that they are interested primarily in learning from "artistically inspired abstraction retaining only those aspects of an agent that are essential to express personality."[23]. As a contrast to this I pro-

pose a method of functionally inspired abstraction, retaining primarily those behavior essential to a smooth face-to-face conversation between human and character. The second major difference is in the physical manifestation of the characters. Many of the characters are not particularly human-like. For example, woggles are small round creatures with eyes which speak via text bubbles; a human user controls a woggle by choosing various actions and emotional states via a mouse[3]. Thus, in terms of integrating perception into language and behavior generation, woggles need only understand their virtual environment and other woggles, and in terms of generating behaviors tied to speech, they operate with very different constraints. In particular they have no arms for gesture or voice for intonation, and the timing of turn-taking behaviors is significantly altered by having text rather than speech as the linguistic medium.

3.1.3 Improv

Another system which approaches the creation of interactive characters in terms of believability is Improv[31]. Improv is a system designed to allow authors of characters to define actions and behavior scripts. Characters in the Improv system follow layered scripts and interact with each other in real time.

While the characters in Improv tend to be more human-like in form than woggles, there are still some major differences between this approach and ours. To begin with, characters primarily interact with each other [4], giving off a life-like appearance from a distance but not in sufficiently realistic manner to allow for a conversation between a human user and a character. An example of this is that one character

[3]It looks as if a user will eventually be able to type text for a woggle as well, but at the time of the [23] paper this was limited to keyword matching of yes and no.

[4][31] does very briefly mention interactive uses of the Improv equipment. However, in one of these the user is represented in the virtual world as a bat, not a humanoid. The other seemed much more relevant, although the interaction is not really a conversation. The interaction is with a character, Sam, who knows how to play "Simon Says". Success with this experience was attributed to three factors (1) participants talked directly with Sam (2) Participants know the character was not being puppeteered, and (3) Sam's motions were relatively lifelike and never repeated themselves precisely. The first two of these are a given in our approach, and the third suggests that in our modeling from life we allow for some variability.

might follow a "tell a joke" script, invoking other scripts to do "joke gestures" and cue other characters to call "listen to joke" scripts. From a distance this looks as if the characters are interacting and reacting to each other, but it is at a relatively course level. Information is not actually encoded in the behaviors of an agent and interpreted by other agents. They are just instructed to run appropriate react scripts.

All of these scripts eventually terminate in actual actions, depending on personality features of the characters and herein lies second difference between our work and theirs. Since the system is aimed at giving creative control to an author, all of these actions and scripts must be carefully hand crafted until they look realistic enough. Actions are created with many variables which fire probabilistically, and "the influence of the author lies in carefully tuning ... such probabilities." It is possible that with the addition of speech capabilities Improv might be able to serve as a background to research in conversational behaviors, but as such it is really designed as a platform for authors, and thus does not address the issues specific to a particular domain such as conversation. Also it is unclear just how difficult it would be to include linguistic abilities and reactivity at a more detailed level.

I believe that the strongest lessons to take out of this system and apply to Rea are the ability to allow for variability in a character's actions and the ability to allow a non-programmer to modify the behavior sets. In our case, since the focus is not on creating an environment for many characters with different personalities, it is less important that everything be authorable. Yet even so, the ability to change the way the character performs various functions will be very useful, particularly when it comes to testing the various behaviors.

3.2 Characters as Interfaces

Rather than creating characters from a dramatic point of view, some researchers are working on anthropomorphic characters as social user interfaces. In this section, I will discuss two projects, one which uses lifelike behaviors in computer generated humanoids as an interface between humans, and one which uses a character as an

interface between a human and a CD player.

3.2.1 BodyChat

BodyChat is a online graphical chat system. Each user is represented by an avatar, a cute character with upper torso, head, and arms, in a shared virtual 3D space. However, rather than using the avatars as nothing more than creative place markers for people, as in current chat systems, BodyChat animates them with realistic conversational phenomenon such as turn-taking gazing and greeting behavior based on human-human interactions.

Like the directed characters discussed above, each character in BodyChat is a joint persona, made up of a user and an avatar. However, rather than directing at a high level, the human user is the primary source of all text and movement around the space, as well as understanding the behaviors of other avatars. What the avatars provide is a set of automatic non-verbal interactional behaviors which accompany the selection of conversational partners, utterances entered while conversing, and the intention to break away. Thus when looking at the other avatars a user sees a combination of verbal and non-verbal signals just as in real life. The strength of this work comes in comparing it to graphical chat systems which do not provide life-like body movements to accompany conversation or ask users to explicitly select the low level behaviors themselves. The problem with these traditional systems is that the avatars do not provide the often involuntary and spontaneous non-verbal cues we are accustomed to seeing and interpreting from humanoid forms. An avatar which periodically checks its watch to show it is "alive" may instead be unwittingly conveying the information that the user of the avatar is uninterested in the conversation he is having. BodyChat seems to provide aliveness and cues which can be appropriately interpreted, even if not noticed as entertaining behavior. In addition it does this with almost no reference to emotion[5]

[5]BodyChat has one switch which could be argued to fall under the domain of emotion, or mood: Available or Not Available(to chat). However, this variable is not used once avatars are engaged in conversation.

Table 3.1: Behaviors in BodyChat

Conversational Phenomena	Communicative Behavior
Approach and Initiation	
Reacting	ShortGlance
ShowWillingnessToChat	SustainedGlance, Smile
DistanceSalutation	Looking, HeadToss/Nod, RaiseEyebrows, Wave, Smile
CloseSalutation	Looking, HeadNod, Embrace or OpenPalms, Smile
While chatting	
Planning	GlanceAway, LowerEyebrows
Emphasize	Looking, HeadNod, RaiseEyebrows
RequestFeedback	Looking, RaiseEyebrows
GiveFeedback	Looking, HeadNod
AccompanyWord	Various
GiveFloor	Looking, RaiseEyebrows
BreakAway	GlanceAround
When Leaving	
Farewell	Looking, HeadNod, Wave

Although the parameters of the system are very different than in a system designed to be an interface between a human and computer, there is one aspect of BodyChat that is particularly useful. This is the clear distinction between conversational functions and behavior used to serve those functions. Thus when behaviors are sent to the animation system, that portion of the architecture need not know, for instance, why or when one would want to make a beat gesture. However, it does know how to move the body to make that beat gesture. Table 3.1 is reprinted from Hannes Vilhjalmsson's Masters thesis[41]. This distinction has served in a large part as a basis for much of the work in this thesis.

3.2.2 Microsoft Persona

The last related work I would like to mention is the Persona project at Microsoft Research[1]. The prototype system described is an animated parrot, Peedy, who operated a CD Player when a human make request for music. In many ways, the Persona project is much more similar to Rea than the above mentioned character

projects, both in overall goal and in stages reached. Both projects are concerned primarily with the interaction between human user and single computer character and have as a goal mixed initiative conversation using spoken natural language. Also, both are attempts to collect many of the traditionally isolated research pieces together into a working system, even though many of the pieces are still areas of active research (such as natural language recognition and understanding).

However, in the process of collecting together the different pieces of technology, the Persona project team and the Gesture and Narrative Language Group have focused on different aspects. For instance, since we are interested in multi-modal behaviors, it is important for us to use output tools with which we can generate non-verbal behaviors. Thus, we are using a 3D humanoid figure in which all major head, face, arm, hand, and finger movements can be generated from code as well as using a text-to-speech system capable of using text with intonation annotations. The Persona team, however, chose to begin by pre-recording all of Peedy's utterances, as well as crafting many short animation sequences by hand, and then working on ways of dynamically scheduling and playing those animations. They indicate that they found this very limiting, and next want to move towards a system with more generation capabilities, both with the speech and the animation. On the other hand, Peedy's creators have focused more on input natural language processing and dialogue management than we have to date, though one of the goals of the Rea project is to include this important aspect. As we incorporate this into Rea, it may be useful to look at what problems they ran into here, though I think that our use of non-verbal conversational cues may help greatly in co-ordinating mixed initiative.

Chapter 4

Happy or Envelope - A Wizard of Oz Experiment

4.1 Description of Study

When attempting to build a computer character which interacts with humans, one important step is to actually test how people interact with computer characters. Do they find some engaging, others helpful, and still others persuasive? Should we try to give them emotions, personalities, or communincative body language? It is difficult to test all of the possibly variables, particularly when they are implemented in different systems with different goals and means of interaction.

Although we cannot answer all these questions, we can design experiments which begin to address these issues. This chapter describes one study in which we attempt to test the effects of smiling versus simple interactional communicative behaviors on the dialogue between a computer character and a human, and on the human's impressions of the character.

The study was a Wizard of Oz experiment in which rather being autonomous, many of the behaviors the character was able to produce were initiated by me. The face of the character was that of Gandalf, the interactive character created by Kris Thórisson[39].

4.1.1 Task and Protocol

The subjects' task was to complete a Desert Survival Problem with the help of the character. For simplicity let us assume one particular subject. First the subject was seated in an area from which the computer set up was not visible and asked to rank 12 items in order of importance with the supposition that she had crash landed in the Sonora Desert. Once the subject had completed this portion, she was then informed that she would next discuss each item with a computer character who may or may not have access to the correct answers, but that in either case the character would also provide a list of rankings for the twelve items.

The subject was next shown into an area in which Gandalf's face appeared on a projection screen in front of her. To the right of Gandalf, on a computer screen, there was a list of rankings, which for every item showed the how the subject ranked it and how Gandalf ranked it. Unbeknownst to her, Gandalf's rankings were just a fixed permutation of her own. In addition, one of the experimenters was sitting about 6 feet to the subject's left behind a computer, such that the subject was in (partial) sight and (partial) hearing range. This experimenter controlled the timing of the character's behaviors. The subject was not informed that this person was involved in the experiment, and the environment was a large room with many people working on computers throughout, such that this person was not conspicuous.

The subject was then instructed to make a statement for each item on the list explaining why she placed it where she did, to which Gandalf would respond with a similar statement about where the item should be ranked. After each pair of statements, the subject would be responsible for continuing on to the next item. For every item, Gandalf was pre-programmed to respond with one of two possible answers, depending on whether his ranking was higher or lower. The subject was lead through a practice example and left to interact with Gandalf. This interaction was recorded on videotape.

After going through all twelve items, the subject was taken back into the area away from the computers and given a chance to change her rankings on a piece of paper

Table 4.1: Controlled Behaviors in the Wizard of Oz Experiment

Condition	**Behaviors (During Subject's Turn)**	**Behaviors (During Gandalf's Turn)**
1 Alive	None	Speak
2 Happy	Smile Brows Neutral Mouth Neutral	Smile Speak Brows Neutral Mouth Neutral
3 Envelope	Nods	Turn To Ranking Screen Speak Turn To Subject Beat Gesture Glance Away and Back
4 Happy+Envelope	Smile Nods Brows Neutral Mouth Neutral	Turn To Ranking Screen Speak Smile Turn To Subject Beat Gesture Brows Neutral Mouth Neutral

which showed both her initial ranking and Gandalf's rankings. After completing this task, she was then asked to fill out a questionnaire about her impressions of the interaction and the character. A copy of the questional appears in Appendix C

4.1.2 Behaviors

The two aspects of the character's non-verbal behavior that we were most interested in testing were turn-taking cues, and emotional response. Thus we had four conditions, each with 6 subjects. See table 4.1 for an outline of the behavior in each condition.

In the first condition, Gandalf tapped his fingers, blinked, and opened his mouth when speaking, but otherwise performed no additional behaviors. Both the tapping, and blinking were automatic functions that occured periodically in all conditions. The only thing the experimenter controlled was when Gandalf would begin to speak.

In the second condition, Gandalf was manipulated to smile and raise his brows during the second portion of each of his utterances and a short time after the subject

began her turn, in addition to the above behavior. The transition from smiling to not smiling was done in two steps at the end of bot the subject's turn and Gandalf's turn. First the brows were lowered and then the corners of the mouth were lowered.

In the third condition, Gandalf began each of his turns by turning his head toward the rankings screen. After his first utterance he again faces the subject and continued speaking. During the following speech, Gandalf does a simple up and down gesture with his hand and glances away from the subject. At the end of his turn he glances back at the subject. During the subject's turn, Gandalf nods during pauses and/or in response to the completion of an utterance. We call this behavior Envelope behavior. This condition also included the behavior from condition 1.

The final condition was a composite of both conditions two and three, save the gazing away and toward the user at the end of Gandalf's turn. This was omitted because it was judged to that with both the smiling behavior and the other envelope behaviors, the end of the Gandalf's turn was sufficiently marked, and that the extra movement did not convey the appropriate cue.

Our hypothesis was that subject's impressions of Gandalf and of the interaction would depend on which non-verbal behaviors he exhibited. In addition we were interested to see which condition persuaded subjects to change their answers the most.

4.2 Why a Study

Although studies such as this are limited with respect to the number of variables they can effectively test, they can give us a lot of information that would be otherwise difficult to obtain. This is because we have the advantage of being able to control more of the variables than we might in human-human interaction, without having to solve all of the technical problems of creating a fully autonomous character. In this case, the wizard acted as a reliable input device and decision module. With current technology it is still very difficult to mechanically read all the subtle clues a human may perform, particularly without using distracting devices worn by or placed right

in front of the subject.

4.2.1 Statistical Level

I see three different levels of learning that an experiment like this helps to advance. The first of these levels is a formal one. We design the experiment with particular conditions to test. Using a semi-animated character allows us to fix many of the variables which occur in the non-verbal behavior of humans, without having to built all of the technology necessary to actually create a sufficiently animated humanoid. Once a sufficient number of questionaires are obtained for each condition, we can do statistical analysis on the data. Thus, we may find that one of our behavior conditions leads to strong impression along some axis. For example, in this experiment we are beginning to see trends in the data that suggest that for many related impressions(Helpful, Relevent Informtion, Trustworthy, Informative, Knowledgable, Insightful, Reliable), having either smiling or envelope feedback helps, but that both together do not.

4.2.2 Observation Level

On a slightly less rigorous level, we can learn a lot about how people interact with computer characters by video taping and examining the behavior of the subjects. Some of this information is not easily quantified, or is unexpected behavior that occurs with only a few subjects and thus not statistically relevent. However, this information is useful for designers of such characters. One way that it is useful is in pointing out behaviors or reactions that were unanticipated by the designers. Thus, in future designs, both of experiments and of animated humanoids, the designer can be aware and deal with a wider and more realistic behavior set.

One example of this a situation in which the subject and Gandalf were unable to recover gracefully from a misunderstanding. The problem occured because Gandalf could not say more than one thing about each item, and though the subjects had been warned about this, the subject in the following example expected Gandalf to respond again:

Subject: *looking back and forth between the rankings and Gandalf* Ok, well, my thoughts on the flashlight were that, uh, I'm going to want to be travelling almost exclusively at night and I don't know um whether there's going to be very much light and I think it will be important to be able to read the compass uh at night 'cuse I can't read stars, um and, I notice you made it at an eight um so its pretty close. Why do you.. what are your thoughts thought on the flashlight? *turns toward Gandalf*

Gandalf: The flashlight should be rated lower, the batteries will not last long and the moon will provide enough light to move around at night.

Subject: *turns back to ranking screen and nods* OK so if you think the moon's going to be out then I will ah.. I would agree with you and es...*gestures to Gandalf* that's a good point, I didn't think about the batteries so, um, then I'll probably, probably move that down to even maybe, *glance at Gandalf and back* I'd say less than eight um, *Gandalf smiles* if uh, because if, if there's going to be moonlight anyway there's really no need for it. *looks at Gandalf, pause* Do you agree? *pause* Gandalf? *pause, looks at rankings, glancing at Gandalf* ok, well I guess thats all I have to say on that. Do you want to go on to the Jackknife? *looks at Gandalf, pause* Hello. *pause* Hmmm. *Gandalf ceases smiling, short pause* You have no, ah, no more thoughts? *pause, turns back to ranking screen, pause, turns to Gandalf, Gandalf smiles,* Ok, so, the Jackknife, Gandalf? *short pause, turns back to ranking screen* Alright well maybe I'll...we both rated them really closely... *continues on to talking about the jackknife*

The subject here was in the "happy" condition, thus Gandalf was constrained to just smiling, unable to provide even a nod for feedback. As it turned out, I ended up including an extra transition from smiling to neutral to smiling again, so that the subject could have enough feedback to move on. This sort of break down in the interaction greatly affects the impression formed by the subject. The subject above ranked the interaction with Gandalf a number 1 out of 10 for both of the qualities,

Rational and Competent, while others in the same condition ranged from 5-9[1].

In contrast, another subject who also talked back to gandalf but was in the Envelope condition had this exchange with Gandalf:

> **Subject:** *looking at ranking screen* What about the airmap? *pause* It should probably be a little bit lower... I would ... I would agree with you then. *turns to look at Gandalf, Gandalf turns toward ranking screen, pause*
>
> **Gandalf:** The airmap should be ranked lower. *Gandalf looks back to subject* Determining your location in a desert environment will be virtually impossible, with or without the map.
>
> **Subject:** *looking mostly at ranking screen* But if you have a compass, you will have a better idea, so it should still be a little bit higher then you have. *looks at Gandalf, Gandalf nods, pause* nope? ok. um *looks back to ranking screen and goes on to next item*

In this exchange, two things stand out. First, there is a pause after the subject has finished her turn, but since Gandalf has turned to "look" at the ranking screen, the subject waits patiantly, watching Gandalf's face. Secondly, when she looks at Gandalf after saying "a bit higher than you have", he nods. Thus, after a little while when he does not take the turn, she says "nope? ok". This "nope" does not seem to be a response to Gandalf on a propositional level. It seems unlikely that she interpreted the nod and a "no". Rather she seems to be saying "you don't want the turn? ok I will go on". This contrasts strongly with above, where when Gandalf doesn't give feedback or take the turn even after the subject explicitly asks for feedback. He seems to begin to wonder if Gandalf is even hearing his words, much less interpreting them.

This sort of evidence seems to show how including discourse knowledge about how to take turns and give feedback as the listener can be very helpful in error recovery. However, since most of the subjects did follow instructions and move directly on to the

[1]In fact, save one other subject who did a similar thing, all other subjects ranked these two aspects from 4-9.

next item, this particular result remains on the observational level. Yet, even results at this level will be useful in designing characters and planning other studies. In this case it would be interesting to think about designing a conversational breakdown into the conversation, and testing various ways of getting back on track.

4.2.3 Personal Level

The third level at which I believe this experiment was a good learning experience is more personal. Speaking as the experimenter who actually controlled the Gandalf face, I think that this experiment taught me things about turn-taking behavior which classes and papers and, of course, years of conversations, had not. For example during the design of the experiment, although I knew what facial features I had to work with and had a general idea of which communicative functions are served by which behaviors, the process of determining which behaviors to include and in what order included a large portion of trial and error. Behavior observed in human-human interactions did not always map in a realistic way to Gandalf's face. For example, although a human might gaze around while speaking and then focus on the listener at the end of his turn to indicate he was done speaking, it was very difficult to achieve that effect with the animated face. One of the other experimenters commented that it made Gandalf look shifty. Having control of the features in real time was a useful way of exploring what did look natural and might have a chance of being understood to the subject as communicative behavior.

Even during the running of the experiment, just having the explicit control of one animated face, and having to keep my own face looking as if I was not paying attention, made me much more aware of those communicative behaviors which I usually take for granted. During the first few practice subjects, I often found myself smiling, laughing or nodding in response to the subject rather than transfering that behavior to the Gandalf face. Later, after having fixed specific conditions and specific times within a condition to make Gandalf exhibit those behaviors, there were often times when I wished I could change them to be more appropriately reactive. While these experiences don't yield a prescriptive guide for generating communicative behaviors,

I believe that having the experience of having to generate limited behaviors on a cartoon face under restrictive circumstances will help me to program a life-like communicative agent who will, for some year to come, be limited in the types of input it can understand and output it can generate.

4.3 What was learned

4.3.1 Emotional Behaviors

I believe there are two aspects to the emotional features of Gandalf that seemed to me unrealistic. First, he has basically two states; happy, in which his eyebrows and the corners of his mouth are raised, and neutral, in which both are down. The transition from neutral to happy looks very much like a strong smile, and the reverse looks like a frown. Such rapid transitions had the effect that it seemed like something had made him smile or frown, usually whatever was directly preceding the change in facial expression. In order to mitigate the frown effect, we had Gandalf go from happy to neutral in two stages, first bringing his eyebrows down, then lowering the corners of his mouth. If the goal is to smile a lot and not to frown, I think it would be even better to have a even more gradual transition there.

The second feature which I noticed more and more as the experiment continued was the timing. When the subject was the speaker, it was often the case that the subject was looking at the ranking screen when Gandalf smiled, but was looking at Gandalf at the end of her turn when he went back to the neutral position. The timing was such that the subject almost always saw the frown transition just at the end of her turn, right where it could be viewed as a comment on what he thought about what she had to say. Perhaps this explains why subjects tended to find the happy condition Gandalf less collaborative and more adversarial than in the other conditions.

I would suggest that rather than trying to tie smiling in with the turn-taking structure as I did in this experiement, it be used by the listener only occasionally

and in response to possibly content based cues. Once a smile(or frown) is produced I think it should fade gradually, thus allowing the character to smile (or frown) again.

The second part of the timing that seemed somewhat off was having Gandalf smile for the second portion of his turn. In general, when Gandalf was the speaker, it seemed natural to me to have his smile timed with either something funny or start at the beginning of an utterance and continue for a while as a comment on how he feels about what he is saying. Seeing as he was not really pre-programed to say anything funny, and it was difficult to time his smile exactly with the beginning of his second utterance, his smile worked best when he smiled just before starting his turn.

After running this experiment, I believe that smiling or frowning, or puzzlement or other such emotional facial expressions, are best considered as meta-level commentary on what is being said, coloring it or responding to it, but in either case, not particularly tied to the turn-taking structure. In addition, it seems that it is important that these facial expressions are seen (or heard).

4.3.2 Envelope Behaviors

Nodding

The most useful of the envelope behaviors, I felt, was the ability to give backchannel feedback via nodding, thus showing that that the subject was heard. I used it in a mixture of cases (1) when the subject had completed a propositional phrase (not necessarily with a following pause), (2) when the subject paused for a long time, and (3) when the subject looked at Gandalf. It worked well in both those cases in which the subject was giving up the turn and in those where the subject continued to speak. In the case that the subject did not pause and was already continuing to speak, it was good because it seemed to indicate reception of the information without distracting the subject from continuing to speech. In the cases of prolonged pauses in the middle of and utterance, it seemed to mean "Go on, I'm listening, I will not take the turn." Whereas, at the end of a turn, it served as acknowledgment that could easily be followed by taking the turn if the subject did not make a cue to keep the

turn.

Head Turning

In all conditions, the subjects spent much of their time while speaking facing the screen with the rankings. However, almost all of them turned to face Gandalf when done with their turn, even subjects who stared off into space or looked at the audio speakers half a second later(as some people in the null condition did). This suggests that, as in human-human conversations, if the speaker has paused, but has not turned to face his listener, it is probably the case that he has not relinquished the turn.

Having Gandalf turn his head to look at the task screen therefore seemed to work well as a strong cue for taking the turn, since it was behavior they themselves did, and they are often looking at him when it occurs. In at least one case where I accidentally misjudged the subject's turn to be over and turned Gandalf's head to began his turn, the subject obligingly stopped speaking in midsentence.

This behavior also seemed to have one other interesting effect as was pointed out by some of the subjects comments in the questionnaire. One subject in the happy condition said, "*There seemed to be an obvious lag between the completion of my statements and the beginning of his-somewhat akin to an extraorinary "awkward silence" b/t two humans*". Whereas two subjects who experienced Gandalf with envelope feed back said he was *quick to respond* and *purely reative with no thinking process.* I suspect that the act of turning his head toward the ranking screen at the begining of his turn satisfied the need for a interactional/reactive cue to indicate that he had taken the turn. Thus, subjects would then give him more time to actually begin to talk before becoming impatient. Perhaps they even assumed that since he was "looking" at the ranking screen, he should also take time to analyze the screen, which Gandalf did not do. (He went directly into his speech after turning.)

Gesture

The gesture that Gandalf performed was a relatively simple movement. His hand changed from being palm face down near the bottom of the screen, to palm face out,

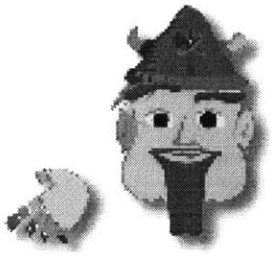

Figure 4-1: Gandalf's Beat Gesture

little fingers slightly curled in. See figure 4-1. The gesture, along with the nodding, was assigned a less specific time to occur than other behaviors tied to turn-taking. The only requirement was that it occur during the content portion of Gandalf's speech. Thus I was amazed when I looked at how much its meaning changed depending only on its duration and when it occured with relation to the speech. When it occured after an utterance, it looked like a shrug. If accompanying the words "you" it seemed like a deictic gesture. If accompanying or just before a pitch accent, it seemed to be marking important information (namely that information with the pitch accent). If accompanying or just before a new statement beginning with "Also", it functioned as a discourse marker. However, it did not look realistic if it was unnaturally short or in same phrase as a pitch accent but not accompanying it.

This suggests that as we implement gesture for a conversational agent, we should pay particular attention to the co-ordination of the words and the gesture, particularly in the case of interactional gestures where the form may not add significantly to the meaning, but the timing could change the meaning from *emphasis* to *shrug.*

4.3.3 General Lessons

Both the Happy condition and the Envelope condition seemed to lead to signifigantly better impressions that the null condition, with the envelope yeilding higer results for friendly, collaborative, and cooperative. However one message in the responses to the questionnaire, was that the condition with both Happy and Envelope behaviors was too mechanical. One subject found Gandalf's behaviors *too automated to be very realistic or believble.* Another said he was *scripted*, a third, *humorless.* I suspect

Table 4.2: Wizard of Oz Lessons for Conversational Agents

- Smiling is not the same as happy *or "What goes up, must come down."*
- Feedback is looked for, and can be useful in recovery situations *or "Gandalf? You have no more thoughts?"*
- Turn-taking and head turning are stongly related *or "I'm not looking, so you can't talk."*
- Gesture timing is important *or "Beats for all purposes."*
- Too many repititive behaviors look mechanical *or "Variability is a good thing."*

part of this came from the repetition of behaviors that was build into the structure of the interaction. It is not often that we can count on knowing the exact ordering of turns and know that the person we are talking with will only say three or four phrases about each item. But even so, I believe that this argues for having a variety of behaviors which can be used for the same function, as well as potentially having some mechanism for insuring that the agents is not overusing some behaviors.

Overall, even though this experiment set out to test the differences in subjects' responses to a character's emotional or envelope behaviors, I learned many other lessons for agent building as well. Table 4.2 summerizes these lessons.

Chapter 5

Conversational Character Projects

In this chapter I discuss two previous group projects in which conversational characters were built. Both of these draw from research in real life multi-modal interaction. However, they have a different domain and focus, thus changing what types of non-verbal behaviors are appropriate and possible to generate.

The first project is Animated Conversation[8][9], in which two characters, Gilbert and George, participated in a task-oriented discussion about how to withdraw $50. The result was an animation which included appropriate speech, intonation, gesture, and gaze generated from an underlying goal planner. The second project in conversational characters is Gandalf[39]. He utilized multimodal turn-taking cues in interactions with humans about the solar system.

As I discuss these projects in the next few sections, I will give a general overview of the projects and address the overall goal of creating an animated interactive conversational humanoid, with particular attention to the actual behaviors implemented in these projects.

5.1 Animated Conversation

Animated Conversation was a project completed by Cassell et al. [8][9] at the University of Pennsylvania in 1994. One of its primary goals was to derive both speech and appropriate multimodal behavior from one underlying representation. The result

was a animated conversation with two characters: George, a bank teller; and Gilbert a bank customer. During their discussion, they worked towards achieving the goal of withdrawing fifty dollars. A domain planner was run to determine the task step, which was then processed by a discourse planner to allow the characters to generate communicative actions.

5.1.1 Important Features

The great strength of the Animated Conversation project was the ability to generate speech and accompanying intonation and gestures from one underlying representation, called information structure. Information structure included not only the discourse entities for each utterance but also additional information such as whether an entity had been previously mentioned or was new information to the hearer. Using this information, timing for gestures and appropriate intonation could be generated to yield a realistic multimodal animation.

One of the features about Animated Conversation that differentiates it from our current work in the Rea project is that it was a dialogue between two computer agents rather than one between a human and a computer. This means that Gilbert and George did not have to deal with the input problems of sensing and understanding; they could communicate by passing well-formed information structures. Nor did they need to understand the functions of interactive cues. They needed only to produce them to look realistic. So even though they each had their own (differing) representations about what beliefs they held, they were at least certain to be able to understand how information was coded.

Another difference between Animated Conversation and Rea is that Animated Conversation did not run in real time. In fact, it took nearly 17 hours to render. Real time production is a serious constraint, not just because it allows the computer less processing time, but also because some data must be evaluated before all data is received. Take for example a system in which a speech recognizer is used to get input for a character. Speech recognizers, in general, do a poor job at recognizing individual words in a stream of speech. However, when programmed with a grammar, they are

better able to recognize whole utterances. Thus the speech recognizer will not provide information on what was said until after the utterance is complete. However, in order to create a realistic human-like character, action must be taken before the speech is complete, even if it is just a change in posture or facial expression to show that the character is paying attention.

5.1.2 Generation of Multi-modal Behavior

The behaviors included in Animated Conversation include intonation, manual gesture, gaze, and facial expression.

Intonation was generated automatically from the information structure. Each utterance was annotated with theme and rheme indications, roughly telling whether that information was new(rheme) or given(theme) in the discourse. In addition, certain key words were marked as focused, including words which were important to indicate contrastive stress. Then, from these markings, associated intonational tunes were chosen to go with the different parts of the speech to convey the meaning to a listener.

Gesture was also hypothesized to have a connection with information structure. Rhematic verb phrases and entities new to the hearer were accompanied by a gesture. Which type of gesture was used depended on whether the information could be represented spatially as an iconic, metaphoric, or deictic. Beats accompanied new information (either to the hearer or the discourse) which could not be represented spatially. These rules about when to include a gesture plus timing rules for aligning the stroke of the motion with the pitch peak in the accompanying utterance, yielded quite realistic timing for the gesture. The actual shape of the semantic gestures were looked up in a dictionary of gestures compiled manually by observing what types of gestures occured when humans were asked to speak about banking.

Facial expression was broken into three functional categories; syntactic, semantic, and dialogic. The syntactic category included facial expressions whose placement could be tied to the occurrence of generic features in the speech. Examples used in Animated Conversation include raising the eyebrows or nodding when accenting

a syllable, and blinking during the pause at the end of and utterance. Semantic functions were those which emphasized words. Dialogic function consisted of those behaviors tied to turn-taking and organizing the structure of dialog, such as gaze.

Gaze was determined probabilistically. A character was more likely to look at the other if he was taking a short turn, accenting a word, at the end of a speaking turn, asking a question, or receiving feedback during within-turn pauses. Conversely, he was more likely to glance away if in the planning phase at the beginning of a turn, answering a question, or giving a continuation signal after a within-turn pause.

Nods were used to indicate yielding the turn, while replying affirmatively, when accenting a word, and to give feedback when listening. In my opinion this yielded far more head nodding than seemed natural.

5.2 Gandalf

Gandalf is a communicative humanoid designed by Kris Thórisson[39] to interact with a human interlocutor about the Solar System. Appearing on his own monitor, he had the ability to face either the user or a larger projection screen upon which a model of the solar system was visible. The user could interact with Gandalf using a limited vocabulary and multimodal cues to ask questions or direct Gandalf to move the solar system model.

The architecture behind Gandalf, Ymir, utilizes blackboards and decision modules which reside in three different reactivity levels; the Reactive Layer, the Process Control Layer and the Content Layer. Each decision module monitors input on the blackboards, and if properly triggered sends a suggested behavior to an Action Scheduler, which is in charge of determining which suggested behaviors actually get performed. A decision module belongs to the Reactive Layer if its actions can be quickly performed and are in some sense direct low level responses to input data. For example, a module which suggests looking in the direction indicated by a pointing gesture would be part of the Reactive Layer. The Content Layer contains those decision modules which utilize specialized database knowledge and can take much longer

to process. Speech processing occurs in this level. Lastly, the Process Control Layer contains modules which can use the state of other modules as input in addition to behaviors which take longer than the approximately one second limit for the Reactive Layer. For example, generating filler speech such as "Um, let's see", while waiting for the content layer to produce content speech.

5.2.1 Important Features

Gandalf is the project most closely related to the goal proposed in the introduction in exterior appearances and approach. The project's highest priority is to interact with a human in real time, utilizing multimodal cues on the input and the output. However, while very good at utilizing interactional turn-taking cues, Gandalf failed to include sufficient speech understanding and generation to be truly useful as a platform for research on the relation between intonation and propositional gesture with speech. Although theoretically designed to be expandable, the Ymir architecture proved to be difficult to built on, because of unpredictable interactions among modules. Someone wishing to modify one section would need to know the rest of the system as well in order to be able to anticipate negative repercussions.

Another drawback of Gandalf was that the initiative was entirely on the human user's side. The human would utter one of a fixed set of sentences and then Gandalf would respond. While this may be sufficient in a system in which the character need only answer questions and perform actions at the human's request, it is a very different interaction than a conversation in which both parties provide information and learn from the other's speech.

Two other features which are less important to the theory of behavior generation but notable in the appearance of the system (see Figure 5-1 are the system of sensors used for input data collection and the output animated represention of the character. When speaking with Gandalf, a user was required to don a body suit full of position sensors. This included an uncomfortable head piece with a reflective panel and camera for eye-tracking. While this provided more precise position and gaze information, it cost significant time in pre-interaction calibration and greatly decreased

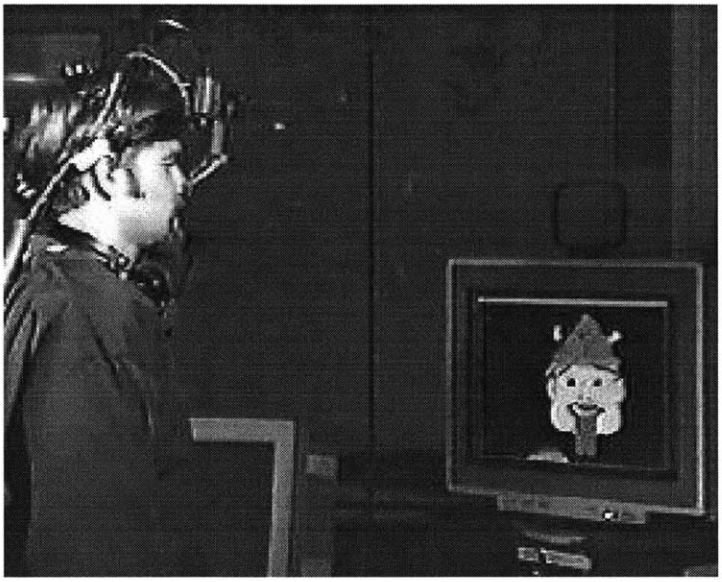

Figure 5-1: Gandalf's Hardware

the approachability of the character, which makes it much less feasible as an actual computer-human interface.

Gandalf's body consisted entirely of a 2D face and hand on a separate monitor from the solar system model, thus allowing him to "look" at the screen from a similar perspective as the human, sharing a task area. Although somewhat limited in realistic gestures by his floating hand, the cartoon nature of his appearance was a conscious decision. The more realistically human the figure, the higher people's expectation of that character's ability to understand, reason, move and behave as human. Although some of these expectations are useful, all of these are active areas of research and since we cannot fully simulate a human, it can be unwise to raise expectations to the point that the deviations are more detrimental to the interaction than the benefit gained.

5.2.2 Generation of Multi-modal Behavior

Thórisson [39] discusses many subcategories of facial gesture, gaze, and gesture as well as different types of utterances and turn-taking cues. Of these, only a representative sampling were implemented in Gandalf. I will now discuss each of these areas in turn,

pointing out which behaviors have actually been tried.

In regards to Gandalf's facial behaviors, a smile was utilized to indicate that Gandalf knew he was being addressed or greeted, in addition to being part of his "happy" face. He wore his neutral face when not in a dialogue and when performing an action in the solar system model. Show-listen, look-pensive, look-aloof, and look-puzzled were also used at the process control level as feedback to indicate the internal state of Gandalf. For example, if Gandalf knew he had received some speech input, but had not yet processed it to the point of being ready to say something, he may show a look-pensive face. Thus the user would be less likely to wonder if he had even been heard and repeat himself. Gandalf could also use nods for backchannel feedback between speaker's utterences. Blink was included as a periodic function.

Gaze was used extensively and effectively in the output behavior set of Gandalf. Morphologically, Gandalf was able to face the user or the task screen. He could also look at the screen, the user, or elsewhere. Functionally, he turned to or at least looked at the user when delivering speech, when indicating that he knew he was being addressed, when done changing something in the work place, and when yielding the turn. He turned to the work place when the user was facing it and when he was changing something there. Taking the turn was signified in part by a quick glance sideways and back. Hesitation was sometimes signified by gazing up.

Speech was relatively limited; a user's utterance was parsed as one of 8 types: confirmation, action request, question, name of a planet or moon, back-channel, name, greeting, goodbye, or was rejected. For each type of utterance there was an appropriate output behavior. In cases where this behavior was speech, a pre-coded string annotated for intonation was run through a text-to-speech system.

Gandalf's output gestures were not very extensive. Even though the user's gestures were considered to some extent in Gandalf's turn-taking processing, he performed primarily content based gesture.[1] Of the propositional gestures, he could perform a beat with a variable duration, and a deictic which took a direction. However,

[1]The one exception to this was his impatient finger drumming which was triggered when the user was not facing Gandalf.

since he only had a small fixed number of pre-determined utterances, his propositional gestures were similarly hard-coded.

Process level cues that Gandalf could understand to some extent were when the user was taking, giving or wanting the turn, or when wanting back-channel feedback. Similarly, Gandalf himself could take the turn (by raising his eyebrows and quickly glancing sideways), give the turn (by turning to user and showing a neutral face), or provide back-channel feedback (by saying "aha" or nodding). In addition, he could hesitate when he had the turn but was not yet ready to speak (by saying "ahh", gazing up, or looking pensive).

5.3 Overview

Both projects discussed here use multi-modal behavior drawn from life. Yet, when compared, it can be seen that they use non-verbal behavior for in different ways. Animated Conversation concentrated on how integrate modalities such as intonation and gesture in the speech generation process. It successfully demonstrated how to perform the interactive function of highlighting propositional content by using non-verbal channels along with speech. However, there was no way to evaluate its treatment of turn-taking and backchannel behavior since the agent co-operated through channels other than sight and sound to co-ordinate the actual flow of the conversation.

Gandalf, on the other hand, was capable of taking turns with a human and producing backchannel behaviors appropriately. However, since the system could only recognize and produce language, but not really understand or generate it, there was little way or reason to generate interactional or propositional behaviors which accompany speech content.

Chapter 6

Rea

The current communicative humanoid research in the Gesture and Narrative Language Group at the MIT Media Lab involves designing and building an embodied conversational agent, Rea. The name of the character stems from her domain knowledge as a Real Estate Agent. However, while she is far from being an expert in selling houses, she will have far more knowledge about conversational cues than the average computer character. In this chapter, I will first summarize the goals and the modules of Rea, and then concentrate on the Generation Module.

6.1 Goals of the Rea Project

Rea is the current instantiation of the goals set out in the introduction, namely to create a conversational humanoid who is capable of understanding and conveying propositional information in a variety of channels and who is able to utilize interactional multi-modal cues to help structure the flow of the conversation. Our end goal is to allow computers to successfully hold conversations with humans using non-verbal signals such as gaze, intonation, gesture, and facial expression along with speech. Not only do we want to be able to detect conversational cues in these various modalities, we want to be able to understand the functions they serve, and be able to produce them in a meaningful manner consistent with human expectations of such behaviors.

In particular, we aim to bring together the successful aspects of both Animated

Conversation and Gandalf in a new architecture that facilitates both testing of hypotheses about multi-modal behavior, and the expansion of the character's abilities. This means that we wish to have a character, like Gandalf, who can interact in real time with a human using gaze, intonation, and gesture to discover and create the flow of a conversation with smooth turns and back-channel feedback. We also wish to be able to create realistic gesture, intonation and speech from the same propositional material, as in Animated Conversation.

One of the difficulties in reaching this goal is that the constraint of running in real time will require a trade off between linguistic processing and reactivity. Our goal is to design Rea such that her reactions are aptly timed, yet retain as many of the types of information found in the information structure of Animated Conversation as possible, particularly theme and rheme for appropriate intonation generation. To begin with, we will likely not include as much in-depth linguistic processing as the creators of Animated Conversation did. However, part of our design process has been focused on making the architecture both straight-forward and modular enough to be able to upgrade incrementally and allow more in-depth processing in later versions of conversational agents.

The ability to use this architecture to test psycholinguistic theories is important to us as well. One example of a hypothesis that we would like to study with the Rea project is that the addition of interactional information helps in the creation of mixed initiative dialogue. We would like Rea to be able to participate in small talk as well as respond to user requests and questions. It is towards this end that we have chosen the domain of a real estate agent for our new character, since a real estate agent is someone with whom the interaction regularly consists of a mixture of social talk and goal oriented discussion.

6.2 Physical Set-Up

Rea will be able to see and hear her human interlocutor using two video cameras and a microphone. She will use information such as whether the user's hands are in gesture

space, which direction the user is facing, when the user begins and ends speaking, and what the speech recognition system comprehends to understand what the user is trying to convey. As we are able to better detect and categorize intonational patterns and different types of gesture, those features will also be included in the input analysis. We are using Stive, a specialized computer vision system for tracking hands and face which was developed in the MIT Media Lab Vision and Modeling Group, and a research speech recognizer from British Telecom.

On the production side, Rea will also utilize the channels of speech, intonation, gesture, gaze, and facial expression to communicate to the user. The body and house environment were created in 3D Studio MAX and Cosmo Worlds and are animated using C++ calls to OpenInventer. Speech is produced through Entropic's text-to-speech system, TrueTalk, which allows a wide variety of intonational contours to be produced through annotation of the text.

6.3 Architecture

The conversational character architecture is being developed jointly by Lee Campbell, Joey Chang, Erin Panttaja, Mukesh Singh, Hannes Vilhjalmsson and myself under the direction of Professor Justine Cassell of the MIT Media Lab Gesture and Narrative Language Group, and by Scott Prevost and Tim Bickmore under the direction of Joseph Sullivan at the FX Palo Alto Lab Office Avatars Project [16].

The architecture contains six major modules: the Input Manager, the Understanding Module, the Reaction Module, the Response Planner, the Generation Module, and the Action Scheduler, as well as a knowledge base and a discourse model. (See figure 6-1.) Information primarily flows from left to right along two different paths, depending on whether the nature of the information is reactive or deliberative.

6.3.1 Input Manager

The input manager is responsible for interfacing with all devices which collect input from the user's area. Once it has determined a significant change in state, it passes

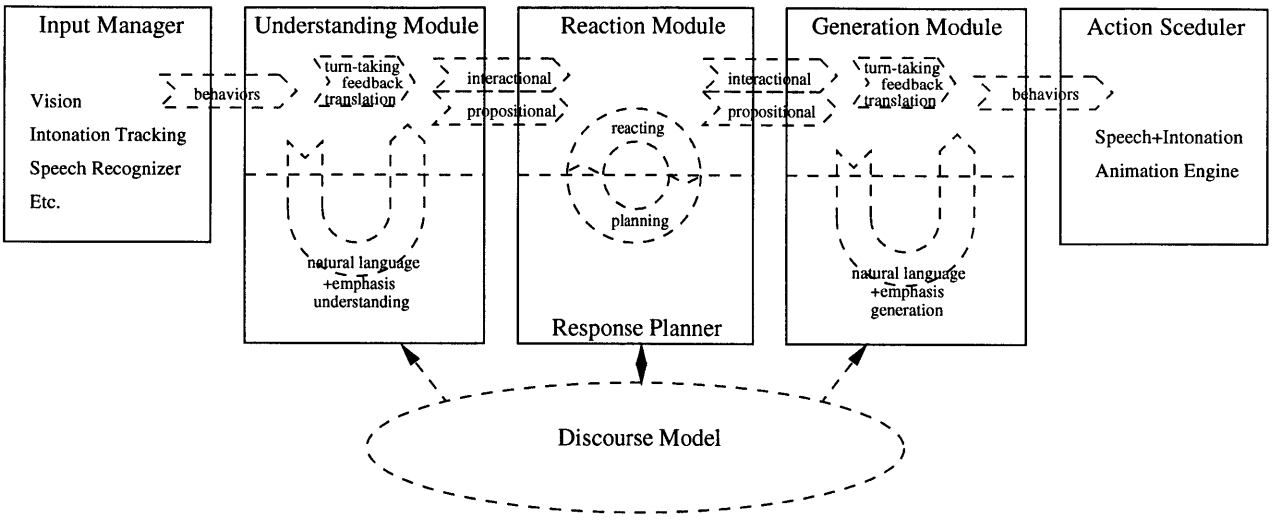

Figure 6-1: Conversational Character Architecture

that information on to the Understanding Module. This includes simple signals such as speech-on and hands-in-gesture-space as well as whole utterance parses with included gestures, as speech recognition and gesture tracking occur in this module. The reason we pass on the simple, seemingly redundant information such as speaking-on is that we wish our character to be reactive to a user, not only at the end of an utterance when the speech recognizer produces a parse, but also when the user is speaking.

6.3.2 Understanding Module

Once a package is received by the Understanding Module, the multi-modal information therein is interpreted with respect to the discourse model into interactional and propositional information. And example of an interactional translation may be the following rule.

> **hands-in-gesture-space** (*from the input manager*) + **speaking-turn-open** (*from the discourse model*) $\Rightarrow$ **user-taking-turn**

In general, interactional information will be processed quite quickly and be sent immediately on to the reactive module. However, some information that results in propositional material such as complex speech parses may take longer to process.

"It's not bigger than <deictic 300,600> that < intonation rise>"

For example the above parse from the input manager may require inquiry of the discourse model to de-reference "it", an inquiry of the knowledge base to determine what object is situated at location x = 300, y = 600, and analysis of the intonation to determine that it is a question. The resulting propositional message would be a query about the relative sizes of two objects.

6.3.3 Reaction Module and Response Planner

The Reaction Module and Response Planner are the central nervous system and brain of Rea. The Reaction Module receives interactional and propositional information and determines the appropriate reaction, whether this be to start a process in the Response Planner, or to send an immediate interactive response through the later half of the system. An example of a decision that this module might make is if it received a **user-wants-turn** interactional signal from the Understanding Module and it knows that the character is still in the middle of a speech, it could issue a **yield-turn** message or a **hold-floor** message. The result of this decision would be that Rea would stop speaking or perhaps increase its volume and not pause between utterances. Nor does the Reactive Module need to wait for input to initiate action. Its one big constraint is that it should be able to process new information and make decisions quickly as all messages in and out of the system will pass through it, and thus will be needed to keep the character reactive.

The Response Planner is used to generate both linguistic and task-oriented plans for the character, and to chart the character's progress towards achieving them. The planner has access to the knowledge base and the discourse model. In the early implementations of the architecture, these two modules will likely reside in the same piece of software, although with at least two threads of operation so that planning, which takes a while in the Response Planner, does not hold up processing in the Reactive Module.

6.3.4 Generation Module

The Generation Module has the reciprocal task to the Understanding Module, translating high level interactional commands such as **provide-backchannel-feedback** into the actual actions, such as **nod** and **say "uh huh"**, which can be sent to the Generation Module. I will provide a more in depth look at the Generation Module in section 6.5.

6.3.5 Action Scheduler

The Action Scheduler is the section responsible for the realization of the motor skills. It runs the graphics and sends text to the text-to-speech system. It receives two types of messages, which roughly correspond to propositional and interactional: those that are cued, and those that are not. For actions which are not cued, the action scheduler will be provided with a list of possible ways to realize the higher level function (such as **nod**, **say "oh yeah?"**, or **raise brows** for the interactive function **give-backchannel-feedback**) so that should the modality of choice be otherwise busy, the Action Scheduler may still perform an appropriate behavior in a timely fashion. Actions which are cued will primarily be speech utterances and their accompanying non-verbal behaviors. Thus the response planner and Generation Module need not produce a whole turn's worth of speech at once. Rather, they can construct one utterance at a time, doing part of the work while the first sentences are actually being spoken.

One of the important tasks that the Action Scheduler will need to be able to do is synchronize speech and animation. This is important both for the mouth movements and when producing co-occurring speech and gesture. In addition, the Action Scheduler will be responsible for the smooth co-articulation of gestures which occur one after another without relaxing to a canonical position in between, as well as the preparation and relaxation phases of isolated gestures.

6.3.6 Intermodule Communications

Each of the modules is an independent process, able to be started or stopped without affecting the other state of the other running modules. Communications between the modules is accomplished by passing messages across sockets, of which the top level information corresponds to a KQMLPerformative. One of the reasons for this was so that any module could be written in a different language on a different type of machine and still be able to communicate with the rest of the system.[1] The specification for the syntax of a frame is still being modified. However, since many of the major pieces are in place, I have included the current version in Appendix B.

The performative is a multi-layer frame which was designed to be able to be passed as plain text between modules, thus further facilitating the use of differing computer architectures and programming languages. At the top level, a performative can be of type **tell** or **ask**, and it contains fields for specifying the sender and recipient modules as well as the time it was composed and a time after which it should be disregarded. The content makes up the rest of the performative. Within Rea, most of the messages will be **tell** messages which contain a **commact** (communicative act) frame. **Commact** frames, in general, contain either optional *:prop* and *:intr* fields for high level propositional and interactional information, or a list of low level *:input* or *:output* **behavior**s, as well as fields for indicating the *:sender* and *:recipient*[2]. At this level the system is symmetric with regard to the types of messages being passed around. On the input side, the Input Module sends **behavior** frames to the Understanding Module which then interprets these and passes on *:prop* and *:intr* information to the Reaction Module. Similarly, on the output side, the Action Scheduler receives **behavior** frames from the Generation Module, which receives *:prop* and *:intr* information from the Reaction Module.

[1] Currently, all of the modules in the Rea project are being implemented in C++. However, the FX Palo Alto researchers are implementing much of their system in Java, while conforming to the same intermodule specifications.

[2] The *:sender* and *:recipient* fields have values such as USER and REA and are included here primarily for readability. In our current system with only two participants, the same information can be inferred from looking at whether the modules sending and receiving the performative are on the input or output branch of the system.

The value associated with *:prop* is a frame which indicates what type of propositional material is being conveyed. To date, portions of the system dealing with propositional content have not yet been implemented, and this part of the specification is still being modified. However, one of the important aspects we are including is a way of designating what propositional material is theme and what is rheme.

The value associated with the *:intr* keyword is a list of keywords which indicates interactional level functions. Currently, this list can contain any of: wantingturn, givingturn, droppingturn, takingturn, attending, present, wantingfeedback, givingfeedback, leaving, and notattending. These functions are used on both the understanding and generation sides.

The **behavior** frames are more complicated, as it is necessary at the input and output stages to allow compositions of actions occurring simultaneously and sequentially. Thus, the value of a behavior frame can be a string representation of text, an indication of a single modal action, or a combination of these. It is here that we really begin to see the differences between current output and input technology. Examples of input actions include (**gesturing** *:state* true *:motion* true) and (**speaking** *:state* false), whereas output actions include one such as (**head** *:state* looktowards), (**head** *:trans* glancetowards *:dur* 2) and (**inton**[3] *:tonetype* accent *:tone* H* *:prominence* 2.0).

In addition, there may be cases in which the Input Manager will want to pass on multiple possible parses, or the Generation Module will send along many different ways of realizing some discourse function. For these possibilities, we have included a way of indicating that there is a **choice** of options. (For exact syntax see Appendix B.)

6.4 Example

As mentioned in the overviews of the various modules, most of the modules have a way of processing some interactional information quickly, as well a component which can spend more time on task such as speech processing and planning. Roughly speaking, there is a fast track and a deliberative track. The following example will demonstrate

[3] **inton** is short for intonation.

how these two tracks interact.

To begin with, let us assume that our conversational agent, Rea, and her prospective buyer are in the kitchen looking out the window. Rea has just pointed out what a great virtual view there is outside. Neither participant has spoken in a while and the discourse modal indicates that neither has the turn. Now the user says to Rea while looking at the view, "I see what you mean. That is nice. *Short pause while user looks briefly at Rea and then away.* UPSTAIRS is next?"

The first frame the Input Manager would send is one containing the information (**speaking** *:state* true). While the speech recognizer is still waiting for the end of the first utterance, the Understanding Module receives this **speaking** frame and will inform the Reaction Module with a frame containing *:intr* [takingturn]. The Reaction Module will then update the discourse model to indicate it is the user's turn, as well as issuing a frame containing *:intr* [attending] to indicate that Rea should pay attention to the user. The Generation Module will check and find that the state of Rea's body is facing away from the user and that Rea is looking out the window, so it will issue a (**choice** [[(**body** *:state* attend) (**head** *:state* looktowards)] [(**eyebrows** *:trans* raise :dur 2) (**head** *:state* looktowards)]]) message. The Action Scheduler will then choose the first item on the list it can perform, which in this case is [(**body** *:state* attend) (**head** *:state* looktowards)] and Rea will turn and face to the user.

Meanwhile, the user continues to speak. When he gets to a break and glances at Rea, the Input Module sends on (**speaking** *:state* false) and (**attention** *:facing* true). The Understanding Module checks the discourse model and finds that it is the user's turn, so the user could either be giving the turn or asking for feedback. It can then issue a *:intr* [wantingfeedback], which will pass through the system, eventually resulting in Rea nodding or saying "mm hmm". Thus far, all of the messages passed through the system have followed the fast interactional track.

At this point the Input Module again indicates that the user is speaking, and passes along the parse of the first utterance with co-occurring gesture or intonation information[4]. For the first time, the Understanding Module has something to deliber-

[4]The parsing and understanding of speech and co-occurring intonation and gesture has not yet

ate about so it passes on the parse to the natural language understanding portion of the module. This portion tries to understand both the current speech parse and which part of that speech is rheme and which part is theme from a combination of looking at non-verbal information and previous discourse. It then sends on a propositional frame indicating that the user shares the opinion that the view is nice. Once the Reaction/Planning Module receives this it will update the discourse model and perhaps decide that because the user still has the turn, and the last propositional material was not an imperative or interrogative, and did not conflict with any information in Rea's knowledge base, it can be ignored.

Meanwhile, the user has again ceased to speak, which results in a quick message though the system and some more feedback from Rea. This time, however, the Understanding Module pieces together the cues that the user has continued to look at Rea for longer than a brief glance, is remaining silent, and has last uttered something with a question intonation[5], and so it passes on a *:intr* [givingturn] message. Once the Reaction Module receives this interactional information, it can issue a message indicating Rea that should take the turn (*:intr* [takingturn]). This results in a (**choice** [[(**eyes** em :trans lookaway) "um"] [(**face** *:state* think)]]). Rea will then either glance away and say "um", or show a face which indicates she is thinking.

While Rea has reactively taken the turn, the content of the user's last utterance is still being processed. First the Input Module passes along the words and the fact that "upstairs" was intonationally emphasized, as well as the fact that the utterance ended in a final rise. Then the Understanding Module forms a propositional **query**, to which the Reactive Module can find an answer and issue a frame containing both a propositional **assert** and an interactional [givingturn]. The **assert** will contain enough information for the Generation Module to formulate an utterance which puts contrastive stress on the word "downstairs". Once the Generation Module has finished

been implemented. Parsing of speech should be implemented by October 1998. However, the identification of meaningful gesture and intonational features are as yet unsolved problems and are currently being researched by members in our and other research labs.

[5]A situation like this is a good example of why we might want to track intonation quickly, and use the information before we even have a speech parse. Panttaja[29] discusses a real-time algorithm for determining the difference between questions and statements.

generating the speech, it will send a frame to the Action Scheduler including both the speech and the multi-modal cues for giving the turn, which in this case are looking at the user and adding low boundary tones to the utterance. Rea then says "Yes, we have seen all of the DOWNSTAIRS". Thus concludes this example.

6.5 A Closer Look at the Generation Module

One of the Generation Module's functions is to serve as an abstraction barrier between the Reaction Module and the Action Scheduler, allowing the Reaction Module to operate in high level discourse concepts while the Action Scheduler functions just with physical operation of the body. This is achieved in two ways. The first is to perform the translation from high level turn-taking and feedback functions into lists of ways to perform those functions using multi-modal actions. As in the other modules, this function must be performed relatively quickly so that the reactive behaviors which pass through it appear soon after the stimulus.

The second is to do the word selection and surface generation for the propositional information coming through. This processing may also involve the generation of propositional multi-modal behaviors. For example, should the user ask, "where is the nearest outlet?", an appropriate reply would be "There is one just over < deictic> there." which includes a propositional deictic gesture. In addition, determining how to realize emphasis will occur along with the generation of speech, whether the final form involves a raising of an eyebrow, a beat gesture, intonational stress or even just a change in syntax. As we expand the speech processing capabilities of the system, the text generation in this module will likely be a very major portion of the module. However, for the initial implementation and for the rest of this section, I will focus primarily on the generation of the turn-taking and feedback behaviors.

As demonstrated by the message, in section 6.4 indicating that Rea should give the turn, propositional and interactional information need not always be in individual messages. If the Reaction Module has determined that an utterance by Rea will be the last before relinquishing the turn, it could send both *:prop* material and *:intr*

[givingturn] in one message. Then, the Generation Module could choose to decrease the pitch range towards the end of the last statement or indicate to the Action Scheduler that a gaze toward the user should be performed just before finishing the string.

Thus, even though many of the interactive functions such as **takingturn** and **givingfeedback** are relatively independent of speech and result in behaviors passed immediately on to be scheduled as soon as possible, not all are. Another example of this would be the addition of "y'know" in the middle of a turn with the accompanying interactive gesture. The interactive function being served may be to elicit back-channel feedback, but the execution is tied to speech and thus is executed in relation to it.

6.5.1 Behavior File

To facilitate the testing of different behaviors, I have designed the Generation Module such that none of the associations between the functions and actions is hard-coded. Rather, these associations reside in an external file which can be modified separately from the code. Thus, a researcher interested in studying different behaviors need not also be a programmer. All that is needed is to change the mapping file and restart the Generation Module, which will read in the file as it initializes[6]. The ability to alter the mapping by simply providing a file could also be useful if we wanted to develop different personalities or nationalities. However, in terms of implementation, the mapping between turn-taking/backchannel functions and multi-modal behaviors is relatively simple. It is more or less a look up table. I suspect that the complexity of this process will increase significantly with the addition of verbal capabilities, in which case it may not be desirable to keep all of mapping information in an external file. This remains to be examined in more detail.

Another feature of the current behavior file is that associated with every interac-

[6]This feature, along with the ability to start and stop any module without affecting the operation of the others has already been extremely valuable when trying to make alterations in the behavior set during operation.

tional function is a list of different ways in which to realize it, each with a probability. (See table reftable:reab for an example of mappings with probabilities.) So, rather than always repeating the same actions, the character has some variability. First, the Generation Module is able to randomly choose behaviors, and then it can also provide a ranked list of them so that the Action Scheduler can try to perform the next behavior on the list if it has a conflict with the previous one.. This will begin to address the problems we saw in the Wizard of Oz experiment, in which the actions seemed distractingly scripted.

As work progresses, it seems like it would also be useful to have a small set of state variables corresponding to those body parts which can be requested to change state for longer than some pre-defined amount of time. For example, in the case where Rea takes the turn before actually having some propositional content to say, the Generation Module may send a (**face** *:state* think) frame. However, when it does actually have some speech ready for production, it will likely want to change the face back into a neutral face. This could either be accomplished by having the Generation Module "remember" to turn off the thinking face, or by having the Action Scheduler "know" that anytime speech is produced, the face should be reset. This is an issue that has not yet been fully worked out, and is closely related to the issue of being able to send the Action Scheduler a list of behaviors so that if one modality is busy another can be used. On the one hand, the Action Scheduler has the best information about which animation motors are free and available to be used. However, the Generation Module has the information about the function of the multi-modal actions, which allows the Action Scheduler to be ignorant of discourse level phenomenon.

Table 6.1 shows a section of the current function/action mapping, as it appears in the behavior file. The remainder of the translations appear in Appendix A.

Table 6.1: Part of the Function-Action Translation Table

attending	[[0.4	[(eyes :trans glancetowards :dur .500) (pause :dur .550) (head :state looktowards)]]
	[0.4	[(head :state looktowards)]]
	[0.2	[(head :trans nod :dur 2.000) (eyebrows :trans raise :dur 1.900)]]]
notattending	[[0.4	[(head :state lookaway)]]
	[0.4	[(eyes :state lookaway)]]
	[0.2	[(body :state idle)]]]
givingfeedback	[[0.3	[(head :trans nod) (eyes :state looktowards)]]
	[0.3	[(head :trans nod) (inton :rate 3.0) (inton :tonetype phrase :tone H :prominence 1.1) "mm" (inton :tonetype accent :tone H* :prominence 1.4) "hmm." (inton :rate 1.0) (eyes :state looktowards)]]
	[0.075	["I see."]]
	[0.075	["Yes."]]
	[0.075	["Right."]]
	[0.075	[(inton :rate 3.0) (inton :tonetype phrase :tone H :prominence 1.1) "mm" (inton :tonetype accent :tone H* :prominence 1.4) "hmm." (inton :rate 1.0)]]
	[0.05	[(inton :tonetype phrase :tone H :prominence 1.9) "uh" (inton :tonetype accent :tone H* :prominence 1.4) "huh."]]
	[0.05	[(inton :tonetype phrase :tone H) "uh" (inton :tonetype accent :tone H* :prominence 1.9) "huh."]]

Chapter 7

Future Work

With the implementation of Rea only just begun, many of the issues I have touched upon in this thesis have yet to be turned into integrated code. Currently, we have some reactive interactional behaviors implemented, in particular, the ones most useful in negotiating the exchange of turns and backchannel feedback. By August 1998, we hope to have in place some speech recognition and some natural language processing, allowing us to look into processing and generation appropriate emphasis. However, there are future directions which I believe are beyond the first design of Rea, and these I would like to mention here. In addition, there are places in which the design decisions already made have drawbacks as well as advantages. These I will also point out.

7.1 Possible Directions for Rea and the Generation Module

As we work towards learning which behaviors actually aid the flow of conversation between humans and computers, I believe it will become more apparent which ones can be exaggerated or deemphasized in order to give agents different personalities without making them unnecessarily dramatic. I think that the Generation Module is an ideal place in the architecture to begin combining the two different lines of

research, since it acts as a buffer between the basic communicative functions and the actions used to perform those functions. As an example, the Generation Module could make a personality more nervous by increasing the number of ways in which feedback is requested, such as by including more interactional gestures, "unsure" intonational patterns, and questioning facial expressions.

Another large area of research is in the actual generation of propositional non-verbal behaviors. For instance, we know that gesture can convey important propositional information, such as shape or motion[12], but we do not know how to decide what information should be conveyed in gesture and what should be conveyed in speech. In addition, we must learn how to generate the gesture. Animated Conversation addressed this issue to some extent, but it can be more fully explored.

One method that was utilized in Animated Conversation to learn about topic specific gesture was to videotape people talking about banking. We intend to do a similar thing with actual Real Estate agents giving house tours. This will not only provide us with sample gestures, but also with examples of appropriate propositional speech.

Even interactional gesture can be improved through more studies from real life, as can other aspects of Rea. Actions such as phrase level conduit gestures, gesture cohesion, and individual word emphasis vs. whole phase emphasis, are still open areas for implementation.

7.2 Drawbacks to Rea

Although the Rea project provides a good platform for many areas of research in human computer interaction, there are drawbacks with the approach that we are taking. The first is that with so many research issues brought together in one project, it is difficult to give each the attention is deserves. Therefore, some aspects must wait longer than others to be implemented. A current example of this is the propositional language processing. Although the final goal is to be able to combine the functional aspects of Animated Conversation and Gandalf, we have, so far, done much more

work on the turn-taking capabilities of the system than on the speech content and manner of emphasis.

The problem of having such an overwhelming research objective as face-to-face conversation is mitigated in part by the strong modularization of the system. Vision capabilities can be upgraded independently of language skills. Changes in speech generation do not affect the Action Scheduler, and so on. Now that the framework is set up, the different modules can be incrementally improved. However, though strong modularization is good in some ways, it can isolate decision making modules from the information they need. For example, we can envision that in a noisy environment, gesture understanding might actually help speech recognition, but since Rea's understanding happens after recognition of gesture and speech, the understanding process cannot direct and aid the recognition process. Another case was mentioned in section 6.5.1 in which the Generation Module knows how discourse functions map to cues, but the Action Scheduler has the most reliable information about what motors are busy.

Other drawbacks are less important from a research perspective, but are important should a system like this be commercialized. These include places in which the input must be of a fairly constrained nature. For example, the current vision system tracks skin, thus a person with a bald spot or a short sleeved shirt can easily confuse it. In addition, at this point it still looks as if users will need to have an American accent and use a constrained vocabulary. Another prohibitive factor is that with some 3D computer vision and animation, and commercial speech recognition, the system requires quite a bit of computing power.[1]

7.3 Experiments

It is important as we do research in human-computer interaction that we continue to test the results of our work. The experiment described in chapter 4 is one example of the tests performed with Gandalf. With the Rea project, we are still very much in

[1]The current system runs on 2 SGI O2's, one SGI Octane, and one HP.

the implementation phase. But as work progresses, we need to consider what sorts of experiments can be used to evaluate our ideas. As I mentioned in section 6.5.1 I envision that keeping some of the behaviors in an external file will be useful in the testing of those behaviors.

Some of the experiments that we may want to run are ones which explicitly test the effects of having back-channel, feedback, and emphasis behaviors, using an agent both with and without a graphical embodiment. Another possible experiment would be to see if it is more useful to have immediate reactive feedback or to wait until the system has had a chance to "understand" the input and provide feedback with semantic information. It may be that a combination of both of these would work best, relying on more explicit feedback in the face of frequent misunderstandings and breakdowns in conversation.

7.4 Summary

In this thesis I have presented an approach to the determination of non-verbal behaviors for conversational characters that is based on the study of behaviors in human-human interactions. I have suggested that one way of dividing up these behaviors is by looking at the functions they perform in conversation, in particular into categories of turn-taking, feedback, and emphasis. I have discussed how this approach differs from other approaches to building characters, namely approaches in which drama and believability are more important than conversational interaction between the computer character and a human.

Next I described an experiment in which we are testing the effects of using communicative envelope behaviors on human-computer interaction.

I have also presented past and current research within our group which is based on this approach in the form of summaries of the Animated Conversation, and Gandalf projects, and an introduction to a new character, Rea. I discussed in more detail the Generation Module of the Rea project showed one way in which the modularity of the system aids in the study of communicative behaviors by providing a single location

in which someone can associate functions with actions.

Finally, I have pointed out ways the Rea project may be limited and future directions to take it, both in terms of implementation and testing.

Appendix A

Function-Action Translation Table

Table A.1: Function-Action Translation Table

takingturn	[[0.3	[(eyes :trans glancetoward :dur .5) (eyes :state lookaway)]]
	[0.3	[(eyes :trans glancetoward :dur .5) (eyes :state lookaway) (face :state think)]]
	[0.3	[(head :state lookaway) (inton :tonetype phrase :tone H :prominence 1.0) "well," (eyes :trans glancetoward :dur 1)]]
	[0.1	[(eyes :trans glaceaway :dur 2) (inton :rate 1.5) (inton :tonetype phrase :tone H :prominence 1.2) "um," (inton :rate 1.0)]]]
leaving	[[1.0	[(body :state leave) (head :state lookaway)]]]
givingturn	[[0.2	[(body :state attend) (head :state looktoward) (face :state neutral)]]
	[0.2	[(body :state attend) (eyes :state looktoward) (face :state neutral)]]
	[0.2	[(eyebrows :trans raise :dur 2.0) (head :state looktowards)]]
	[0.2	[(eyebrows :trans raise :dur 2.0) (head :state looktowards) (gesture :trans drop)]]]

attending	[[0.4	[(eyes :trans glancetowards :dur .500) (pause :dur .550) (head :state looktowards)]]
	[0.4	[(head :state looktowards)]]
	[0.2	[(head :trans nod :dur 2.000) (eyebrows :trans raise :dur 1.900)]]]
notattending	[[0.4	[(head :state lookaway)]]
	[0.4	[(eyes :state lookaway)]]
	[0.2	[(body :state idle)]]]
givingfeedback	[[0.3	[(head :trans nod) (eyes :state looktowards)]]
	[0.3	[(head :trans nod) (inton :rate 3.0) (inton :tonetype phrase :tone H :prominence 1.1) "mm" (inton :tonetype accent :tone H* :prominence 1.4) "hmm." (inton :rate 1.0) (eyes :state looktowards)]]
	[0.075	["I see."]]
	[0.075	["Yes."]]
	[0.075	["Right."]]
	[0.075	[(inton :rate 3.0) (inton :tonetype phrase :tone H :prominence 1.1) "mm" (inton :tonetype accent :tone H* :prominence 1.4) "hmm." (inton :rate 1.0)]]
	[0.05	[(inton :tonetype phrase :tone H :prominence 1.9) "uh" (inton :tonetype accent :tone H* :prominence 1.4) "huh."]]
	[0.05	[(inton :tonetype phrase :tone H) "uh" (inton :tonetype accent :tone H* :prominence 1.9) "huh."]]

present	[[0.5	[(body :state arrive)]]
	[0.5	[(body :state arrive) (head :state lookat)]]]
wantingturn	[[0.75	[(gesture :trans raise)]]
	[0.10	[(inton :rate 1.5) (inton :tonetype phrase :tone H :prominence 1.2) "but," (inton :rate 1.0)]]
	[0.10	[(inton :rate 2) (inton :tonetype phrase :tone H :prominence 1.2) "Um," (inton :rate 1)]]
	[0.05	[(inton :rate 2) (inton :tonetype phrase :tone H :prominence 1.2) "Um." (pause :dur .75) (inton :rate 1.5) (inton :tonetype phrase :tone H :prominence 1.7) "but." (inton :rate 1)]]]
droppingturn	[[0.5	[(inton :state halt) (gesture :trans drop) (gesture :state relax) (head :state looktoward)]]
	[0.5	[(inton :state halt) (head :state looktoward) (gesture :state relax)]]]

Appendix B

KQMLPerfomative Specification

```
The REA Performative Syntax Specification (v0.2) - 4/25/98

==[TOP-LEVEL]=============================================

(tell :sender WORD
      :recipient WORD
      :starttime TIME
      :endtime TIME
      :content  (commact ..)
                |(taskact ..)
                |(statechange ..)    )

(ask  :sender WORD
      :recipient WORD
      :starttime TIME
      :endtime TIME
      :content (query ..)  )

==[CONTENT]===============================================
```

```
(commact  :sender WORD
  :recipient WORD
  :input [ (behavior ..) .. ]
          :prop  (decl ..)
                |(int ..)
                |(imp ..)
          :output [ (behavior ..) .. ]
          :intr LIST           )

(taskact :action WORD )

(statechange :object WORD :state WORD )

==[BEHAVIOR]==============================================

(behavior :queued BOOL
          :content VALUE
          :starttime TIME
          :durtime TIME
          :probability DOUBLE
          :priority LONG
          :ext VALUE           )

VALUE = STRING | LIST | CHOICE | BEH_FRAME

LIST = [ VALUE ... ]

CHOICE = (choice :options LIST)

BEH_FRAME =
```

```
  // ** input behaviors **

        (speaking :state BOOL)
         (gesturing :state BOOL :motion BOOL)
         (attention :presence BOOL :facing BOOL)

  // ** output behaviors **

(head :state BEH_LOOKTOWARDS | BEH_LOOKAWAY
       :trans BEH_GLANCETOWARDS | BEH_GLANCEAWAY | BEH_GLANCEAROUND |
      BEH_NOD | BEH_SHAKE
       :dur TIME)

(eyes :state BEH_LOOKTOWARDS | BEH_LOOKAWAY
       :trans BEH_GLANCETOWARDS | BEH_GLANCEAWAY | BEH_GLANCEAROUND
       :dur TIME)

(eyebrows :state BEH_RAISE | BEH_LOWER
   :trans BEH_RAISE | BEH_LOWER
           :dur TIME)

(body :state BEH_ATTEND | BEH_IDLE | BEH_LEAVE | BEH_ARRIVE)

(gesture :state BEH_RELAX
 :trans BEH_TOWARDS | BEH_RAISE | BEH_DROP
          :dur TIME)

(face :state BEH_SMILE | BEH_FROWN | BEH_NEUTRAL |
      BEH_PUZZLED | BEH_THINK)
```

```
(inton :tonetype BEH_PHRASE | BEH_BOUNDARY | BEH_ACCENT |
 BEH_DEACCENT | BEH_LASTACCENT
       :tone BEH_BOUNDARYTONE_HIGH | BEH_BOUNDARYTONE_LOW |
     BEH_PITCHACCENT_HIGH | BEH_PITCHACCENT_LOW |
     BEH_PITCHACCENT_LOWH | BEH_PITCHACCENT_HLOW |
     BEH_PITCHACCENT_LHIGH | BEH_PHRASEACCENT_HIGH
     BEH_PHRASEACCENT_LOW
       :prominence DOUBLE
       :rate DOUBLE)

==[PROPOSITION-TYPES]====================================

(decl :speech STRING
      :given LIST
      :new   (assert ..)
            |(opinion ..)
    |(ritual ..) )

(int :speech STRING
     :given LIST
     :new   (query ..)
           |(reveal ..)
           |(compare ..)
           |(boolean ..)   )

(imp :speech STRING
     :given LIST
     :new   (travel ..)
           |(move ..)
```

```
           |(operate ..)     )

==[PROPOSITION]=======================================

(ritual   :type WORD)
(assert   :object WORD :property WORD :value VALUE)
(opinion :assertion (assert ..) :bias WORD)

(travel   :place WORD)
(move     :object WORD :place WORD)
(operate :object WORD :state WORD)

(query    :object WORD :property WORD)
(reveal   :object WORD)
(compare :comparison WORD :object1 WORD :object2 WORD :measure WORD)
(boolean :conclusion WORD :query( ..) )
```

Appendix C

Wizard of Oz Experiment Questionaire

blankpage

How well did you do on this task?

Not well • • • • • • • • • • very well

How much did you like Gandalf?

not at all • • • • • • • • • • very much

Would you enjoy working with Gandalf again?

not at all • • • • • • • • • • very much

How much did you cooperate with Gandalf?

not at all • • • • • • • • • • very much

How much did Gandalf cooperate with you?

not at all • • • • • • • • • • very much

How much did you trust the information from Gandalf?

not at all • • • • • • • • • • very much

How relevant was Gandalf's information?

not at all • • • • • • • • • • very much

How helpful was Gandalf's information?

not at all • • • • • • • • • • very much

How insightful was Gandalf's information?

not at all • • • • • • • • • • very much

To what extent do the following words describe the computer humanoid you worked with on the Desert Survival Task?

	Describes very poorly									Describes very well
Agreeable	•	•	•	•	•	•	•	•	•	•
Analytical	•	•	•	•	•	•	•	•	•	•
Competent	•	•	•	•	•	•	•	•	•	•
Credible	•	•	•	•	•	•	•	•	•	•
Expert	•	•	•	•	•	•	•	•	•	•
Friendly	•	•	•	•	•	•	•	•	•	•
Helpful	•	•	•	•	•	•	•	•	•	•
Imaginative	•	•	•	•	•	•	•	•	•	•
Informed	•	•	•	•	•	•	•	•	•	•
Insightful	•	•	•	•	•	•	•	•	•	•
Intelligent	•	•	•	•	•	•	•	•	•	•
Knowledgeable	•	•	•	•	•	•	•	•	•	•
Likeable	•	•	•	•	•	•	•	•	•	•
Pleasant	•	•	•	•	•	•	•	•	•	•
Rational	•	•	•	•	•	•	•	•	•	•
Reliable	•	•	•	•	•	•	•	•	•	•
Responsive	•	•	•	•	•	•	•	•	•	•
Useful	•	•	•	•	•	•	•	•	•	•
Warm	•	•	•	•	•	•	•	•	•	•

To what extent do the following words describe your <u>interaction</u> with Gandalf?

	Describes very poorly									Describes very well
Adversarial	•	•	•	•	•	•	•	•	•	•
Collaborative	•	•	•	•	•	•	•	•	•	•
Combative	•	•	•	•	•	•	•	•	•	•
Cooperative	•	•	•	•	•	•	•	•	•	•
Engaging	•	•	•	•	•	•	•	•	•	•
Enjoyable	•	•	•	•	•	•	•	•	•	•
Friendly	•	•	•	•	•	•	•	•	•	•
Frustrating	•	•	•	•	•	•	•	•	•	•
Fun	•	•	•	•	•	•	•	•	•	•
Interesting	•	•	•	•	•	•	•	•	•	•
Involving	•	•	•	•	•	•	•	•	•	•
Satisfying	•	•	•	•	•	•	•	•	•	•
Successful	•	•	•	•	•	•	•	•	•	•
Tedious	•	•	•	•	•	•	•	•	•	•

1. On a scale of 1 to 10, assuming that a human gets a score of 10, the smoothness of the interaction with Gandalf gets a score of ____.

2. Compared with interacting with a dog, the smoothness of the interaction with Gandalf is ...

   o...much more interesting.
   o...somewhat more interesting.
   o...about equal.
   o...somewhat less interesting.
   o...much less interesting.

3. Compared with interacting with a fish in a fishbowl, interacting with Gandalf is ...

   o... much more interesting.
   o...somewhat more interesting.
   o...about equal.
   o...somewhat less interesting.
   o...much less interesting.

4. Compared with any real animal (excluding humans), Gandalf seems ...

5. Compared to the most life-like character in any computer game or program you have seen, Gandalf seems ...

   o...incredibly life-like.
   o...very life-like.
   o...somewhat life-like.
   o...not very life-like.
   o...not life-like at all.

6. If the video screen with Gandalf's face had been turned off for the whole time, do you think your interaction with the computer would have been different?

   O Yes
   O Yes, perhaps, but not significantly.
   O No, probably not.

7. If Yes, how would it be different? (For each line mark A or B or leave blank)

	A		B
a)	O More fun	OR	O Less fun
b)	O More difficult	OR	O Less difficult
c)	O More efficient	OR	O Less efficient
d)	O Smoother	OR	O Less smooth
e)	O Other: _____.		

8. How helpful to the interaction did you find ...

a) ...the contents of Gandalf's speech?
   o Very helpful.
   o Somewhat helpful
   o Not helpful or unhelpful
   o Unhelpful
   o Counterproductive

b) ...Gandalf's head motion?
   o Very helpful.
   o Somewhat helpful
   o Not helpful or unhelpful
   o Unhelpful
   o Counterproductive

c) ...Gandalf's expressions?
   o Very helpful.
   o Somewhat helpful
   o Not helpful or unhelpful
   o Unhelpful
   o Counterproductive

d) ...Gandalf's gaze?
   o Very helpful.
   o Somewhat helpful
   o Not helpful or unhelpful
   o Unhelpful
   o Counterproductive

e) ...Gandalf's hand gestures?
   o Very helpful.
   o Somewhat helpful
   o Not helpful or unhelpful
   o Unhelpful
   o Counterproductive

blankpage

Bibliography

[1] Ball, G., Ling, D., Kurlander, K., Miller, J., Pugh, D., Skelly, T., Stankosky, A., Theil, D., Van Dantzich, M., Wax, T. (1998) "Life-like Compuer Characters: the Persona project at Microsoft Research" *http://www.research.microsoft.com/research/ui/*

[2] Bavelas, J. B. "Gestures as Part of Speech: Methodological Implications." University of Victoria.

[3] Bavelas, J. B., Chovil N., Lawrie, D., Wade A.(1992) "Interactive Gestures." *Discourse Processes* 15 (pp 469-489)

[4] Blumberg, B., Galyean, T. (1995) "Multi-Level Direction of Autonomous Creatures for Real-Time Virtual Environments" *Computer Graphics Proceedings*Annual Conference Series.

[5] Bolinger, D. (1972) "Accent is predictable (if you're a mind reader)" *Language* 48-3

[6] Beckman, M. E., Elam, G. A. (1997) "Guidelines for ToBI Labeling." Version 3, March 1997. The Ohio State University Research Foundation. *http://ling.ohio-state.edu/Phonetics/E_ToBI/etobi_homepage.html*

[7] Brennen, S., Hulteen, E. (1995) "Interaction and feedback in a spoken language system: a theoretical framework" *Knowledge-Based Systems* Volume 8 Numbers 2-3 April-June

[8] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket T., Douville, B., Prevost, S., Stone, M. (1994) "Animated Conversation: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conersational Agents." *Proceedings of SIGGRAPH '94.* (ACM Special Interest Group on Graphics)

[9] Cassell, J., Stone, M., Douville, B., Prevost, S., Achorn, B., Steedman, M., Badler, N., Pelachaud, C. (1994) "Modeling the Interaction between Speech and Gesture." *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society.* Georgia Institute of Technology.

[10] Cassell, J., Prevost, S. (1998) "Embodied Natural Language Generation: A Framework for the Generation of Speech and Gesture." in preparation.

[11] Cassell, J. "A Framework for Gesture Generation and Interpretation." Cipolla R., Pentland A.(Eds.) *Computer Vision in Human-Machine Interaction.* Cambridge University Press.

[12] Cassell, J., McNeill, D., McCullough, K. (1998) "Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Nonlinguistic Information" *Pragmatics and Cognition* 6:2, 1998.

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

[14] Duncan, S. (1974) "On the Structure of Speaker-Auditor Interaction During Speaking Turns" *Language in Society* Volume 3, 161-180.

[15] Ekman, P. (1992) "Facial expressions of emotion: an old controversy and new findings." *Philosophical Transactions: Biological Sciences (Series B),* 35(1273):63-69.

[16] FX Palo Alto Lab Office Avatars Project and MIT Media Lab Gesture and Narrative Language Group (1997) "Conversational Character Software Architecture" internal document.

[17] Goffman, E. (1983) *Forms of Talk* Chapter 3. University of Pennsylvania Press, Philadelphia.

[18] Halliday, M., Hasan, R. (1976) *Cohesion in English* Chapters 1 &7. Blackwell, Oxford UK & Cambridge USA.

[19] Hiyakumoto, L., Prevost, S., Cassell, J. (1997) "Semantic and Discourse Information for Text-to-Speech Intonation" *Concept to Speech Generation Systems* Procceedings of a workshop sponcered by the Association for Computational Linguistics. Madrid, Spain.

[20] Kantrowitz, M. (1990) "GLINDA: Natural Language Text Generation in the Oz Interactive Fiction Project" Technical Report CMU-CS-90-158, Carnegie Mellon University, Pittsburgh PA USA, July.

[21] Kantrowitz, M., Bates, J. (1992) "Integrated Natural Language Generation Systems" Technical Report CMU-CS-92-107, Carnegie Mellon University, Pittsburgh PA USA, April.

[22] Kendon, A. (1980) "Gesticulation and Speech: Two Aspects of the Process of Utterance." In M.R. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207-227). The Hauge: Mouton.

[23] Loyall, A., Bates, J. (1997) "Personality-Rich Believable Agents That Use Language" *Agents '97* Marina del Rey CA USA

[24] McClave, E. (1991). "Intonation and gesture." Doctoral dissertation, Georgetown University, Washington, DC.

[25] McClave, E. (1994). "Gestural Beats: The Rhythm Hypothesis." *Journal of Psycholinguistic Research*, Vol. 23, No. 1 (pp45-66).

[26] McNeill, D. (1992). *Hand and Mind.* Chicago: University of Chicago Press.

[27] McNeill, D., Cassell, J., Levy, E. (1993) "Abstract deixis" *Semiotica* 95-1/2, 5-19

[28] Nakatani, C. H., Hirschberg, J. (1994). "A Corpus-based study of repair cues in spontaneous speech." *Journal of the Acoustical Society of America*, 95 (3), March (pp 1603-1616)

[29] Panttaja, E. (1998) "Recognizing Intonational Patterns in English Speech" M.I.T. Masters Thesis, to appear

[30] Pelachad, C., Badler, N., Steedman, M. (1994) "Generating Facial Expressions for Speech." *Cognitive Science*

[31] Perlin, K., Goldberg, A.(1996) "Improv: A System for Scripting Interactive Actors in Virtual Worlds" SIGGRAPH 1996.

[32] Picard, R., (1997) *Affective Computing* Chapter 6. The MIT Press, Cambridge USA, London UK.

[33] Pierrehumbert, J., Hirschberg, J. (1990) "The Meaning of Intonational Contours in the Interpretation of Discourse." In Cohen, Morgan, and Pollack (Ed.), *Intentions in Communication* (pp. 271-311)

[34] Pope, L., Smith, C. (1994) "Brief Report on the Distinct Meanings of Smiles and Frowns" *Cognition and Emotion.* 8(1), 65-72

[35] Prevost, S. (1996) "An Information Structural Approach to Spoken Language Generation." *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics.*

[36] Prevost, S., Steedman, M. (1994) "Specifying Intonation from Context for Speech Synthesis" *Speech Communication* 15, 139-153

[37] Schiffrin, D. (1987) *Discourse Markers* Cambridge Univercity Press, Chaps 3&10.

[38] Schiffrin, D. (1994) "Interactional Socialinguistics" *Approaches to Discourse* Chapter 4, Blackwell, Oxford UK & Cambridge USA.

[39] Thórisson, K. R. (1996) "Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills." Doctoral thesis Massachusetts Institute of Technology

[40] Tuite, K. (1993). "The production of gesture." *Semiotica* 93-1/2 (pp 83-105).

[41] Vilhjalmsson, H. (1997). "Autnomous Communicative Behaviors in Avatars." Master's Thesis MIT Media Lab.