Steve Goes to Bosnia: Towards a New Generation of Virtual Humans for Interactive Experiences

Jeff Rickel, Jonathan Gratch, Randall Hill,
Stacy Marsella, and **William Swartout**
University of Southern California

Abstract

Interactive virtual worlds provide a powerful medium for entertainment and experiential learning. Our goal is to enrich such virtual worlds with virtual humans – autonomous agents that support face-to-face interaction with people in these environments in a variety of roles. While supporting face-to-face interaction in virtual worlds is a daunting task, this paper argues that the key building blocks are already in place. We propose an ambitious integration of core technologies centered on a common representation of task knowledge, and we describe an implemented virtual world and set of characters for an Army peace-keeping scenario that illustrates our vision.

1 Introduction

Interactive virtual worlds provide a powerful medium for entertainment and experiential learning. Navy personnel can become familiar with the layout and operation of a ship to which they will be assigned before they ever set foot on it. Students can learn about ancient Greece by walking its streets, visiting its buildings, and interacting with its people. Scientists can envision life in a colony on Mars long before the required infrastructure is in place. The range of worlds that people can explore and experience is unlimited, ranging from factual to fantasy, set in the past, present, or future.

Our goal is to enrich such worlds with virtual humans – autonomous agents that support face-to-face interaction with people in these environments in a variety of roles. Existing virtual worlds such as military simulations and computer games often incorporate virtual humans with varying degrees of intelligence. However, the ability of these characters to interact with human users is usually very limited; most typically, users can shoot at them and they can shoot back. Those characters that support more collegial interactions, such as in children's educational software, are typically very scripted, and offer human users no ability to carry on a dialogue. In contrast, we envision virtual humans that cohabit virtual worlds with people and support face-to-face dialogues situated in those worlds, serving as guides, mentors, and teammates.

Though our goals are ambitious, we argue in this paper that many of the key building blocks are already in place. Early work on embodied conversational agents (Cassell *et al.* 2000) and animated pedagogical agents (Johnson, Rickel, & Lester 2000) is laying the groundwork for face-to-face dialogues with users. Our prior work on Steve (Rickel & Johnson 1999a; 1999b; 2000) is particularly relevant; Steve cohabits three-dimensional virtual worlds with students, appearing as a graphical human figure (Figure 1), and collaborates with them on tasks as either an instructor or teammate. Section 2 provides a brief overview of Steve as background. Section 3 then identifies four key areas where we believe Steve must be extended – a better body, better natural language capabilities, a model of emotions and personality, and more human-like perception – and argues that the core technology in each of these areas is available. Finally, Section 4 introduces a new project aimed at integrating these capabilities into Steve and describes the first result of the project: an implemented Army peace-keeping scenario that illustrates our vision.

2 Background: Steve

Steve supports many of the capabilities required for face-to-face collaboration with people in virtual worlds. Like earlier intelligent tutoring systems (Wenger 1987), he can help students by answering questions such as "What should I do next?" and "Why?" and by providing feedback on student actions. However, because he has an animated body, and cohabits the virtual world with students, he can interact with them in ways that previous disembodied tutors cannot. For examples, he can lead them around the virtual world, demonstrate tasks, guide their attention through his gaze and pointing gestures, and play the role of a teammate whose activities they can monitor.

Steve's behavior is not scripted. Rather, Steve consists of a set of general, domain-independent capabilities operating over a declarative representation of domain tasks. Steve can be applied to a new domain by simply giving him declarative knowledge of the virtual

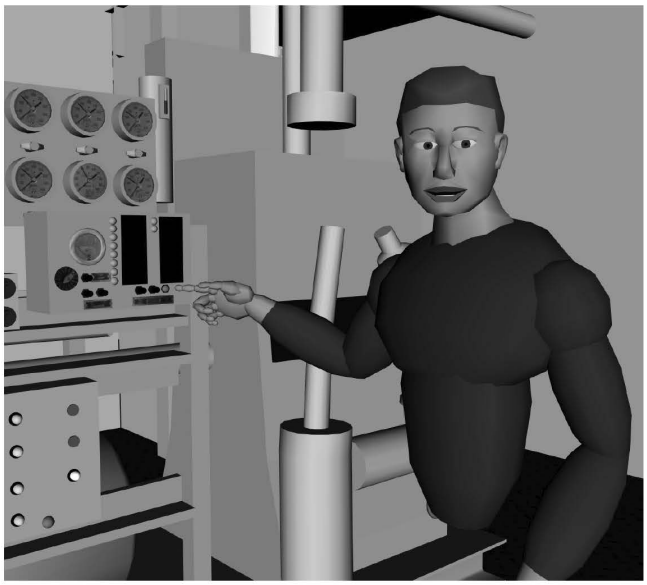

Figure 1: Steve describing a power light

world (i.e., its objects, their relevant simulator state variables, and their spatial properties) and the tasks that can be performed in that world. Task knowledge is given to Steve using a relatively standard hierarchical plan representation. Each task consists of a set of steps (each a primitive action or another task), a set of ordering constraints on those steps, and a set of causal links. The causal links describe the role of each step in the task; each one specifies that one step achieves a particular goal that is a precondition for a second step (or for termination of the task). Steve's general capabilities use such knowledge to construct a plan for completing a task from any given state of the world, revise the plan when the world changes unexpectedly, and maintain a collaborative dialogue with his student and teammates about the task (Rickel & Johnson 1999a).

3 Steve's Key Limitations

Virtual Human Bodies

Research in computer graphics has made great strides in modeling human body motion. Most relevant to our objectives is work that focuses on real-time control of human figures. Within that area, some work uses inverse kinematics to dynamically synthesize body motions that achieve desired end positions for body parts while avoiding collisions with objects along the way (Badler, Phillips, & Webber 1993). Other work focuses on dynamically sequencing motion segments that were created by keyframe animation or motion capture (Lester, Stone, & Stelling 1999); this approach achieves more realistic body motions at the expense of less flexibility. Both approaches have reached a sophisticated level of maturity, and much current work focuses on combining them to achieve both realism and flexibility (Hodgins & Popovic 2000).

Steve was designed to easily accommodate different bodies. His motor control module accepts abstract motor commands from his cognition module and sends detailed commands to his body through a generic API. Integrating a new body into Steve simply requires adding a layer of code to map that API onto the API for the body. Steve's current body, developed by Marcus Thiebaux at the USC Information Sciences Institute, generates all motions dynamically using a simple and efficient set of algorithms. However, it cannot generate obstacle avoidance motions (e.g., reaching around objects), it does not have legs (Steve moves by floating around), and its face has a limited range of expressions with no support for synchronizing lip movements to speech. By integrating a new, more state-of-the-art body onto Steve (or perhaps different bodies for different purposes), we expect to achieve more realistic motions with little or no modification to Steve's other modules.

Spoken Task-Oriented Dialogue

Spoken dialogue is crucial for collaboration. Students must be able to ask a wide range of questions of their virtual human instructors. Teammates must communicate to coordinate their activities, including giving commands and requests, asking for and offering status reports, and discussing options. Without spoken dialogue capabilities, virtual humans cannot fully collaborate with people in virtual worlds.

Steve uses commercial speech recognition and synthesis products to communicate with human students and teammates, but he has no true natural language understanding capabilities. To allow him to understand a new phrase, that phrase must be added to the speech recognizer's grammar, and the speech recognizer must map that phrase to an appropriate semantic representation when it is recognized. This poses two problems. First, the range of utterances that Steve understands is too small. Second, interpretation of utterances is done within the speech recognizer, which, unlike Steve, does not maintain a representation of the current task and dialogue context to guide the interpretation.

While unrestricted natural language dialogue is still a difficult research problem, spoken task-oriented dialogue is becoming practical. In task-oriented dialogue, the computer is given a representation of the tasks on which it can collaborate with human users, and it uses the task knowledge to guide its interpretation of utterances. Multiple research labs have demonstrated robust spoken dialogue systems of this sort (Allen *et al.* 1996; Smith & Hipp 1994). Moreover, these systems rely on the same basic task knowledge that Steve uses, suggesting that an integration of their algorithms into Steve is both feasible and promising.

Emotions and Personality

Steve has no emotions. While this makes him a patient and tolerant collaborator, it leads to two serious limitations. First, because his teaching is emotionally flat, it

is not motivational, and Steve is unable to distinguish mundane instructions (e.g., "To check the oil level, pull out the dipstick") from important ones (e.g., "Whatever you do, don't push that red button!"). We believe that students will show more enthusiasm for Steve's instruction, and will be more likely to retain important points, if Steve injects more emotion into his teaching (Elliott, Rickel, & Lester 1999). Second, Steve is unrealistically rational as a teammate. In many tasks, students must learn how their teammates are likely to react under stress, since learning to monitor teammates and adapt to their errors is an important aspect of team training. Thus, Steve's lack of emotions hampers his performance as an instructor and teammate, and of course it also makes him less engaging for interactive entertainment applications.

Fortunately, research on computational models of emotion has exploded in recent years. Gratch's work on task-oriented emotions (Gratch 2000) is particularly relevant to our interests. In Gratch's model, emotions arise naturally from the status of an agent's plans, and these emotions in turn affect the agent's subsequent decisions. Gratch's model uses the same basic task representation as Steve, with only a few extensions, so we expect an integration of the model into Steve to be relatively straightforward. Moreover, Gratch's model includes a set of personality parameters, allowing us to model a variety of different characters.

Of course, emotional state is conveyed not only by the impact it has on the task-related actions an agent takes, but also by the agent's non-verbal behaviors. This includes discourse-related gestures, as well as various other kinds of body language such as fidgeting with hands, averting gaze, clenching a hand, rubbing a shoulder or slumping in a chair. Incorporating such behaviors into Steve will be critical to making him more realistic and engaging. To address this issue, we plan to extend Gratch's model to include work by Marsella on regulating competing demands on an agent's physical behavior in a behaviorally, emotionally realistic fashion (Marsella, Johnson, & LaBore 2000).

Human-like Perception

The goal of virtual reality is to increase the perceptual fidelity of virtual worlds. For entertainment, the increased perceptual fidelity leads to an increased feeling of immersion and realism. For training and education, the increased perceptual fidelity can help students learn how to use perceptual cues to guide task performance. Virtual humans like Steve can contribute to these goals by teaching students which perceptual cues are relevant and by illustrating the likely perceptual limitations of teammates for which the student must compensate.

Unfortunately, Steve cannot currently help in these areas because he is omniscient. That is, he receives messages from the virtual world simulator describing every state change that is relevant to his task model, regardless of his current location or state of attention. Without a realistic model of human attention and perception, there is no principled basis for limiting his access to these state changes.

Recent research may provide that principled basis. Hill (1999; 2000) has developed a model of perceptual resolution for autonomous agents based on psychological theories of human perception. His model predicts the level of detail at which an agent will perceive objects and their properties in the virtual world, and he has applied his model to synthetic fighter pilots in simulated war exercises. Complementary research by Chopra (Chopra-Khullar & Badler 1999) provides a model of visual attention for virtual humans. Her work, which is also based on human psychological research, specifies the types of visual attention that are required for a variety of basic tasks (e.g., locomotion, object manipulation, and visual search), as well as the mechanisms for dividing attention among multiple such tasks. Together, the work of Hill and Chopra provide a solid foundation for adding more human-like perception to Steve.

4 Status and Plans

To illustrate our vision for virtual humans that can collaborate with people in interactive virtual worlds, we have implemented an Army peace-keeping scenario, which was viewed by several hundred people at the recent grand opening of the new USC Institute for Creative Technologies. As the simulation begins, a human user, playing the role of a U.S. Army lieutenant, finds himself in the passenger seat of a simulated HMMWV speeding towards a 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 2). 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 platoon downtown, 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. The lieutenant carries on a dialogue with the sergeant and medic to assess the situation, issue orders (which are carried out by the sergeant through four squads of soldiers), and ask for suggestions. His decisions influence the way the situation unfolds, culminating in a glowing news story praising his actions or a scathing news story exposing the flaws in his decisions and describing their sad consequences.

This sort of interactive experience clearly has both entertainment and training applications. The U.S. Army is well aware of the difficulty of preparing officers to face such difficult dilemmas in foreign cultures under similarly stressful conditions. By training in engaging, immersive, realistic virtual worlds, officers can gain valuable experience. The same technology could

Figure 2: An interactive peace-keeping scenario featuring (left to right) a sergeant, a mother, and a medic

power a new generation of games or educational software, allowing people to experience exciting adventures in roles that are more rich and interactive than current software supports.

The current implementation of the scenario includes many elements of a general approach, but it also includes a variety of scripted elements. The visual scene is projected onto an 8 foot tall screen that wraps around the viewer in a 150 degree arc (12 foot radius). Immersive audio software provides two tracks of spatialized sounds, one for general ambience (e.g., crowd noise) and another for triggered sounds (e.g., explosions); these sounds are played through ten speakers located around the user and two subwoofers. The graphics, including static scene elements and special effects, are rendered by Multigen/Paradigm's Vega; special effects are currently triggered by a human operator at appropriate times using a graphical user interface, as are radio transmissions (voice clips) from the platoon downtown, the medevac helicopter, and a command center. There are three Steve agents: the sergeant, the medic, and the mother. All other virtual humans (a crowd of locals and four squads of soldiers) are scripted characters implemented in Boston Dynamics's PeopleShop (Boston Dynamics 2000). The three Steve agents use speech recognition to understand the human lieutenant's utterances, but they have no general natural language understanding, so the lieutenant is limited to a small, fixed set of phrases. Only the sergeant uses speech synthesis; all other characters use pre-recorded voice clips. The bodies for all characters, including the three Steve agents, are animated dynamically using PeopleShop; the primitive motions were created using motion capture, and the Steve agents sequence these motions dynamically in response to the situation by sending commands to the PeopleShop run-time software. The medic and sergeant include expressive faces created by Haptek (www.haptek.com) that support synchronization of lip movements to speech. The mother includes a preliminary integration of Gratch's emotion model into Steve, as well as Marsella's model for regulating emotional behavior; her emotional state ebbs and flows dynamically in response to the events around her. The Steve agents perceive the unfolding state of the virtual world through messages they receive from Vega; they do not yet include a model of limited perception. While the system is still far less general than we envision, its development served as an excellent catalyst for understanding the many research challenges that remain, and it will serve as a target as we work towards the new generation of virtual humans outlined in this paper.

5 Acknowledgments

We are grateful to all the people that contributed to the demo described in the final section. Marcus Thiebaux and Ben Moore played a number of valuable roles, including developing the Vega application, the speech recognition software (using Entropic's GrapHvite), and the Vega special effects. Lewis Johnson and Richard Whitney developed the synthetic voice for the sergeant. Kate LaBore created all the animations for the scripted (non-Steve) characters. Marc Raibert and Adam Crane developed the extensions to the PeopleShop software for the virtual human bodies, Whitney Crane processed the motion-capture data for the bodies, and Marlon Veal helped integrate the Haptek heads onto those bodies. Jacki Morie, Erika Sass, and Michael Murguia created most of the graphics for the Bosnian town. Chris Kyriakakis and Dave Miraglia created the sound effects. 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. Jay Douglas and Trevor Hawkins spent many tiring hours running the demo for hundreds of people; Jay

triggered simulation events at appropriate times, and Trevor acted as the lieutenant. Finally, Ramon Gonzalez kept our SGI running. This research is funded by the Army Research Office, with program management at STRICOM.

References

Allen, J. F.; Miller, B. W.; Ringger, E. K.; and Sikorski, T. 1996. Robust understanding in a dialogue system. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 62–70. San Francisco, CA: Morgan Kaufmann.

Badler, N. I.; Phillips, C. B.; and Webber, B. L. 1993. *Simulating Humans*. New York: Oxford University Press.

Boston Dynamics. 2000. *PeopleShop 1.4 User Manual*.

Cassell, J.; Sullivan, J.; Prevost, S.; and Churchill, E., eds. 2000. *Embodied Conversational Agents*. Boston: MIT Press.

Chopra-Khullar, S., and Badler, N. I. 1999. Where to look? automating attending behaviors of virtual human characters. In *Proceedings of the Third International Conference on Autonomous Agents*. ACM Press.

Elliott, C.; Rickel, J.; and Lester, J. 1999. Lifelike pedagogical agents and affective computing: An exploratory synthesis. In Wooldridge, M., and Veloso, M., eds., *Artificial Intelligence Today*, volume 1600 of *Lecture Notes in Computer Science*. Berlin: Springer-Verlag. 195 212.

Gratch, J. 2000. Emile: Marshalling passions in training and education. In *Proceedings of the Fourth International Conference on Autonomous Agents*, 325–332. New York: ACM Press.

Hill, R. 1999. Modeling perceptual attention in virtual humans. In *Proceedings of the 8th Conference on Computer Generated Forces and Behavioral Representation*.

Hill, R. 2000. Perceptual attention in virtual humans: Towards realistic and believable gaze behaviors. In *Proceedings of the AAAI Fall Symposium on Simulating Human Agents*.

Hodgins, J. K., and Popovic, Z., eds. 2000. *Animating Humans by Combining Simulation and Motion Capture*. New Orleans, LA: SIGGRAPH 2000 Course 33 Notes.

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

Lester, J. C.; Stone, B. A.; and Stelling, G. D. 1999. Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments. *User Modeling and User-Adapted Interaction* 9:1–44.

Marsella, S. C.; Johnson, W. L.; and LaBore, C. 2000. Interactive pedagogical drama. In *Proceedings of the Fourth International Conference on Autonomous Agents*, 301–308. New York: ACM Press.

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

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

Rickel, J., and Johnson, W. L. 2000. Task-oriented collaboration with embodied agents in virtual worlds. In Cassell, J.; Sullivan, J.; Prevost, S.; and Churchill, E., eds., *Embodied Conversational Agents*. Boston: MIT Press.

Smith, R. W., and Hipp, D. R. 1994. *Spoken Natural Language Dialog Systems*. New York: Oxford University Press.

Wenger, E. 1987. *Artificial Intelligence and Tutoring Systems*. Los Altos, CA: Morgan Kaufmann.