Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds

David Traum
University of Southern California
Institute for Creative Technologies
13274 Fiji Way
Marina del Rey, CA 90292
traum@ict.usc.edu

Jeff Rickel
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
rickel@isi.edu

ABSTRACT

Immersive virtual worlds are increasingly being used for education, training, and entertainment, and virtual humans that can interact with human users in these worlds play many important roles. However, current computational models of dialogue do not address the issues that arise with face-to-face communication situated in three-dimensional worlds, such as the proximity and attentional focus of others, the ability to maintain multi-party conversations, and the interplay between speech and nonverbal signals. This paper presents a new model that integrates and extends prior work on spoken dialogue and embodied conversational agents, and describes an initial implementation that has been applied to training in virtual reality.

Categories and Subject Descriptors

I.2 [**Computing Methodologies**]: Artificial Intelligence

General Terms

Design

Keywords

Human-computer interaction, multi-agent systems, multimodal communication, spoken dialogue, virtual humans, virtual reality

1. INTRODUCTION

Immersive virtual worlds offer exciting potential for rich interactive experiences. Human users can cohabit three-dimensional graphical environments with virtual humans for entertainment, education, and training. They can have adventures in fantasy worlds. They can learn about history or other cultures by experiencing life in distant places and times. They can practice tasks, make mistakes, and gain experience without the consequences of real-world failure. In all these applications, virtual humans can play a wide variety of roles, including mentors and guides, teammates, companions, adversaries, and the local populace.

Perhaps the greatest challenge in creating virtual humans for interactive experiences is supporting face-to-face communication among people and virtual humans. On one hand, virtual worlds are an ideal application for current spoken language technology: they provide a microworld where conversation can legitimately be restricted to the events and objects within its confines. On the other hand, they raise issues that have received relatively little attention in computational linguistics. First, face-to-face communication in virtual worlds requires attention to all the nonverbal signals (e.g., gaze, gestures, and facial displays) that accompany human speech. Second, conversations that are situated in a 3D world raise a host of issues, including the attentional focus of the conversants, whether and to what degree they can see and hear one another, and the relative locations of conversants and the objects they are discussing. Finally, since there will typically be multiple real and virtual people, virtual worlds require support for multi-party conversations, including the ability to reason about the active participants in a conversation as well as who else might be listening or unaware of what happens in a conversation. While there has been some early work in the area of embodied conversational agents [10, 21], and some of this work has addressed human-agent dialogues situated in 3D virtual worlds [37], there is currently no general model of such dialogues.

In this paper, we present a candidate model of multi-party dialogue in immersive virtual worlds, drawing on prior models of collaborative dialogue from computational linguistics, as well as work on embodied conversational agents and the social psychology literature on the nonverbal signals that accompany human speech. The model is organized as a set of dialogue management layers, each including an information state and a set of dialogue acts that change that state. The layers include traditional ones, such as turn-taking and grounding, as well as several novel layers addressing the issues of multi-party dialogue in immersive worlds.

2. EXAMPLE SCENARIO

The test bed for our embodied agents is the Mission Rehearsal Exercise project at the University of Southern California's Institute for Creative Technologies. The project is exploring the integration of high-end virtual reality with

AAMAS'02, July 15-19, 2002, Bologna, Italy.

Figure 1: An interactive peacekeeping scenario (with sergeant, mother, and medic in foreground).

Hollywood storytelling techniques to create engaging, memorable training experiences. The setting for the project is a virtual reality theatre, including a visual scene projected onto an 8 foot tall screen that wraps around the viewer in a 150 degree arc (12 foot radius). Immersive audio software provides multiple tracks of spatialized sounds, played through ten speakers located around the user and two subwoofers. Within this setting, a virtual environment has been constructed representing a small village in Bosnia, complete with buildings, vehicles, and virtual characters. This environment provides an opportunity for Army personnel to gain experience in handling peacekeeping situations.

The first prototype implementation of a training scenario within this environment was completed in September 2000 [42]. To guide the development, a Hollywood writer, in consultation with Army training experts, created a script providing an overall story line and representative interactions between a human user (Army lieutenant) and the virtual characters. In the scenario, the lieutenant finds himself in the passenger seat of a simulated Army vehicle speeding towards the Bosnian village to help a platoon in trouble. Suddenly, he rounds a corner to find that one of his platoon's vehicles has crashed into a civilian vehicle, injuring a local boy (Figure 1). The boy's mother and an Army medic are hunched over him, and a sergeant approaches the lieutenant to brief him on the situation. Urgent radio calls from the other platoon, as well as occasional explosions and weapons fire from that direction, suggest that the lieutenant send his troops to help them. Emotional pleas from the boy's mother, as well as a grim assessment by the medic that the boy needs a medevac immediately, suggest that the lieutenant instead use his troops to secure a landing zone for the medevac helicopter. Figure 2 shows a small excerpt from the original script. Our virtual characters include enough knowledge and capabilities to support the types of interactions in the script as well as many variations, depending on user actions and external event triggers.

The interaction in Figure 2 illustrates a number of issues that arise for embodied agents, some going beyond capabilities of current implemented systems. First, at a broad level, the agents must concern themselves with multiple characters and multiple conversations. The main conversation is between the lieutenant and sergeant, but the medic is also brought in, and the mother is an important overhearer. Other platoon members and townspeople may also be potential overhearers. There is also a separate conversation between the sergeant and the squad leaders, starting at the end of the excerpt given here. In other parts of the scenario, the lieutenant engages in radio conversations with his home base, another platoon, and sometimes a medevac helicopter. Some of these conversations have explicit beginning and ending points (especially the radio conversations), while others, which are more focused on specific tasks, end without remark as the local purpose of the interaction is established and resolved and attention of the conversants shifts to other matters. In all cases, agents must reason about who they are talking to, who is listening, and whether they are being addressed or not.

In this immersive virtual world, the agents must also coordinate speech with other communicative modalities. In many cases, gestures and other nonverbal cues are important in carrying some of the communicative function. Some examples here are the way the lieutenant approaches the sergeant to initiate conversation, the way that the sergeant glances at the medic to signal that he should take the turn and respond to the lieutenant's question, and the way the medic glances at the mother while formulating a less direct answer about the boy's health — focusing on the consequence of his condition rather than directly stating what might be upsetting to her.

3. PRIOR WORK

Our work builds on prior work in the areas of embodied conversational agents [10] and animated pedagogical agents [21]. Several systems have carefully modeled the interplay between speech and nonverbal behavior [9, 11, 8, 34], but these systems have focused exclusively on dyadic conversation, and they did not allow users and agents to cohabit a virtual world. The Gandalf system [11] allowed an agent and human to cohabit a real physical space, and to use gaze and

actor	speech	nonverbal
LT		Drive up, exit vehicle, approach SGT
SGT		Look at LT
LT	Sergeant, what happened here?	
SGT	They just shot out from the side street sir.	Gesturing towards the civilian vehicle
	The driver couldn't see 'em coming.	
LT	How many people are hurt?	
SGT	The boy and one of our drivers.	Gesturing toward the boy
LT	Are the injuries serious?	
SGT		Makes eye contact with medic and nods
MEDIC	Driver's got a cracked rib but the kid's –	Glancing at the mother
	Sir, we gotta get a medevac in here ASAP.	
LT	We'll get it.	
LT	Platoon Sergeant, secure the area.	
SGT	Yes Sir!	
SGT	(Shouting) Squad leaders, listen up!	Raises arm, looks around at squad leaders
SGT	I want 360 degree security here now!	
SGT	First squad 12-4	Looks at 1st squad leader and gestures
SGT	Second squad 4-8	Looks at 2nd squad leader and gestures
SGT	Third squad 8-12	Looks at 3rd squad leader and gestures

Figure 2: Multi-modal, multi-character interaction excerpt (many nonverbal behaviors omitted).

gesture to reference an object (i.e., a wall-mounted display screen) in that space, but the agent's presence was limited to a head and hand on a 2D computer monitor. Similarly, the Rea agent [8] can transport herself to and into virtual houses and apartments, and the user can point to some objects within those virtual environments, but the user is not immersed in the environment, and Rea's movement and references within those environments is very limited. The Cosmo agent [27] includes a sophisticated speech and gesture generation module that chooses appropriate deictic references and gestures to objects in its virtual world based on both spatial considerations and the dialogue context, but the agent and its environment are rendered in 2D and the user does not cohabit the virtual world with Cosmo.

In contrast, Steve [37, 39, 38] cohabits 3D virtual worlds with people and other Steve agents, so it has addressed both multi-party and immersive aspects of dialogue in virtual worlds. Steve agents use path planning algorithms to move around in virtual worlds, they are sensitive to where human users are looking, they can use gaze and deictic gestures to reference arbitrary objects in those worlds, and they can use gaze to regulate turn-taking in multi-party (team) dialogues. However, while Steve includes a dialogue model built on ideas from computational linguistics [39], it falls far short of the models in state-of-the-art spoken dialogue systems. Moreover, the model focuses primarily on the context of dyadic conversations between a Steve agent and his human student; there is very little dialogue context maintained for the multi-party dialogues between a Steve agent and his human and agent teammates.

Work in computational linguistics has focused on a complementary set of issues: it has largely ignored issues of embodiment and immersion in virtual worlds, but has produced relatively sophisticated models of spoken dialogue that include a variety of hooks for multiple modalities. We follow the framework of the Trindi project [26], using *dialogue acts*[1] as abstract input and output descriptions for the dialogue modeling component. This serves particularly well for considering multi-modal communication, since it allows maximum flexibility of description, including acts that must be realized using a specified modality, acts that could be realized by any of a set of modalities, or acts that require realization using a combination of multiple modalities. We also view the dialogue acts (and the affiliated information states) as segmented into a number of layers, each concerning a distinct aspect of information state, and using different classes of dialogue acts. Moreover, there is no one to one correspondence between dialogue acts and atomic communication realizations: a single utterance (or gesture) will generally correspond to multiple (parts of) dialogue acts, and it may take several communications (sometimes split into multiple modalities) to realize some dialogue acts.

As a starting point, we use the dialogue layers developed in the TRAINS and EDIS dialogue systems [45, 35, 30]. These included layers for turn-taking, grounding, core speech acts, and argumentation acts (later termed forward and backward-looking acts [18]). There has also been other work on dialogue layers that, while not fully implemented within natural language dialogue systems, become important for dealing with multi-character, multi-conversation domains such as our Mission Rehearsal Exercise. This includes work by Novick on meta-locutionary acts, including an attention level [33], work by Allwood [2] and Clark [14] on basic communicative functions, work by Bunt on interaction management functions [7], and work on multi-level grounding in an extended multi-modal task interaction [17].

4. MULTI-MODAL DIALOGUE MODEL

Our agents are designed to run within the Mission Rehearsal Exercise environment [36, 42]. This environment includes a message-passing event simulator, immersive sound, and graphics including static scene elements and special effects, rendered by Multigen/Paradigm's Vega.

Our agent model is based on Steve, as described in the previous section. Within the peacekeeping scenario, Steve

[1] termed *dialogue moves* in the Trindi publications

- contact
- attention
- conversation
  - participants
  - turn
  - initiative
  - grounding
  - topic
  - rhetorical
- social commitments (obligations)
- negotiation

Figure 3: Multi-party, Multi-conversation Dialogue Layers

agents are given dynamically animated bodies from Boston Dynamics' PeopleShop [6]; some body movements come directly from motion capture, while other movements (e.g., gaze and gestures) are generated dynamically through procedural animation, and the agent controls both types of movements by sending commands in real-time to the animation software. The medic and sergeant include expressive faces created by Haptek (www.haptek.com) that support synchronization of lip movements to speech. We are augmenting Steve's dialogue model with the new dialogue model presented here.

Our dialogue model currently consists of the layers shown in Figure 3. Each of these is modeled from the perspective of an agent involved in the interaction. For each layer, there is both an *information state*, representing the current status of that layer, and a set of *dialogue acts*, each of which will correspond to well-defined changes to the information state. There are also conventional *signals* – behaviors that can be associated with the performance of dialogue acts, given the right context. In addition to this abstract model, an implementation will also have *recognition rules*, associating observed behavior (speech and/or other modalities) with performance of one or more of these dialogue acts, *selection rules*, which allow the agent to choose to generate one or more of these acts given certain conditions of the information state as well as other aspects of the agent's internal state, and *realization rules*, which specify what kinds of output behavior will perform or signal the performance of the selected acts in the current context. In this section, we focus on the information state and acts for each level, giving occasional examples of behavior that corresponds to acts. More detail on other aspects of the implementation is given in the next section.

We will first briefly describe each of these, and then give details of the layers and how the associated acts may be realized using the palette of multi-modal communicative abilities. The *contact* layer concerns whether and how other individuals can be accessible for communication. Modalities include visual, voice (shout, normal, whisper), and radio. The *attention* layer concerns the object or process that agents attend to. Contact is a prerequisite for attention. The *Conversation* layer models the separate dialogue episodes that go on during an interaction. A conversation is a reified process entity, consisting of a number of sub-fields. Each of these fields may be different for different conversations happening at the same time. The *participants* may be active speakers, addressees, or overhearers [15]. The *turn* indicates the participant with the right to communicate (using the primary channel). The *initiative* indicates the participant who is controlling the direction of the conversation. The *grounding* component of a conversation tracks how information is added to the common ground of the participants. The conversation structure also includes a *topic* that governs relevance, and *rhetorical* connections between individual content units. Once material is grounded, even as it still relates to the topic and rhetorical structure of an ongoing conversation, it is also added to the social fabric linking agents, which is not part of any individual conversation. This includes *social commitments* — both obligations to act or restrictions on action, as well as commitments to factual information. There is also a *negotiation* layer, modeling how agents come to agree on these commitments. We now turn to the layers in more detail.

The contact layer is modeled as a vector for all participants that the agent may interact with, each element indicating whether the participant is in contact in the media specified above. There are dimensions for whether someone is in contact to *send* or *receive* communications by this modality. The actions influencing this layer include **make-contact**, which could be established by turning on a radio or walking over to within eye contact or earshot, and **break-contact**, which could be established by walking out of hearing, turning out of view (or moving behind something), or turning off the radio. Contact is not generally realized verbally, although one might indicate a desire for contact, e.g., by shouting for someone to come over. An example of a make-contact action is shown at the beginning of our example in Figure 2, where the lieutenant drives up to the sergeant and walks out of the vehicle, to initiate contact (for the purpose of starting a conversation).

The attention layer is modeled by a similar vector to that of contact, though also including an entry for the agent itself, and attention is a one-way phenomenon, rather than having (potentially) distinct send and receive dimensions. The actions affecting this layer are divided into those that an agent performs concerning its own attention, and those related to the attention of other agents. **Give-attention** involves paying attention to some process, person, or object, as well as signalling this attention. This can be accomplished both verbally (e.g., saying "yes") or nonverbally (gazing at the object of attention). **Withdraw-attention** removes the current object from the attention entry of the agent. It can be implicit in giving attention to something else, or performed explicitly, by looking away in some cases (when this look does not serve some other purpose, such as planning a turn or indicating turn-taking in conversation). **Request-attention** signals to an agent that its attention is desired – changing the attentional state will also require a give-attention action by this agent. Request-attention can be signalled by a call on the radio, a shout, or using an agent's name, but also by gestures, such as raising an arm or waving. A **release-attention** act indicates that attention is no longer required. It occurs by default when a process or action that is the object of attention ends. It can also be explicit, in the form of a dismissal, or gesture indicating lack of attention (looking away). Attention of the released agent may still persist, however, until withdrawn or given

to something else. **Direct-attention** signals that attention should be given to another object or event, rather than the signaller. This can be accomplished with a deictic gesture, or with an utterance such as "look up!"

Conversation is one common reason for desiring attention. In this case, attention will be assumed (unless explicitly withdrawn) for the duration of the conversation. There are also explicit indicators of conversational openings and closings [40, 24]. Conversations are often opened with verbal greetings, but nonverbal actions can be very important as well. Kendon found a variety of nonverbal actions involved in the initiation of conversation [24]. The interaction typically starts with a "sighting" before the orientation and approach. Individuals who do not know each other well and have no special reason to greet each other will "catch the eye" by gazing longer than normal. Next, a "distance salutation" involves definite eye contact. This is followed by an approach, which typically involves gaze avoidance. Finally, a "close salutation" involves resumed eye contact. The BodyChat system [12] was the first to model these acts in animated agents. Conversational openings and closings are very formalized in the military radio modality, e.g., saying "out" to close a conversation. Conversation-maintenance actions such as **open**, **continue** and **close** can be performed either explicitly or implicitly. For instance, when teammates have a conversation on an urgent task, after solution of the task, they may directly move to other tasks, without explicitly closing or continuing the conversation. There are also actions for maintaining and changing the presence and status of participants. An example is in Figure 2, the way the medic is elevated from being an overhearer to active participant status.

Turn-taking actions model shifts in the turn holder. Most can be realized verbally, nonverbally, or through a combination of the two. **Take-turn** is an attempt to take the turn by starting to speak. **Request-turn** (e.g., signalled by various speech preparation signals such as opening the mouth or raising the hands into gesture space, or by avoiding a speaker's gaze at phrase boundaries) is an attempt to request the turn without forcibly taking it [3]. **Release-turn** (e.g., signalled by an intonational boundary tone, removal of the hands from gesture space, or a sustained gaze at a listener at the end of an utterance) is an attempt to offer the turn to the listener [3, 19]. **Hold-turn** (e.g., signalled verbally by a filled pause, or nonverbally by gaze aversion at phrase boundaries or hands in gesture space) is an attempt to keep the turn at a point where a listener might otherwise take it [3, 19]. These four turn-taking acts have been modeled in embodied conversational agents since the earliest systems [9]. In multi-party dialogue, there is one more act: **assign-turn**, which can be used to explicitly select the next speaker [3]. **Assign-turn** can be signalled verbally by a vocative expression or nonverbally by a speaker's gaze at the end of an utterance. Among embodied conversational agents, only Steve includes this act.

We use *initiative* to model how the agent should plan contributions. Even though the turn may shift from speaker to speaker, in many parts of a dialogue a single agent controls the flow of the contributions while others only respond to the initiative of that agent. For some *mixed-initiative* dialogues, initiative may shift from one participant to another. Initiative is sometimes pre-allocated by role for specific tasks. Otherwise, it starts with the agent who opened the conversation, and can be managed with **take-initiative**, **hold-initiative**, and **release-initiative** actions. These acts can often be signalled by performing (only) appropriate core-speech acts in context, e.g., as proposed by [47, 46]. We are not currently considering nonverbal signals of initiative.

Following [16, 43, 31], we treat *grounding* as occurring in discrete bundles of dialogue-introduced information that are added to the common ground together. Common Ground Units (CGUs) are modeled as information stores with state, which can be updated by the performance of the *grounding acts* from [45, 43]: **initiate**, **continue**, **repair**, **request-repair**, **display**, **acknowledge**, **request-acknowledge**, and **cancel**. See previous work for details of all but display, which is an explicit signal of what was understood (e.g., repeating a word, performing an action), leaving it to the original speaker to decide if this act functions as a repair, request-repair, or acknowledge [22]. Embodied conversational agents typically include nonverbal actions for request-acknowledge (e.g., gaze at addressee at grammatical pauses) and acknowledge (e.g., gaze at speaker and nod), and some include request-repair (e.g., when speech recognition fails, Peedy the parrot cups his wing to his ear and says "Huh?" [5]).

Topic actions include **start-topic** and **end-topic**. Topic structure can also be complex, when new topics are started before old ones have completed. Topic shifts of various sorts can be signalled explicitly with cue phrases (e.g., "now," "anyway"), but also with nonverbal cues. Head movements can signal topic shifts; a related sequence of utterances by a speaker typically uses the same basic head movement, and the speaker will often employ a new type of head movement to mark the start of a new topic [23]. Topic shifts are also frequently accompanied by shifts in the speaker's body posture [23].

Our negotiation model, broadly similar to models such as [4, 41], concerns the social process of deciding on future courses of action to perform. Each task that an agent might consider is represented as having certain features such as preconditions and effects. When tasks are bundled into a plan, more information is also represented such as causal links, and individual goals and intentions about the task and beliefs about its feasibility. Team plans also have specific responsibility and authority roles. Both of the agents that fill these roles[2] for a particular task must sufficiently endorse the task in order to have confidence it will be done (although we allow for the possibility that an agent may take initiative and perform an unauthorized task). Negotiation stances include **committed, endorsed, mentioned, unmentioned, disparaged**, and **rejected**.

Our model also includes layers for rhetorical structure of topical elements within a conversation [28], and obligations and commitments [44, 1]. We will, however, skip detailed discussion of these layers for the present, for matters of space, and also because they have received a fair amount of attention in the previous literature.

5. IMPLEMENTATION AND PLANS

We are implementing the multi-modal dialogue model described above and using it to control virtual agents engaged

[2]It is possible for an agent to hold both roles, in which case this agent may autonomously decide without negotiation to do it.

in the peacekeeping scenario described in section 2. Currently, we have three characters (the sergeant, medic, and mother) using this model to communicate with a human lieutenant as well as each other. These agents accept speech input from a human, and produce both speech and gestural output. In addition, the sergeant and lieutenant communicate with other (simple) agents to carry out tasks such as securing the area, or sending a medevac helicopter. The speech vocabulary includes a few hundred words, combined productively into thousands of distinct phrases, and the agent's internal domain model consists of about 20 entities, several locations, and over 90 distinct world conditions and 35 distinct actions (many tied together in a hierarchical task model). The agent characters can

- move and gesture in the virtual world
- answer questions about states of the world and events that have happened or are happening
- suggest appropriate future courses of action
- respond to orders or requests to act by acting or negotiating, which may involve rejecting the order or counterproposing a more opportune alternative
- engage in a clarification sub-dialogue to elaborate missing information from an interpreted utterance
- issue orders to subordinates in service of a team task, and open a conversation with others when it needs to talk to them to accomplish a task
- attend (including visual gaze) to multiple concerns, including monitoring multiple tasks and engaging in multiple conversations.

The agents are built on top of the Steve architecture, which in turn is built on top of Soar [25, 32]. We have retained Steve's basic architecture, as well as many of its core modules, such as its planner. However, we have replaced and extended most of Steve's language processing modules. Previously, the output of speech recognition was mapped directly into a simple semantic representation. Now, a module external to Steve uses more sophisticated natural language understanding techniques to convert the output of speech recognition into a richer semantic representation, and a new module within Steve uses Steve's knowledge of the virtual world to perform reference resolution and disambiguation over this representation. On the generation side, we have replaced Steve's template-driven approach with a more sophisticated natural language generator [20]. Additional code annotates the output of the generator with appropriate body movements, and the speech and body movements are synchronized using BEAT [13]. Finally, Steve's previous dialogue manager has been replaced with the dialogue model described in this paper. In the rest of this section, we briefly outline the current state of implementation of the model presented in section 4.

We have only partially implemented the two lowest layers of the dialogue model. In the contact layer, domain-specific inference rules directly update the information state from the agent's beliefs (e.g., about the locations of other characters), and the agents use the information state to decide how to contact other characters. However, they do not yet actively try to change the state (e.g., by moving into or out of contact). In the attention layer, only (visual) **give attention** is currently modeled, on the output side.

At the conversation level, we have the ability to model and engage in multiple conversations. A conversation is represented as a data structure with the following fields:

- participants (active and overhearers)
- a turn-holder
- an initiative-holder
- a dialogue history of utterances and actions that contributed to this conversation
- a grounding structure consisting of a set of CGUs
- depending on the type of conversation, a "style" and "purpose" marker. Conversations with an urgent-task style will end as soon as the participants realize the task is complete, without explicit verbal closings. The agent has the ability to begin such tasks by calling for the attention of the agent(s) needed for them. Future work includes modeling of radio conversations, explicit closings, conversations with multiple topics, and attention management in the face of multiple open conversations.

Turn-taking actions are used to set the turn-holder for each conversation. Currently implemented are recognition rules for **take-turn** and **assign-turn**, and modeling of **take-turn**, **assign-turn**, **release-turn** and **hold-turn**. The current agent is polite and will not speak out of turn, unless there is an urgent communication needed.

Initiative is currently only vestigially modeled. It starts with the conversation initiator. A **take-initiative** act is recognized when someone asks a (non-clarification) question, or when someone provides non-reactive information. We plan to model the rest of the acts as well as investigating the relation of initiative to social and role status, task urgency, and emotional coping [29].

Grounding is modeled as a set of Common Ground Units (CGUs) each of which contains a set of core speech acts (e.g., **order**, **assert**, **request** [45]) and a representation of the effects on the social state (obligations, commitments, negotiation stances). Each CGU also has a state marker and an initiator flag (indicating the initiator of the grounding unit, not the conversation), following the theory in [43]. CGUs can be active or cancelled (in which case there is some memory of speech events, but no effects to the social state). The following grounding acts are currently recognized and generated in the system: **initiate**, **continue**, **repair**, **acknowledge**, **request-repair**, and **cancel**. These acts and the ability to maintain multiple CGUs allows for performing multi-functional utterances (e.g., acknowledging a prior CGU while initiating a new one, acknowledging multiple utterances together, and repairing and cancelling of problematic sequences).

The topic layer is currently only vestigially implemented, for urgent conversations. Rhetorical structure is not currently explicitly modeled. However, our generation content planning does have rhetorically motivated behaviors, such as temporal sequencing of counterproposals, and multi-utterance answers to produce emotional effect.

We maintain a social state representation, consisting of commitments to propositions, obligations to act, as well as

the social relationships of other agents (e.g., subordinates, superiors, and teammates). Obligations provide one of the most powerful motivating factors for engaging in dialogue. A typical example is the motivation to answer a question [44]. This is still a motivation, however, not a hard-wired system reaction, and the answer act is interruptible, both by external events and by further dialogue from the human. If the motivation still exists, the agent will eventually resume the dialogue behavior, but if the motivation is removed, the agent will not.

The negotiation model is mostly implemented, although sophisticated moves such as providing and undercutting reasons are not currently dealt with. Also, we are not currently recognizing specific negotiation dialogue acts, but rather model introduction and elimination of negotiation stances as effects of forward and backward acts at the social commitment level. Acts having negotiation-level effects include **order**, **accept**, **reject**, **suggest**, and **counterpropose**.

While the model presented in Section 4 seems to be appropriate to handle many of the complexities found in both constructed dialogues such as Figure 2 and naturally occuring dialogues, we have plans for more formal evaluation along a couple of dimensions. This involves both annotating spontaneous and Wizard of Oz dialogues in our and similar domains, as well as continued implementation and evaluation of the adequacy of agents using this model to participate in complex conversational interactions.

6. SUMMARY

The dialogue model presented in this paper integrates and extends prior work on spoken dialogue and embodied conversational agents, providing a broad foundation for multiparty dialogues in immersive virtual worlds. Collectively, the different layers of our model cover many of the phenomena that arise in such dialogues. Our model separates the abstract dialogue acts at each layer, along with their effect on the dialogue state, from the various utterances and nonverbal signals that can realize them, supporting a more modular implementation that can be easily extended to include new communicative modalities. While not yet complete, our current implementation includes the core portions of our dialogue model, integrated into an embodied agent with a wide range of capabilities. We are currently completing the implementation of the core dialogue manager, as well as extending the set of utterances and nonverbal signals that the agent can map into the abstract dialogue acts on both the input and output sides.

Acknowledgements

We would like to thank members of the MRE project for interesting discussions about the complexities of interaction in this scenario, as well as our many colleagues involved in building the agent, artificial world, and simulation environment. Larry Tuch wrote the script with creative input from Richard Lindheim and technical input on Army procedures from Elke Hutto and General Pat O'Neal. The work described in this paper was supported in part by the U.S. Army Research Office under contract #DAAD19-99-C-0046. The content of this article does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

7. REFERENCES

[1] J. Allwood. Obligations and options in dialogue. *Think Quarterly*, 3:9–18, 1994.

[2] J. Allwood, J. Nivre, and E. Ahlsen. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9, 1992.

[3] M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge, 1976.

[4] M. Baker. A model for negotiation in teaching-learning dialogues. *Journal of Artificial Intelligence in Education*, 5(2):199–254, 1994.

[5] G. Ball, D. Ling, D. Kurlander, J. Miller, D. Pugh, T. Skelly, A. Stankosky, D. Thiel, M. van Dantzich, and T. Wax. Lifelike computer characters: the persona project at microsoft. In J. Bradshaw, editor, *Software Agents*. AAAI/MIT Press, Menlo Park, CA, 1997.

[6] Boston Dynamics. *PeopleShop 1.4 User Manual*, 2000.

[7] H. Bunt. Interaction management functions and context representation requirements. In *Proceedings of the Twente Workshop on Language Technology: Dialogue Management in Natural Language Systems (TWLT 11)*, pages 187–198, 1996.

[8] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan. Conversation as a system framework: Designing embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.

[9] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proceedings of ACM SIGGRAPH '94*, pages 413–420, Reading, MA, 1994. Addison-Wesley.

[10] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.

[11] J. Cassell and K. R. Thórisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13:519–538, 1999.

[12] J. Cassell and H. Vilhjálmsson. Fully embodied conversational avatars: Making communicative behaviors autonomous. *Autonomous Agents and Multi-Agent Systems*, 2:45–64, 1999.

[13] J. Cassell, H. Vilhjálmsson, and T. Bickmore. Beat: the behavior expression animation toolkit. In *Proceedings of ACM SIGGRAPH*, pages 477–486, New York, 2001. ACM Press.

[14] H. H. Clark. Managing problems in speaking. *Speech Communication*, 15:243 – 250, 1994.

[15] H. H. Clark. *Using Language*. Cambridge University Press, Cambridge, England, 1996.

[16] H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive Science*, 13:259–294, 1989.

[17] P. Dillenbourg, D. Traum, and D. Schneider. Grounding in multi-modal task-oriented collaboration. In *Proceedings of the European Conference on AI in Education*, 1996.

[18] Discourse Resource Initiative. Standards for dialogue coding in natural language processing. Report no. 167,

Dagstuhl-Seminar, 1997.

[19] S. Duncan, Jr. Some signals and rules for taking speaking turns in conversations. In S. Weitz, editor, *Nonverbal Communication*, pages 298–311. Oxford University Press, 1974.

[20] M. Fleischman and E. Hovy. Emotional variation in speech-based natural language generation. In Proceedings of The Second International Natural Language Generation Conference (INLG'02), June 2002.

[21] W. L. Johnson, J. W. Rickel, and J. C. Lester. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11:47–78, 2000.

[22] Y. Katagiri and A. Shimojima. Display acts in grounding negotiations. In *Proceedings of Gotalog 2000, the 4th Workshop on the Semantics and Pragmatics of Dialogue*, pages 195–198, 2000.

[23] A. Kendon. Some relationships between body motion and speech. In A. Siegman and B. Pope, editors, *Studies in Dyadic Communication*, pages 177–210. Pergamon Press, New York, 1972.

[24] A. Kendon. A description of some human greetings. In R. Michael and J. Crook, editors, *Comparative Ecology and Behavior of Primates*, pages 591–668. Academic Press, New York, 1973.

[25] J. E. Laird, A. Newell, and P. S. Rosenbloom. Soar: An architecture for general intelligence. *Artificial Intelligence*, 33(1):1–64, 1987.

[26] S. Larsson and D. Traum. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6:323–340, September 2000.

[27] J. C. Lester, J. L. Voerman, S. G. Towns, and C. B. Callaway. Deictic believability: Coordinating gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence*, 13:383–414, 1999.

[28] W. C. Mann and S. A. Thompson. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, USC, Information Sciences Institute, June 1987.

[29] S. Marsella and J. Gratch. A step towards irrationality: using emotion to change belief. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, June 2002.

[30] C. Matheson, M. Poesio, and D. Traum. Modelling grounding and discourse obligations using update rules. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics*, 2000.

[31] C. H. Nakatani and D. R. Traum. Coding discourse structure in dialogue (version 1.0). Technical Report UMIACS-TR-99-03, University of Maryland, 1999.

[32] A. Newell. *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA, 1990.

[33] D. Novick. *Control of Mixed-Initiative Discourse Through Meta-Locutionary Acts: A Computational Model*. PhD thesis, University of Oregon, 1988. Also available as U. Oregon Computer and Information Science Tech Report CIS-TR-88-18.

[34] C. Pelachaud, N. I. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1), 1996.

[35] M. Poesio and D. R. Traum. Towards an axiomatization of dialogue acts. In *Proceedings of Twendial'98, 13th Twente Workshop on Language Technology: Formal Semantics and Pragmatics of Dialogue*, 1998.

[36] J. Rickel, J. Gratch, R. Hill, S. Marsella, and W. Swartout. Steve goes to Bosnia: Towards a new generation of virtual humans for interactive experiences. In *AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment*, 2001.

[37] J. Rickel and W. L. Johnson. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence*, 13:343–382, 1999.

[38] J. Rickel and W. L. Johnson. Virtual humans for team training in virtual reality. In *Proceedings of the Ninth International Conference on Artificial Intelligence in Education*, pages 578–585. IOS Press, 1999.

[39] J. Rickel and W. L. Johnson. Task-oriented collaboration with embodied agents in virtual worlds. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.

[40] E. A. Schegloff and H. Sacks. Opening up closings. *Semiotica*, 7:289–327, 1973.

[41] C. L. Sidner. An artificial discourse language for collaborative negotiation. In *Proceedings of the Fourteenth National Conference of the American Association for Artificial Intelligence (AAAI-94)*, pages 814–819, 1994.

[42] W. Swartout, R. Hill, J. Gratch, W. Johnson, C. Kyriakakis, K. Labore, R. Lindheim, S. Marsella, D. Miraglia, B. Moore, J. Morie, J. Rickel, M. Thiebaux, L. Tuch, R. Whitney, and J. Douglas. Toward the holodeck: Integrating graphics, sound, character and story. In *Proceedings of 5th International Conference on Autonomous Agents*, 2001.

[43] D. R. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, Department of Computer Science, University of Rochester, 1994. Also available as TR 545, Department of Computer Science, University of Rochester.

[44] D. R. Traum and J. F. Allen. Discourse obligations in dialogue processing. In *Proceedings of the 32^{nd} Annual Meeting of the Association for Computational Linguistics*, pages 1–8, 1994.

[45] D. R. Traum and E. A. Hinkelman. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3):575–599, 1992.

[46] M. A. Walker and S. Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings ACL-90*, pages 70–78, 1990.

[47] S. Whittaker and P. Stenton. Cues and control in expert-client dialogues. In *Proceedings ACL-88*, pages 123–130, 1988.