Toward Virtual Humans

**William Swartout[1], Jonathan Gratch[1], Randall Hill[1], Eduard Hovy[2],
Stacy Marsella[2], Jeff Rickel, David Traum[1]**

**[1]USC Institute for Creative Technologies
13274 Fiji Way
Marina del Rey, CA 90292**

**[2]USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 902**

swartout@ict.usc.edu, gratch@ict.usc.edu, hill@ict.usc.edu, hovy@isi.edu, marsella@isi.edu, traum@ict.usc.edu

Abstract

This paper describes the virtual humans developed as part of the Mission Rehearsal Exercise project, a virtual reality based training system. This project is an ambitious exercise in integration, both in the sense of integrating technology with entertainment industry content, but also in that we have joined a number of component technologies that have not been integrated before. This integration has not only raised new research issues, but it has also suggested some new approaches to difficult problems. We describe the key capabilities of the virtual humans, including task representation and reasoning, natural language dialogue, and emotion reasoning, and show how these capabilities are integrated to provide more human-level intelligence than would otherwise be possible.

Introduction

Achieving human-level intelligence in virtual characters requires a number of core capabilities, including planning, belief representation, communication ability, emotional reasoning, and most importantly, a way to integrate these capabilities. For many researchers, software integration is often regarded as a kind of necessary evil – something to make sure that all the research components of a large system fit together and interoperate properly – but not something that is likely to contribute new research insights or suggest new solutions. We have found, on the contrary, that the conventional wisdom about integration does not hold: as we describe in this paper, the integration process has raised new research issues and at the same time has suggested new approaches to long-standing issues. We begin with a brief description of the background behind our work in training and the approach we have taken to improving training. We then describe the technology components we have developed, the system architecture we use, and we conclude with some of the insights we have gained from the integration process.

Virtual Humans for Training

Virtual humans are software artifacts that look like, act like and interact with humans but exist in virtual environments. We have been exploring the use of virtual humans to create social training environments, environments where a learner can explore stressful social situations in the safety of a virtual world.

For example, we designed the Mission Rehearsal Exercise (MRE) system to demonstrate the use of virtual human technology to teach leadership skills in high–stakes social situations. MRE places the trainee in an environment populated with virtual humans. The training scenario we are currently using is situated in a small town in Bosnia. It opens with a lieutenant (the trainee) in his Humvee. Over the radio, he gets orders to proceed to a rendezvous point to meet up with his soldiers to plan a mission to assist in quelling a civil disturbance. When he arrives at the rendezvous point, he discovers a surprise (see Figure 1). One of his platoon's Humvees has been involved in an accident with a civilian car. There's a small boy on the ground with serious injuries, a frantic mother, and a crowd is starting to form. A TV camera crew shows up and starts taping. What should the lieutenant do? Should he stop and render aid? Or should he continue on with his mission? Depending on decisions he makes, different outcomes will occur.

Our virtual humans build on prior work in the areas of embodied conversational agents (Cassell, Sullivan, Prevost, & Churchill, 2000) and animated pedagogical agents (Johnson, Rickel, & Lester, 2000), but they integrate a broader set of capabilities than any prior work. For the types of training scenarios we are targeting, the virtual humans must integrate three broad influences on their behavior: they must perceive and

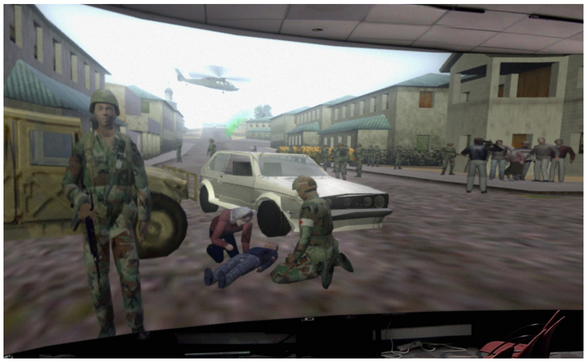


Figure 1: The Mission Rehearsal Exercise System, showing, from the left, the platoon sergeant, the injured boy and his mother, a medic, and a crowd.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **OCT 2004**	2. REPORT TYPE	3. DATES COVERED **00-10-2004 to 00-10-2004**

4. TITLE AND SUBTITLE
Toward Virtual Humans

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of California,Institute for Creative Technologies,13274 Fiji Way,Marina del Rey,CA,90292

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**8**	

act in a 3D virtual world, they must engage in face-to-face spoken dialogues with people and other virtual humans in such worlds, and they must exhibit human-like emotions. Classic work on virtual humans in the computer graphics community focused on perception and action in 3D worlds (Badler, Phillips, & Webber, 1993; Thalmann, 1993), but largely ignored dialogue and emotions. Several systems have carefully modeled the interplay between speech and nonverbal behavior in face-to-face dialogue (Cassell, et al, 2000; Pelachaud, Badler, & Steedman, 1996) but these virtual humans did not include emotions and could not participate in physical tasks in 3D worlds. Some work has begun to explore the integration of conversational capabilities with emotions (Lester, et al 2000; Marsella, Johnson, & LaBore, 2000; Poggi & Pelachaud, 2000), but still does not address physical tasks in 3D worlds. Likewise, prior work on Steve addressed the issues of integrating face-to-face dialogue with collaboration on physical tasks in a 3D virtual world (Rickel & Johnson, 2000), but Steve did not include emotions and had far less sophisticated dialogue capabilities than our current virtual humans. The tight integration of all these capabilities is one of the most novel aspects of our current work.

The virtual humans, which include the sergeant, medic, and mother in the scenario described in the previous section, are implemented in Soar, a general architecture for building intelligent agents (Newell, 1990) and build on the earlier Steve system. As such, their behavior is not scripted; rather, it is driven by a set of general, domain-independent capabilities discussed below. The virtual humans perceive events in the simulation, reason about the tasks they are performing, and they control the bodies and faces of the characters to which they have been assigned. They send messages to one another, to the character bodies, and to the audio system via the Communications Bus shown in Figure 2.

Architecture

In order for virtual humans to collaborate with people and each other in scenarios like the peacekeeping mission, they must include a wide variety of capabilities, such as perception, planning, spoken dialogue, and emotions. Thus, we desired a flexible architecture for our virtual humans that would allow us to easily experiment with the connections between the individual components.

A blackboard architecture, in which individual components have access to the intermediate and final results of other components by default, provides such flexibility. The alternative, in which each module would explicitly pass specific information to other components, would require constant revision as we made progress understanding the interdependencies among components.

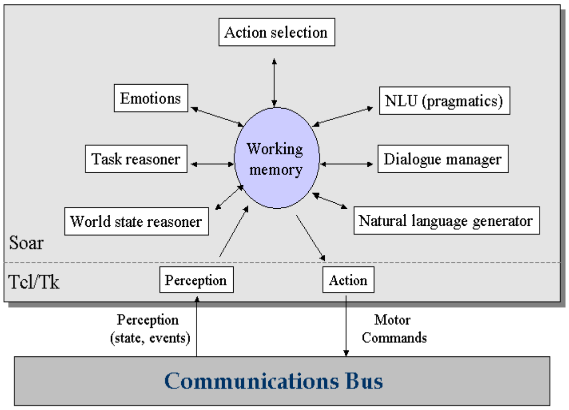


Figure 2: Virtual Human architecture

For our integrated architecture, we chose Soar, because it allows each component to be implemented with production rules that read from and write to a common working memory, which acts as the desired blackboard. Soar further breaks computation into a sequence of intermediate *operators* that are proposed in parallel but selected sequentially via an arbitration mechanism. This allows for tight interleaving of operators from individual components and flexible control over their priority.

All components of the virtual humans are implemented in Soar, with several exceptions: speech recognition, natural language understanding (syntactic and semantic analysis), synchronization of verbal and nonverbal components of output utterances, and speech synthesis. It was less practical to implement these four components in Soar because each was built on top of existing software that would have been difficult to reimplement. These modules also work more or less as pipe-lines, with well-defined inputs and outputs, and do not require as much detailed interaction as the core reasoning components implemented within Soar.

Task Representation and Reasoning

To collaborate with humans and other synthetic teammates, virtual humans need to understand how past events, present circumstances, and future possibilities impact team tasks and goals. For example, the platoon sergeant agent in Figure 1 must be able to brief the trainee on past events that led to the accident as well as how the victim’s current injuries impact the platoon’s future mission. More generally, agents must understand task goals and how to assess whether they are currently satisfied, the actions that can achieve them, how the team must coordinate the selection and execution of those actions, and how to adapt execution to unexpected events.

To provide this understanding, agents use domain-independent reasoning algorithms operating over a domain-specific declarative representation of team tasks. The representation incorporates elements of decision-theoretic plan representations (allowing agents to reason about the value and likelihood of future possibilities) with an explicit representation of beliefs and intentions (important for multi-agent reasoning). This representation is divided into explicit representations of past episodes, present state and future task-related information: The *causal history* maintains a sequence of past observed steps (including unexpected and non-task events) and interdependencies between past steps and present or future states (e.g., causal links). The *current world description* represents the current state of the world through a list of propositions. The *task description* includes of a set of possible future steps, each of which is either a primitive action (e.g., a physical or sensing action in the virtual world) or an abstract action which must itself be further decomposed. Abstract actions give tasks a hierarchical structure. Interdependencies are represented as a set of ordering constraints, causal links and threat relations.

In addition to understanding the structure of tasks, agents must understand the roles of each team member. Each task step is associated with the team member that is responsible for performing it as well as a possibly different agent that has authority over its execution; that is, the teammate responsible for a task step cannot perform it until authorization is given by the specified teammate with authority (Traum, Rickel, Gratch & Marsella 2003). This is required to model the hierarchical organizational structure of some teams, such as in the military.

An agent's task model represents its understanding of the task in general, independent of the current scenario conditions (different agents may have different representations of the same task). Agents continually monitor the state of the virtual world via messages from the simulator that are filtered to reflect perceptual limitations (Rickel et al., 2002) and update their plans accordingly. The result of this planning algorithm specifies how the agent privately believes that the team can collectively complete the task, with some causal links specifying the interdependencies among team members' actions.

A key aspect of collaborative planning is negotiating about alternative ways to achieve team goals (Traum, Rickel, Gratch & Marsella, 2003). To support such negotiation, the decision-theoretic planner can reason about alternative, mutually exclusive courses of action (recipes) for achieving tasks, their likelihood, and the utility of certain consequences, allowing the system to assess the relative strengths and weaknesses of different alternatives. These courses of action are self-contained hierarchical tasks in the sense defined above, and subject to the same dynamic task reasoning. For example, one might evacuate someone to a hospital by using either a medevac helicopter or an ambulance. Depending on the circumstances, only one option might be possible (e.g., the medevac may be unavailable or the injuries may be too severe for an ambulance), but if both are valid options, they must be ranked through some reasoned analysis of their relative costs and benefits.

Natural Language Dialogue

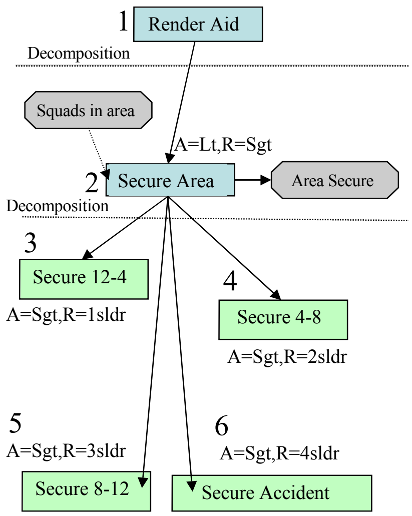
In many ways, our natural language processing components and architecture mirror fairly traditional dialogue systems. There is a speech recognizer, semantic parser, dialogue manager, NL generator, and speech synthesizer. However, the challenges of the MRE project, including integration within an immersive story environment as well as with the other virtual human components required innovations in most areas. Here we briefly describe the natural language processing components and capabilities; we will return later to some of the specific innovations motivated by this integration.

The Speech recognizer was built using Sonic (Pellom, 2001), with a domain specific n-gram language model and with locally trained acoustic models (Wang & Narayanan, 2002). Output is currently the single best interpretation, as well as indications of when the user starts and stops speaking, to manage gaze control and turn-taking behavior of agents.

Speech recognition output is processed by the semantic parser module, which produces a semantic representation of the utterances. The parser uses a hybrid between finite-state transducers and statistical processing to produce a best-guess at semantic information from the input word stream (Feng 2003). In cases in which imperfect input is given, it will robustly produce representations which may possibly be incomplete or partially incorrect. The module will provide addressee information (if vocatives were present), sentence mood, and semantic information corresponding to states and actions related to the task model (Traum, 2003).

The Soar-module for each agent receives the output of the speech recognizer and semantic parser. This information is then matched against the agent's internal representation of the context, including the actions and states in the task model, current expectations, and focus to determine a set of candidate interpretations. Some of these interpretations may be underspecified, due to impoverished input, or overspecified in cases of incorrect input (either an out of domain utterance by the user, or an error in the speech recognizer or semantic parser). In some cases, underspecified elements can be filled in with reference to the agent's knowledge; if not, the representation is left underspecified and processing continues. The dialogue component of the Soar agent also produces a set of dialogue act interpretations of the utterance. Some of these are traditional speech acts (e.g., assert, request, info-request) with content being the semantic interpretation, while others represent other levels of action that have been performed, such as turn-taking, grounding, and negotiation (Traum & Rickel, 2002).

Dialogue management follows the approach of the TRINDI Project (Larsson & Traum, 2000), and specifically the EDIS system (Matheson, Poesio, & Traum, 2000). Dialogue acts are used to update an *Information State* that is also used as context for other aspects of agent reasoning (Traum & Rickel, 2002). Decisions of how to act in dialogue are tightly coupled with other action selection decisions in the agent. The agent can choose to speak, choose to listen, choose to act related to a task, etc. Aspects of the information state provide motivations to speak, including answering questions, negotiating with respect to a request or order, giving feedback of understanding (acknowledgements, repairs, and repair requests), and making suggestions and issuing orders, when appropriate according to the task model.



```
Focus=1
Lt: U11“secure the area ”
Commited (lt,2), 2 authorized,  Obl(sgt,U11)
Sgt: U12 “yes sir”
Committed(sgt,2), Push(2,focus)
Goal7:Announce(2,{1sldr,2sldr,3sldr,4sldr})
Goal8: Start-conversation(sgt, ,{1sldr,2sldr,   …},2)
Goal8 -> Sgt: U21  “Squad leaders listen up! ”
Goal7 -> U22  “I want 360 degree security  ”
Push(3, focus)
Goal9:authorize 3
Goal9 -> u23 “1st squad take 12-4 ”
Committed(sgt,3), 3 authorized
Pop(3), Push(4)
Goal10: authorize 4
Goal10 -> u24 “2nd squad take 4-8  ”
Committed(sgt,4), 4 authorized
Pop(4)
…
A10: Squads move
A10: grounds U21-U26, …
      ends conversation about 2, realizes 2
```

Figure 3: Sample task model and dialogue interaction

Once a decision is made to speak, there are several phases involved in the language production process, including *content selection, sentence planning*, and *realization*. The final sentence is then augmented with communicative gestures and sent to the synthesizer and rendering modules to produce the speech. Meanwhile, messages are sent to other agents, letting them know what the agent is saying (Fleischman & Hovy, 2002). The speech synthesizer uses Festival and Festvox, with locally developed unit-selection limited-domain voices to provide the emotional expressiveness needed to maintain immersiveness (Johnson et al., 2002).

Figure 3 shows a brief example of how dialogue behavior is integrated with task reasoning. The left side of the figure shows a small fragment of the task model: part of the "Render aid" task involves securing the assembly area, which requires that the squads are in the area; it has a decomposition involving actions of various squads, and has the effect that the area is secure. The figure also shows which agents are responsible (R) for seeing that an action is performed (either doing it themselves or acting as team leader making sure the subtasks are carried out), and which agents have authority (A) to have the action performed. With reference to this piece of the task model, consider the dialogue fragment on the right. Initially the focus is on the render aid task. When the lieutenant issues the command to secure the area (utterance U11), the sergeant recognizes the command as referring to a subaction of Render Aid in the current task model (Task 2). As a direct effect of the lieutenant issuing a command to perform this task, the lieutenant has committed himself to the task, the sergeant has an obligation to perform the task, and the task becomes authorized. Because the sergeant already agrees that this is an appropriate next step, he is able to accept it with utterance U12, which also commits him to perform the action. The sergeant then pushes this task into his task model focus and begins execution. In this case, because it is a team task requiring actions of other teammates, the sergeant, as team leader, must announce the task to the other team members. Thus, the system forms a communicative goal to make this announcement. Before the sergeant can issue this announcement, he must make sure he has the squad leaders' attention and has them engaged in conversation. He forms a goal to open a new conversation so that he can produce the announcement. Then his focus can turn to the individual tasks for each squad leader. As each one enters the sergeant's focus, he issues the command that commits the sergeant and authorizes the troops to carry it out. When the sergeant observes the troops move into action, he can infer that they have understood his order and adopted his plan. When the task completes, the conversation between sergeant and squad leaders finishes and the sergeant turns his attention to other matters.

Emotion

As our agents attempt to realistically model the behavior of humans in high-stress scenarios, it is important to model the role emotion plays in influencing decision-making and behavior. Our work on modeling emotion is motivated by *appraisal theory*, a psychological theory of emotion that emphasizes the relationship between emotion and cognition (Lazarus, 1991). The theory posits two basic processes: *Appraisal* generates emotion by assessing the person-environment relationship (did an event facilitate or inhibit the agent's goals; who deserves blame or credit). *Coping* is the process of dealing with emotion, either by acting externally on the world (problem-focused coping), or by acting internally to change beliefs or attention (emotion-focused coping). Coping and appraisal interact and unfold over time, modeling the temporal character of emotion noted by several emotion researchers (Lazarus, 1991; Scherer, 1984): an agent may "feel" distress for an event (appraisal), which motivates the shifting of blame (coping), which leads to anger (re-appraisal).

In re-casting this theory as a computational model, we have tied appraisals and coping to the agent's task knowledge and reasoning (Gratch & Marsella, 2004). This representation has several advantages for modeling emotion. It makes a clean separation between domain-specific knowledge, it acts as a blackboard architecture, simplifying communication between appraisal and coping to other mechanisms (like planning) that operate on the task knowledge, it facilitates reasoning about blame and indirect consequences of action (e.g., a threat to a sub-goal might be distressing, not because the subgoal is intrinsically important, but because it facilitates a larger goal), and it provides a uniform representation of past and future actions.

Our approach to appraisal assesses the agent-environment relationship via features of this explicit task representation (Gratch, 2000). Speaking loosely, we treat appraisal as a set of feature detectors that map features of this representation into appraisal variables that characterize the consequences of an event from the agent's perspective. These variables include the desirability of those consequences, the likelihood of them occurring, who deserves credit or blame and a measure of the agent's ability to alter those consequences. The result is one or more *appraisal frames* that characterize the agent's emotional reactions to an event.

Our computational model of coping (Marsella & Gratch, 2002) similarly exploits the task representation to uncover which features led to the appraised emotion, and what potential there may be for altering these features. In essence, coping is the inverse of appraisal. To discharge a strong emotion about some situation, one obvious strategy is to change one or more of the factors that contributed to the emotion. Coping operates on the same representations as the appraisals, the agent's beliefs, goals and plans, but in reverse, seeking to make a change, directly or indirectly, that would have the desired impact on appraisal. Coping could impact the agent's beliefs about the situation, such as the importance of a threatened goal, the likelihood of the threat, responsibility for the threat, etc. Further, the agent might form intentions to change external factors, for example, by performing some action that removes the threat. Indeed, our coping strategies can involve a combination of such approaches. This mirrors how coping processes are understood to operate in human behavior whereby people may employ a mix of problem-focused coping and emotion-focused coping to deal with stress.

Action and Body Movements

Internally, the virtual humans are perceiving events, understanding utterances, updating their beliefs, formulating and revising plans, generating emotional appraisals, and choosing actions. Agents manifest the rich dynamics of their cognitive and emotional inner state through external behavior using the same verbal and nonverbal cues that people use to understand one another and these behaviors must be seamlessly integrated across modality and across time.

Here we summarize the model discussed in (Marsella, Gratch, & Rickel, 2003), which drives gaze, facial expressions, and body gestures based on features of the agent's dynamic cognitive state. Gaze indicates a character's focus of attention and is synchronized to the character's inner thoughts. For example, task-related behaviors (e.g., monitoring for an expected effect or action) trigger a corresponding gaze shift, and gaze during social interactions is driven by the dialogue state and the state of the virtual human's own processing (e.g., gaze at an interlocutor who is speaking, gaze aversion during utterance planning to hold the turn). Facial expressions both convey emotion and augment verbal communication. In humans, these behaviors can be used intentionally by an individual to inform or deceive but can also unintentionally reveal information about the individual's mental state and our work integrates these aspects: by tying some expressive behavior to emotional appraisal we reveal "true" mental state, whereas tying other behaviors to coping strategies, we inform intentional displays. Finally, a wide range of body movements emphasize and augment speech.

Our approach plans the utterance, annotates it with nonverbal behavior, then passes it to a text-to-speech system that schedules both the verbal and nonverbal behavior, using BEAT (Cassell, Vilhjálmsson, & Bickmore, 2001), although we augment this to express not only the syntactic, semantic and pragmatic structure of the utterance, emotional appraisal and coping information as well.

Putting it together: the value of integration

We have described the major technical components of the virtual humans. As we pointed out in the introduction, software integration is necessary to make sure that all the various pieces in a system work together properly, but one usually expects that the real research takes place in building the individual components. One doesn't expect to learn much from integration (except perhaps to find that some components don't interface properly). However, in integrating the Mission Rehearsal Exercise system, we have been surprised: we have uncovered new research issues and some new approaches to existing problems have been suggested. In this section we outline some of the things we learned as we brought all the pieces together.

Socially Embedded Multi-speaker Dialogue

The Bosnian scenario of MRE, with a cast of many characters occupying various roles in a rich social fabric, is quite different from the usual case of natural language dialogue with a single human and single computer system interacting. While some aspects of dialogue as social interaction had already been addressed in previous work (e.g, discourse obligations in (Traum & Allen, 1994)), many new issues must be confronted, such as:

- Is the intended addressee paying attention?
- Is he already engaged in conversation?
- How will hearers recognize the addressee?
- How are vocatives and gaze as well as context reasoning used to help this process?
- How are multiple, interleaved, conversations managed (e.g., talking face to face with one character while on the radio to another)?

These issues have implications for agents in both understanding and producing communications, and for representing the dialogue state. Furthermore, there are differences depending on whether the conversation is between virtual humans or between the human trainee and a virtual human, because more limited information is available in the second case.

We have begun to address these issues in several ways. First, the dialogue model has been extended so that who is being addressed is captured as well as the content to be conveyed. Second, we have introduced conventions for marking the start and termination of a conversation with an agent. A conversation begins by addressing the character either by name or by his role. For example the lieutenant might give the sergeant an order by saying: “Sergeant, send first squad to Celic!” Once a conversation has been started, it is assumed to continue until it is terminated, either by the purpose having been fulfilled (for a short task-specific conversation like securing the area), or by an explicit closing (e.g., "out" on the radio).

For conversations between the human trainee and the virtual humans we rely on these conventions to determine who is addressing whom. For conversations between virtual humans, the problem of determining who is being addressed is easier, because it is all represented internally. However, the virtual humans use the same reasoning methods when talking among themselves as they use for interacting with the trainee so their behavior is consistent. We feel this is an important constraint to achieve consistency in interface behavior (Traum & Rickel, 2002). We have also begun to make use of head-tracking data to determine who the trainee is looking at when he speaks.

The Pervasive Effect of Emotion

In humans, emotion has a broad effect on behavior. It affects how we speak, how we gesture, our posture, and even how we reason. And, of course, emotion is indispensable for creating good story and compelling characters. In integrating emotion into our virtual humans, we have found that we need to deal with a similarly broad range of issues. Models of emotion can both affect the behavior of other components of the virtual human, and they can provide additional knowledge that the system can use in reasoning. Below we give an example of each.

Emotionally Appropriate Natural Language Generation. A big challenge for Natural Language Generation in MRE is the generation of emotionally appropriate language, which expresses both the desired information and the desired emotional attitude towards that information. Each expressive variant casts an emotional shade on each representational item it contains (for example, the phrase governed by the verb “ram” as in “They rammed into us, sir” casts the subject in a negative and the object in a positive light). Prior work on the generation of variation expressions, such as (Bateman & Paris, 1989; Hovy, 1990), uses quite simplistic emotional models of the speaker and hearer. In general, these systems simply had to choose among a small set of phrases, and within the phrase from a small set of lexical fillers for certain positions of the phrase, where each alternative phrase and lexical item was pre-annotated with an affective value such as *good* or *bad*.

The presence in MRE of an emotion model provides a considerably finer-grain level of control, enabling principled realization decisions over a far more nuanced set of expressive alternatives. Given many representational items, a rich set of emotional values potentially holding for them, and numerous phrases, each with its own combination of positive and negative fields, the problem was to design a system that can reliably and quickly find the optimal phrasing without dropping content. To compute shades of connotation more accurately and quickly, we created a vector space in which we can represent the desired attitudes of the speaker (as specified by the emotion model) as well as the overall emotional value of each candidate expression (whether noun phrase or whole sentence). Using a standard Euclidean distance measure we can then determine which variant expression most closely matches the desired effect. See (Fleischman & Hovy, 2002) for details.

Using Emotion to Determine Linguistic Focus. In natural language, we often refer to things in imprecise ways. To correctly interpret such referents in a natural language utterance, one needs to understand what is in linguistic focus. Loosely speaking, one needs to understand what is the main subject of discussion. For example, when the lieutenant trainee arrives at the accident scene in the MRE scenario, he might ask the sergeant, “What happened here?” In principle many things have happened: the lieutenant just drove up, the soldiers assembled at the meeting point, an accident occurred, a crowd formed, and so forth. The sergeant could talk about any one of these and be factually correct, but not necessarily pragmatically appropriate. A number of heuristics have been developed to model linguistic focus. One such heuristic is based on the idea of recency. It holds that the entity that is in linguistic focus is whatever was most recently discussed, or occurred most recently. In this case, recency doesn’t work, since the Sergeant would sound quite silly if he responded: “Well, you just drove up, sir.” On the other hand, people are often focused most strongly on the things that upset them emotionally, which suggests an emotion-based heuristic for determining linguistic focus. Because we have modeled the sergeant’s emotions in MRE, the dialogue planning modules have access to the fact that he is upset about the accident can use that information to give the most appropriate answer: describing the accident and how it occurred.

Status and Evaluation

An initial version of the MRE system described in this paper has been implemented and applied to the peacekeeping training scenario described earlier. The system allows the trainee, playing the role of the lieutenant, to interact freely (through speech) with the three virtual humans (sergeant, medic, and mother). The trainee takes action in the virtual world through commands to the sergeant, who in turn commands the squads. Ultimately, the experience terminates with one of four possible endings, depending on the trainee's actions. However, unlike interactive narrative models based on an explicit branching structure, the system does not force the trainee through a predetermined sequence of decision points, each with a limited set of options; the trainee's interactions with the characters is unconstrained and limited only by the characters' understanding and capabilities.

The understanding and capabilities of the virtual humans is limited by the coverage of their spoken dialogue models and their models of the domain tasks. The sergeant's speech recognizer currently has a vocabulary of a few hundred words, with a grammar allowing recognition of 16000 distinct utterances. His natural language understanding module can currently produce semantic representation frames for all of these sentences as well as providing (sometimes partial) results for different or ill-formed input. His natural language generation module currently expresses all communicative goals formed by the dialog module, modulating some of them for affective appropriateness. His speech synthesis module currently has a vocabulary of over 1000 words. The sergeant's domain task knowledge, which is the most complex among all the virtual humans in the scenario, includes about 40 tasks, and about 150 properties of the world. While the tasks represent the full range of actions that the sergeant can understand and carry out, his ability to talk about these tasks and properties (e.g., answer questions and give advice) is broad, limited only by the coverage of the spoken dialogue modules as described above.

Despite its complexity, real-time performance of the system is good, although we are continuing to improve latencies. Given an utterance by the user, a virtual human typically responds within 3 seconds, including speech recognition, natural language understanding, updating dialogue and emotional states, choosing how to respond, natural language generation, planning the voice output and accompanying gestures and visemes, and finally producing the speech. As is typical of humans, the virtual humans are producing communicative behaviors throughout this time delay, including averting gaze from the user during the utterance planning phases to indicate that they are formulating a response (Kendon 1967).

We have tested the system with a variety of users acting as trainees, including subjects with with and without prior knowledge of the military domain. Not surprisingly, subjects with military knowledge were substantially more successful, since they understood the context and how to proceed. Initial evaluation results and metrics of dialogue interaction using military cadets are presented in (Traum, Robinson, and Stephan, 2004)

Human-level intelligence requires a number of core capabilities, including planning, belief representation, communication ability, emotional reasoning, and most importantly, a way to integrate these capabilities. The virtual humans in the MRE project represent a significant step along this path.

ACKNOWLEDGEMENTS

Jeff Rickel created much of the virtual human architecture, and contributed to most aspects of the project. Sadly, he died in July 2003. The authors would like to thank the many people who collaborated on the virtual humans and Mission rehearsal project for their hard work and enthusiasm that contributed greatly to this research. This paper was developed with funds of the United States Department of the Army under contract number DAAD 19-99-D-0046. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the United States Department of the Army.

REFERENCES

Badler, N. I., Phillips, C. B., & Webber, B. L. (1993). Simulating Humans. New York: Oxford University Press.

Bateman, J. A., & Paris, C. L. (1989). Phrasing a Text in Terms the User can Understand. Paper presented at the 11th International Joint Conference on Artificial Intelligence, Detroit, MI.

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., & Yan, H. (2000). Human conversation as a system framework: Designing embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), Embodied Con-versational Agents (pp. 29-63). Boston: MIT Press.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (Eds.). (2000). Embodied Confersational Agents. Cambridge, MA: MIT Press.

Cassell, J., Vilhjálmsson, H., & Bickmore, T. (2001). BEAT: The Behavior Expressive Animation Toolkit. Paper presented at the SIGGRAPH, Los Angeles, CA.

Fleischman, M., & Hovy, E. (2002). Emotional variation in speech-based natural language generation. Paper presented at the International Natural Language Generation Conference, Arden House, NY.

Gratch, J. (2000). Émile: marshalling passions in training and education. Paper presented at the Fourth International Conference on Intelligent Agents, Barcelona, Spain.

Gratch, J., & Marsella, S. (2004). A domain independent framework for modeling emotion. Journal of Cognitive Systems Research.

Hovy, E. H. (1990). Pragmatics and Natural Language Generation. Artificial Intelligence, 43(2), 153-198.

Johnson, W. L., Narayanan, S., Whitney, R., Das, R., Bulut, M., & LaBore, C. (2002). Limited Domain Synthesis of Expressive Military Speech for Animated Characters. Paper presented at the 7th International Conference on Spoken Language Processing, Denver, CO.

Johnson, W. L., Rickel, J., & Lester, J. C. (2000). Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. International Journal of AI in Education, 11, 47-78.

Larsson, S., & Traum, D. (2000). Information state and dialogue management in the TRINDI Dialogue Move Engine Toolkit. Natural Language Engineering, 6, 323-340.

Lazarus, R. (1991). Emotion and Adaptation. NY: Oxford University Press.

Lester, J. C., Towns, S. G., Callaway, C. B., Voerman, J. L., & FitzGerald, P. J. (2000). Deictic and Emotive Communication in Animated Pedagogical Agents. In J. Cassell, S. Prevost, J. Sullivan & E. Churchill (Eds.), Embodied Conversational Agents (pp. 123-154). MIT Press.

Marsella, S., & Gratch, J. (2002). A Step Toward Irrationality: Using Emotion to Change Belief. Paper presented at the First International Joint Conference on Autonomous Agents and Multiagent Systems, Bologna, Italy.

Marsella, S., Gratch, J., & Rickel, J. (2003). Expressive Behaviors for Virtual Worlds. In H. Prendinger & M. Ishizuka (Eds.), Life-like Characters Tools, Affective Functions and Applications: Springer-Verlag.

Marsella, S., Johnson, W. L., & LaBore, C. (2000). Interactive Pedagogical Drama. Paper presented at the Fourth International Conference on Autonomous Agents, Montreal, Canada.

Matheson, C., Poesio, M., & Traum, D. (2000). Modeling Grounding and Discourse Obligations Using Update Rules. Paper presented at the First Conference of the North American Chapter of the Association for Computational Linguistics.

Newell, A. (1990). Unified Theories of Cognition. Cambridge, MA: Harvard University Press.

Pelachaud, C., Badler, N. I., & Steedman, M. (1996). Generating Facial Expressions for Speech. Cognitive Science, 20(1).

Bryan Pellom, "SONIC: The University of Colorado Continuous Speech Recognizer", University of Colorado, #TR-CSLR-2001-01, Boulder, Colorado, March, 2001

Poggi, I., & Pelachaud, C. (2000). Emotional Meaning and Expression in Performative Faces. In A. Paiva (Ed.), Affective Interactions: Towards a New Generation of Computer Interfaces (pp. 182-195). Berlin: Springer-Verlag.

Rickel, J., & Johnson, W. L. (2000). Task-Oriented Collaboration with Embodied Agents in Virtual Worlds. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), Embodied Conversational Agents. Boston: MIT Press.

Rickel, J., Marsella, S., Gratch, J., Hill, R., Traum, D., & Swartout, W. (2002). Toward a New Generation of Virtual Humans for Interactive Experiences. IEEE Intelligent Systems, July/August, 32-38.

Scherer, K. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), Approaches to emotion (pp. 293-317).

Thalmann, D. (1993). Human Modeling and Animation. In Eurographics '93 State-of-the-Art Reports.

Traum, D., & Allen, J. F. (1994). Discourse Obligations in Dialogue Processing. In proceessings of the 32nd Annual Meeting of Association for Computational Linguistics.

Traum, D., & Rickel, J. (2002). Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. Paper presented at the First International Conference on Autonomous Agents and Multi-agent Systems, Bologna, Italy.

David R. Traum, Susan Robinson, Jens Stephan (2004) Evaluation of multi-party virtual reality dialogue interaction, in proceedings, Language Resources and Evaluation Conference (LREC 2004).

Wang, D., & Narayanan, S. (2002). A confidence-score based unsupervised MAP adaptation for speech recognition. Paper presented at the Proceedings of 36th Asilomar Conference on Signals, Systems and Computers.