Lessons Learned from Virtual Humans

William Swartout

■ *Over the past decade, we have been engaged in an extensive research effort to build virtual humans and applications that use them. Building a virtual human might be considered the quintessential AI problem, because it brings together many of the key features, such as autonomy, natural communication, and sophisticated reasoning and behavior, that distinguish AI systems. This article describes major virtual human systems we have built and important lessons we have learned along the way.*

The Institute for Creative Technologies was founded a decade ago to bring together researchers working at the cutting edge of simulation technologies, such as computer graphics, artificial intelligence, and virtual reality to work with people from the entertainment industry who know how to create characters that are compelling and stories that are engaging to work toward the goal of creating the next generation of simulation and training systems. Early on, we decided to focus on training human-oriented skills, such as leadership, negotiation, and cultural awareness.

These skills are based on what is sometimes called *tacit knowledge* (Sternberg 2000), that is, knowledge that is not easily explicated or taught in a classroom setting but instead is best learned through experience. Currently, these training experiences are usually delivered through various human-to-human role-playing exercises. We sought to replace the human role players with virtual humans, which are computer-generated interactive characters that look and act like people but exist in virtual environments. There are several benefits to taking such an approach. Human-based role playing is costly in terms of personnel requirements and is often done at training centers that may be far away from the student's location. In contrast, virtual exercises can be delivered on a laptop, making them available to a student whenever and wherever they are needed, without the need to tie up additional personnel resources.

Vision for Virtual Humans

Our vision for virtual humans is that ultimately they should look and behave as much like real people as possible. Specifically, their behaviors should not be scripted, but instead they should function autonomously, reacting to events in the virtual (and real) world around them and responding appropriately. They should fully perceive their environment including both virtual and real people. They should interact in a fluid, natural way using the full repertoire of human verbal and nonverbal communication. They should model their own and others' beliefs, desires, and intentions and they should exhibit emotions. Finally, they should do all these things in a coherent, integrated fashion.

Achieving the vision for virtual humans outlined above is unquestionably ambitious. Many of the elements in the virtual human vision are consonant with the general vision for AI that McCarthy and others cast for the 1956 Dartmouth Summer Research Project on Artificial Intelligence:

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (reprinted in McCarthy et al. [2006]).

Are our goals too hard? It is certainly a reasonable question. Implementing a virtual human requires integrating a diverse range of AI technologies including speech recognition, natural language understanding, dialogue management, automated reasoning, speech and gesture generation, and animation. Not only are fundamental advances required in some of the subareas but also the technology for integration of the parts is quite complex.

While the ultimate goal remains ahead of us, significant progress has been made, and it is now possible to create systems that successfully use virtual humans in applications. Over the past 15 years or so, a number of research groups have pioneered efforts to explore different themes in virtual humans including body animation and control (Badler, Phillips, and Webber 1993), dialogue and nonverbal behavior (Cassell et al. 1994; Cassell et al. 2000; Hayes-Roth, Gent, and Huber 1997; Pelachaud, Badler, and Steedman 1996), and immersive training with action and dialogue (Rickel and Johnson 1999a). These led to a broad range of applications for health (Bickmore, Pfeifer, and Jack 2009; Marsella, Johnson, and LaBore 2003), interactive entertainment (Mateas and Stern 2003), training (Johnson, Rickel, and Lester 2000; Rickel and Johnson 1999b) and education in a social context (Johnson, Vilhjálmsson, and Marsella 2004; Paiva, Dias, and Aylett 2005; Pelachaud et al. 2002; Zoll et al. 2006)

At the University of Southern California (USC) Institute for Creative Technologies (ICT), we have performed research to understand better how virtual humans interact with real people and how people perceive and react to virtual humans, to develop new technology that extends their capabilities, and to use them in novel applications. We have created characters to help train leadership skills (Swartout et al. 2006; Swartout et al. 2001) and negotiation tactics (Hill et al. 2006; Traum et al. 2005). We have developed virtual patients (Kenny et al. 2008) to help train clinicians in appropriate patient interviewing techniques, and we have even developed virtual humans to help with tasks such as recruitment (Artstein et al. 2009).

As we have constructed these applications, we have learned a number of lessons that make it easier to achieve the vision for virtual humans outlined above. Some of these help facilitate the construction of functioning virtual humans, while others help us identify new areas for research. Some of the lessons come from our collaboration with the entertainment industry, while others stem from good system engineering practices. The main lessons are enumerated as follows: (1) the importance of story, (2) the value of integration, (3) the role of emotion, and (4) the need for a hybrid approach.

In the remainder of this article, I will describe the virtual human applications we have created and elaborate on the lessons learned.

Mission Rehearsal Exercise System

The first training application built at the ICT to use virtual humans was the Mission Rehearsal Exercise (MRE) system (Swartout et al. 2006; Swartout et al. 2001; Traum, Robinson, and Stephan 2004), shown in figure 1. It was started in late 1999 and the project ended in 2003. The goal of this system was to expose junior army officers to thorny dilemmas that might plausibly occur during the course of a mission but that typically are not covered in a standard training manual or course. The emphasis in building this system was to create a prototype that integrated and developed the necessary technologies to show what could be done, rather than to develop all of the content that would be required for a training system that could be deployed.

Our scenario took place in Bosnia. Imagine that you are a second lieutenant. You are part of a peacekeeping force stationed outside a medium-sized town. You have just received orders to travel with your platoon from your base to the downtown area of the town to help another platoon

Figure 1. An Immersive Experience—Mission Rehearsal Exercise System in the ICT Virtual Reality Theater.

quell an uprising. Along the way, you encounter an unwelcome surprise. The platoon's lead Humvee gets into an accident with a car driven by a local woman. Moments later, when you arrive, the scene is already chaotic (figure 2). The following dialogue ensues:

LT: What happened here?

SGT: There was an accident, sir. This woman and her son came from the side street and our driver didn't see them.

LT: Tucci, how is the boy?

Tucci: The boy has critical injuries, sir. Sir, we need to get a medevac in here ASAP.

LT: OK.

LT (on radio): Eagle Base, this is Eagle 2-6, over.

Eagle Base: Eagle 2-6, this is Eagle Base, over.

LT: Requesting a medevac for an injured civilian, over.

Eagle Base: Standby…. Eagle 2-6, this is Eagle Base, medevac launching from operating base Alicia, time now. ETA your location 30. Over.

LT: Roger, 2-6 out.

LT: Sergeant, secure a landing zone.

It will take time to secure the area and set up a landing zone. Meanwhile, a crowd starts to gather, and as luck would have it, a TV camera crew shows up and starts filming everything. As you're pondering this situation, your radio crackles to life. It's the other platoon that's already downtown. They wonder what's been keeping you. They need you now. They're dealing with a riot. What should you do? Split your forces and send one squad forward to help quell the riot, or keep your forces intact? The platoon sergeant acts as a coach for the lieutenant (much as they do in real life) and suggests that it would be a bad idea to split the forces. However, the trainee is free to do what he wants. If he keeps his forces together, the medevac goes smoothly, but if he sends a squad forward he has insufficient forces to clear the landing zone for the helicopter and the evacuation of the boy is substantially delayed. These outcomes were summarized in the form of a TV news report that was presented at the end of the exercise as a form of after action review.

Figure 2. A Dramatized Dilemma.

To enhance the engagement of the simulation, it was presented on a curved, 30-foot-wide screen (see figure 1), and a 10.2 audio system (Kyriakakis 1998) was used to create an immersive sound field.

Creating a system like MRE, or its successor, Stability and Support Operations (SASO), which will

be described next, required us to develop and integrate a broad range of AI technologies, including speech recognition (Lee and Narayanan 2005; Wang and Narayanan 2002); natural language understanding (Bhagat, Leuski, and Hovy 2005); dialogue management (Traum et al. 2008a; Traum and Rickel 2002; Traum et al. 2008b); task and domain reasoning (Hartholt et al. 2008; Traum et al. 2003); emotion modeling (Gratch and Marsella 2004; Marsella and Gratch 2009); natural language generation (DeVault, Traum, and Artstein 2008a, 2008b);[1] speech synthesis;[2] and gesture generation (Thiebaux et al. 2008)

A detailed description of each of these technologies is beyond the scope of this article. I will, however, briefly describe the approach taken by Jonathan Gratch and Stacy Marsella to modeling emotions, since that is one of the pioneering efforts to endow virtual humans with emotions.

While the topic of emotion modeling for AI systems has received some attention, it is fair to say that the great preponderance of AI systems have been concerned with producing intelligent behavior, such as diagnosing a disease, playing a game of chess, or providing clever assistance with office tasks. Relatively little attention has been paid to modeling emotions. However, we have found that because people view virtual humans anthropomorphically, they will ascribe emotional motivations to their behaviors whether the characters have been designed to exhibit emotions or not. If a character is not designed to exhibit emotions, the particular emotions people will see in it will be somewhat difficult to predict. Thus, we found that it was essential to have an explicit model of emotions and design the characters to exhibit those emotions explicitly to avoid giving unintentionally misleading emotional cues.

The virtual humans' emotion model is based upon a psychological theory called appraisal theory (Smith and Lazarus 1990). This theory holds that people form emotions by comparing, or appraising, external events and the resulting states to their own internal goals, beliefs, and desires. If their goals are being facilitated, they will be happy and satisfied, but if they are being thwarted, they will be upset.

More specifically, in appraisal theory, events are assessed in terms of appraisal variables, such as desirability (is this event desired?), expectedness (is the event expected?), controllability (is the event inevitable, or can it be controlled?), and causal attribution (can the event be attributed to someone or something else?).

These appraisal variables are used to derive values for what we think of as conventional emotions, such as sadness or anger. For example, if an event is undesirable, not controllable, and uncertain, it will tend to lead to a feeling of fear. If an event is undesirable and certain, it will give rise to sadness, while if an event is undesirable and someone else can be blamed for it (causal attribution), it will cause anger. The derived emotions can then be used to affect how the character behaves, what it says, and what expressions and other forms of nonverbal communication it may use. Coping strategies can be used by the characters to deal with their emotions and reduce the dissonance between their goals and the events they observe. There are two general classes of coping strategies: problem-focused coping strategies attempt to reduce dissonance by making a change in the world, while emotion-focused strategies make changes to internal beliefs and goals.

The virtual humans not only model their own goals, but they also have models of the assumed goals of the people who interact with them. In our current implementation, the goal models for humans are fixed when the model is created, and that model is not updated during run time. However, based on events that unfold, that model can be used to produce dynamically varying estimates of the emotional reactions of the human trainee during the simulation.

Specifically, let's consider how this model could be applied to the platoon sergeant in the scenario above. The sergeant wants the boy to be healthy, but he is not. At the same time, the blame for the accident has not been resolved. Together, these two appraisals lead to a significant feeling of distress for the sergeant. The sergeant can attempt to cope with the distress by using a problem-solving strategy to make amends, for example by requesting a medevac, or he could use a belief-focused strategy and attempt to blame the mother for the accident.

The emotion model gave our characters believable affective reactions to unfolding events, but as we will see in the next section, it also contributed to cognitive processing by other parts of the virtual human architecture.

SASO-ST and SASO-EN: Negotiating with Virtual Humans

Moving beyond mission rehearsal, we wanted to explore the possibility of creating virtual humans that could negotiate. In the Stability and Support Operations Simulation and Training (SASO-ST) simulation (Traum et al. 2005) and its follow-on, the Stability and Support Operations Extended Negotiations (SASO-EN) simulation (Traum et al. 2008a), the trainee, a captain in the U.S. Army, was given the task of negotiating with a physician running a relief clinic about moving the location of the clinic because military operations were planned in the area and the clinic's safety would be at risk. The captain could not reveal the opera-

tional plans, but he could offer a variety of inducements to move, including support for the move and medical supplies. In SASO-ST, the negotiation was one-on-one between the captain and the doctor (see figure 3). In SASO-EN we added a town elder so that the captain needed to negotiate with both characters about the clinic's location (see figure 4). If the captain was skillful he might be able to convince one character to move and then have that character help persuade the other. But if he negotiated badly, the characters would ally against him.

One unusual aspect of the SASO systems was that the characters used their emotion models to evaluate the statements and proposals that the trainee made. If a negotiating stance or proposition put forth by the trainee resulted in primarily positive emotions, the characters would embrace it, but if negative feelings predominated, it would tend to be rejected. When considering a particular possibility, the characters evaluated it from the current state, but they also envisioned how it might be improved by successful negotiation. Even though the initial appraisal of a suggestion by the captain might be quite negative, the characters would be willing to discuss it if they could see how it could be improved to be a strong alternative through negotiation.

The characters were built to follow psychological models of negotiation (Putnam and Jones 1982; Sillars et al. 1982). Initially, they would attempt to avoid the negotiation entirely, by changing the topic of conversation. If the captain persisted and stayed on topic, the characters would realize that it was necessary to negotiate, and they would initially view the negotiation as a win-lose proposition and hence would point out as many problems and make as many demands as they could to try to place themselves in the strongest negotiating position. If the trainee handled things well, he might eventually be able to get the characters to view the problem from a mutual perspective and enter into an agreement.

Trust was critical throughout the exercise. The trust levels could be set at different values initially and then these initial values would be updated dynamically based on what the trainee said and did. For example, if the captain took time initially to engage in pleasantries or complement the physician on what he was doing, that could help boost trust levels. On the other hand, if the captain asserted that the clinic was unsafe, that would tend to reduce trust since the virtual characters who were not aware of the operational plans would see no reason why the clinic was not safe.

The level of trust is computed as a linear combination of three factors: familiarity, which is assessed based on pleasantries and complements as well as behaving according to conventions; solidarity, which is determined by the degree to which the trainee's statements seem consonant with the character's goals and desires; and credibility, the degree to which the character finds the trainee's statements true or believable.

Figure 3. SASO-ST: Single-Party Negotiation.

Figure 4. Multicharacter Negotiation Provides a Greater Challenge.

If trust were high, the negotiation could go quite smoothly, but if it were low, the negotiation would be very difficult. We found that most test subjects failed in their initial negotiations with the characters because they did not spend enough time on initial trust-building small talk and thus ran into difficulty when they needed to make hard demands later in the scenario. Additional aspects of SASO evaluations are discussed in Traum (2008a, 2008b).

Figure 5. Justina: A Virtual Patient

Moving Beyond Training

Our initial application focus for virtual humans was military training. However, we and others (Kenny et al. 2007; Rich and Sidner 2009) have found that virtual humans can be used in a wide variety of ways that include but go beyond either the military or training.

At the ICT, we have created, and are in the process of creating, a number of virtual human systems that move beyond military training. SGT Star (Artstein et al. 2009) is a character that interactively provides information about career possibilities, benefits, and generally what it's like to be in the army to potential recruits. The SGT Star character is in use and has been deployed with army recruiting teams. Under sponsorship from the National Science Foundation we are collaborating with the Boston Museum of Science to create virtual humans that will function as virtual museum guides, helping visitors select exhibits to see and answering questions about science and technology. In this section, I will discuss another application. We have created a character, called Justina, that acts as a virtual patient (Kenny et al. 2008).

Currently, student clinical psychologists and other care givers are taught interviewing techniques by interacting with standardized patients. Standardized patients are human actors who have been trained about the signs and symptoms of a particular disease. When the student psychologist queries the standardized patient about his symptoms, the patient responds as if he or she had the disease. By asking the right questions and drawing the right inferences, the student is supposed to diagnose the patient correctly.

Unfortunately, standardized patients suffer from some important limitations. There is a significant cost associated with training the actors and making them available to students. As a result, it may be cost prohibitive to train students on rare, infrequently occurring diseases. The actors must be scheduled and will not always be available to students. Finally, actors will not all be able to portray the disease with the same level of accuracy, so each student will not see the same performance.

Justina (shown in figure 5) was designed to play the role of a teenage girl suffering from posttraumatic stress disorder (PTSD) as a result of sexual trauma. Students query Justina about her symptoms, and if they ask the right questions they get responses that should lead them to make the correct diagnosis.

Patients suffering from PTSD exhibit a variety of symptoms such as reexperiencing the traumatic event through dreams or flashbacks, persistent avoidance of stimuli related to the trauma, and persistent symptoms of anxiety or increased arousal, such as hypervigilance or irritability. In all, the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association 2000) lists six major categories for a PTSD diagnosis. Starting with these categories, we developed a set of questions and appropriate answers that a patient with PTSD would give by consulting with psychiatry faculty from the USC Keck School of Medicine and by conducting role-playing exercises with people playing the parts of patient and therapist. These exercises gave us insight into appropriate verbal and nonverbal behavior.

For Justina, it was possible to take a simpler approach to natural language processing than we did in MRE or SASO, due to the nature of the domain. In those systems, there are extended discussions about how to deal with the accident (in the case of MRE) or negotiating an agreement (in the case of SASO). In both cases, the meaning of a statement depends not only on the utterance itself but also on the goals and beliefs of the agents and the dialogue history—what was said before by both parties. For a simple example, if a SASO participant says, "I agree to that," the meaning of that agreement depends completely on what has come before in the conversation.

For the PTSD domain, after some initial general questions, the patient interview largely consists of asking questions about the six diagnostic categories to see the degree to which the patient is experiencing symptoms from a particular category. For example, the clinician might ask: "Are you still thinking about your experience?" to determine to what extent the patient was experiencing flashbacks. Of course, the answers to these questions depend on the degree that the patient is suffering from PTSD, but the answer to a particular question does not depend very much on what has gone before. We found that each question-answer

pair could be viewed largely independently, although some history of what had been said was still useful to avoid having the character repeat lines.

As a result, Justina could use a text classification approach to natural language processing (Leuski and Traum 2008). Based on the role-playing exercises and discussions with medical faculty, the various ways that people asked questions about a particular topic were identified, and the appropriate answers were created and recorded by a voice actor. A statistical text classifier was then trained to identify question categories based on key words and key-word pairs in the input question.

When a student asked Justina a question, a speech recognizer (Pellom 2001) converted the student's utterance to text. The text classifier identified the corresponding question category and the appropriate answer. The answer text was passed to the NonVerbal Behavior Generator (NVBG) (Lee and Marsella 2006), which used a set of heuristic rules to mark up the output with appropriate gestures. The gestures and audio stream for the answer were then synchronized by Smartbody (Thiebaux et al. 2008) and presented to the user.

A preliminary evaluation of Justina with student clinicians (Kenny et al. 2008) showed that it did provide responses across the spectrum of diagnostic criteria and that the appropriateness of the response was reasonable for most of the categories, although responses needed some improvement in three of the categories: increased arousal, effect duration, and life effects.

Lessons Learned

As we have constructed virtual human applications and prototypes, we have learned a number of different kinds of lessons. In this section, I describe these lessons and show how they relate to the applications described above.

Lesson 1: The Importance of Story

The knowledge base for each of our virtual humans is not intended to be completely general purpose, but instead is designed to work in the context of a particular story or scenario. Thus, the knowledge base for a character from the Bosnian MRE simulation would not function well in a SASO negotiation, and vice versa. While the processing mechanisms that the characters use are general purpose and can be reused between scenarios, the knowledge that each character has about tasks that can be performed, its goals and desires, and to some degree its language model are all specialized to the particular scenario that the character is intended to operate within.

It is this specialization that makes it feasible to construct interactive virtual humans.[3] As I mentioned in the introduction, building a virtual human is a challenging problem that requires the integration of a broad range of complex AI technologies, and the integration of these technologies is itself difficult. A good story creates a strong context that will limit what people are likely to say, and thus limiting the scope of the problem makes it more feasible to build virtual humans. For example, in the MRE scenario, it is important for the characters to be able to discuss the health status of the injured child and the feasibility of getting a helicopter to transport him to a hospital, but they do not need to be able to discuss astrophysics or the latest basketball scores, since these topics are extremely unlikely to come up in the context of the scenario.

In addition to making virtual humans more feasible, a good story can perform several other important functions. We have found that a good story can create strong emotional engagement because people come to care about the characters in the simulation, just as empathetic characters in a movie can create greater audience involvement. This effect can be particularly pronounced if the scenario evokes prior experiences. For example, in our first simulation, MRE, which had good (but not photo-realistic) graphics, acceptable (but not flawless) animation, and adequate (but not perfect) natural language interaction, we found that people viewing MRE who had been stationed in Bosnia would often become very emotionally engaged with the simulation and remark that the simulation seemed almost too real. This engagement, we believe, came from the story and the fact that it evoked their prior experiences.

Stories can perform another function. A good story can create a rich social context that raises new research issues. For example, most work in natural language processing has assumed a one-on-one interaction between a computer and a person. However, in SASO-EN, a person would interact with several virtual humans, which raised research issues in how to model attention and how to understand and signal that a conversation with a character was beginning, when it was continuing, and when it had drawn to a close. Additionally, most natural processing research assumes that communication between people and machines is cooperative and that all parties will try to keep on the topic of conversation. However, in both SASO negotiation scenarios, those assumptions were invalid. In a negotiation, participants may try to change the topic of conversation deliberately to avoid the negotiation, and they will try not to be informative and reveal information if it would damage their negotiation goals (Traum 2008a, 2008b).

Finally, if constructed skillfully, a good story can cover technical limitations in the simulation. As

we were designing the SASO negotiation systems, we realized that the natural language understanding modules would occasionally make mistakes or entirely fail to understand what the trainee was saying. Additionally, although text-to-speech technology has improved considerably in the past few years, there are still times when even a good synthesizer will sound mechanical or mispronounce words. We ameliorated both of these problems by choosing to cast the virtual human playing the physician in the scenario as a foreign doctor with a thick Spanish accent. Since he was a foreigner, it was plausible that he would occasionally misunderstand English. The use of a foreign character also reduced expectations for the quality of the speech output, since it is reasonable for foreigners to mispronounce words. We reinforced that point by having the doctor point out that his English was not good when he misunderstood and by inserting common grammatical mistakes that nonnative English speakers make into his speech.

Lesson 2: The Value of Integration

As outlined above, creating a working virtual human ultimately requires integrating a broad range of AI technologies. That integration is itself a daunting task since the interfaces between modules must be carefully specified and any assumptions about how information will be processed must be communicated to all the developers of modules that depend on that data. Despite that, we also found that integration can help make certain problems easier to solve.

Consider a problem from computational linguistics, the problem of resolving the referent of an ambiguous question.[4] In the MRE simulation, when the lieutenant trainee first encounters the accident scene, he may ask a question such as, "What happened here?" An expected response would be for the platoon sergeant to say that an accident had occurred and then go on to describe the details.

To be able to produce such a response, the natural language processing module used by the sergeant character must be able to figure out that the event that the lieutenant is referring to is the accident. That seems easy enough for people to do, but it is more challenging for an AI system since, in fact, a number of events have occurred recently. Thus, potential answers to "What happened here?" would include (1) we got out of our Humvees, (2) the medic started treating the boy, (3) you just drove up, sir, (4) I assembled the troops at the rendezvous point, and (5) there was an accident.

How should one choose from these alternatives? There doesn't seem to be an obvious reason for preferring one over the other.

Computational linguists have recognized the problem of resolving an ambiguous referent for some time, and one of the strongest heuristics that has been proposed is recency: select as the focus the event that happened most recently. In many cases, this heuristic works quite well. However, in this case, we find that when we order the events chronologically and pick the most recent, we get the wrong answer: *You just drove up, sir.*

Even if the system eliminates the first answer since it is likely that the lieutenant knows that he just drove up, the next choice, *the medic started treating the boy,* is still not appropriate.

However, with our integrated virtual human, we have resources available that are not part of a typical computational linguistics system. Consider how the emotion model might help solve a linguistics problem. In people, our emotions serve as a strong focusing mechanism that can affect how we use language. For example, various linguistic "slips" often occur in the context of intense emotions. By looking at the sergeant's emotion model we find (as described above) that he is very upset about the accident. If we then use that information to guide the disambiguation of the referent in the question he was asked, we interpret the question as being about the accident, and hence get the expected answer: There was an accident.

Thus, by integrating several modules in a virtual human that are seemingly concerned with very different things, we can in fact make additional knowledge available that makes it easier to solve certain problems in a particular discipline (such as computational linguistics) than it would be if one looks at the problem from the narrower perspective of that discipline by itself. Fundamentally, the prospect of finding synergies such as this underscores the value of undertaking a multidisciplinary research project, such as building a virtual human, that brings together people from different technical backgrounds (computer science, electrical engineering, and signal processing), the social sciences (linguistics and psychology), and the arts (theater and animation) to work together toward a common goal.

Lesson 3: The Role of Emotion

Because virtual humans look and behave like real people, when real people observe virtual humans they expect the characters to exhibit emotion. Steve (Rickel and Johnson 1999a) was an interactive pedagogical character that would coach trainees in the operation of a large air compressor. It was built without an emotional model and no matter what happened would always maintain a neutral affective tone. Rickel and Johnson found that users felt the lack of emotion was peculiar, and they would sometimes intentionally perform incorrect actions in the simulation in an effort to get an emotional reaction out of the agent.

We thus came to the conclusion that emotion

models were essential. However as we developed emotion models we found that they could play a more extensive role than just making the characters believable.

In fact, we found that the emotion model could have a much broader impact on the cognitive processing in virtual humans. First, the emotion model can provide a strong focusing mechanism. As we have seen in the discussion about the value of integration, Lesson 2, the focus that emotion provides can help natural language processing modules disambiguate ambiguous references. Second, the emotion model can inform decision making. When the SASO characters decide to consider, reject, or accept proposals that the trainee puts forth, that decision is made using the emotion model. We believe that understanding the role of emotion in cognition may be one of the keys to creating truly intelligent characters (Gratch and Marsella 2007).

Lesson 4: The Need for a Hybrid Approach

Throughout its history, AI researchers have often tended to act like "true believers," advocating one particular theory or approach as the way to solve much if not all of the AI problem. At various points logic-based knowledge representations, rule-based systems, neural nets, and Bayesian statistics have all been put forth as the way to make progress in AI. In some ways, this strong focus on a particular approach is good because it allows the field to better understand the strengths and weaknesses of that methodology. However, at the ICT, perhaps in part because of our connection with the entertainment industry, we have come to appreciate the value of a hybrid approach.

In a modern motion picture, filmmakers try to tell a story in the most engaging and cost-effective way possible. Often live action, computer graphics, and models are seamlessly intermixed to produce the desired results. Each has its strengths and weaknesses. Live action is good for capturing realistic performances but is not appropriate for highly dangerous scenes. Computer graphics allow filmmakers to create scenes that would be difficult or impossible to realize in live action but cannot yet compete with the highly nuanced performances provided by skilled live actors. Models allow filmmakers to cast a grand vision without incurring the expense of building a full-scale set. Because each technique has strengths, yet none dominates the other, the most effective strategy usually involves a hybrid approach that mixes these techniques together, sometimes in the same scene.

Some early AI systems that tackled ambitious problems also embraced taking a hybrid approach. For example, the algebraic manipulation system MACSYMA (Moses 1971, 1974) used multiple representations for mathematical expressions. The overhead of converting between representations was worth the cost because having the right representation for a problem made it much easier to solve.

We have also seen the value of a hybrid approach. As described above, we used different approaches to natural language processing in Justina and the MRE and the SASO systems based on differences in the requirements of the characters and scenarios. The deep understanding of MRE and SASO is required to handle the nuances of negotiation, while the statistical approach of Justina is more robust in handling speech-recognition errors. Similarly, some of our characters use text-to-speech to speak while others use prerecorded speech spoken by a voice actor. Text-to-speech works well if the character has a lot of different things to say that may change dynamically from session to session and the character does not need to express much emotion or prosody when speaking, while prerecorded lines can be very expressive but are feasible only if the lines can all be worked out in advance. Taking a hybrid approach allows us to develop the most effective and robust solution given the current state of technology.

Even within a particular simulation or scenario, it is often important to be able to use a hybrid approach. Some characters may have much more sophisticated roles than others and hence require deep reasoning and natural language processing, while simpler techniques may suffice for the "bit players." It is also possible to benefit from a hybrid approach within a single, individual character. For example, a character may use lines recorded by a voice actor for frequently occurring utterances, while relying on text-to-speech for infrequent or unanticipated lines. That approach allows the character to sound natural in most circumstances, while providing greater coverage of the domain than would be feasible if relying on a voice actor alone.

A corollary of using a hybrid approach is that the virtual human architecture must support it. It must be possible to swap out processing modules and replace them with other modules that use a different approach. That implies that the application programming interfaces (APIs) for modules must be well specified. Additionally, greater flexibility can be provided if the modules are written so that they are independent from one another. Thus, ideally the decision to use a text-to-speech synthesizer versus a prerecorded voice should not affect how the characters' gestures are generated.

Summary

Building virtual humans that autonomously behave, reason, and communicate like real people

is certainly one of the grand challenges for AI. In the introduction, I outlined our vision for virtual humans. Over the last decade, we have made substantial progress toward achieving that vision. Within the context of a particular scenario, we have been able to build virtual characters that interact using natural language and that use gestures and eye gaze for nonverbal communication. Our characters can model their own beliefs, desires, and intentions as well as those of others, and they can model emotion. Being able to perceive the actions and gestures of real people is still an active area of research (Chu and Nevatia 2008; Morency, de Kok, and Gratch 2008; Morency et al. 2005) as is the goal of making virtual humans look like real people (Alexander et al. 2009).

Although achieving the ultimate virtual human remains in the future, it is now possible to build working prototypes and useful applications using current virtual human technology. As I have outlined in this article, there are a number of factors that make building a workable virtual human more feasible than it might at first appear. A good story creates a strong context that in essence allows a virtual human to operate in a microworld, but one that is still useful (Lesson 1). Building a virtual human requires the integration of diverse processing modules, which makes it possible to find synergies that might not otherwise be available (Lesson 2). An emotion model can not only improve characters' believability, but can also have a major effect on cognitive processing (Lesson 3). Finally, taking a hybrid, modular approach to the system modules and architectures allows one to match the technology to the task (Lesson 4). It will be exciting to see what the next decade of virtual human research brings.

Acknowledgements

The work described here represents the culmination of a great deal of hard work by a number of very talented individuals both at the ICT, USC, and other research laboratories. Some of the key individuals who helped make these efforts successful include Jonathan Gratch, David Traum, Stacy Marsella, Randall Hill, Anton Leuski, Patrick Kenny, Eduard Hovy, Shri Narayanan, Arno Harthold, Ed Fast, Chad Lane, Mark Core, David DeVault, Diane Piepol, and Skip Rizzo. This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the government and no official endorsement should be inferred.

Notes

1. This example is also described in Swartout et al. (2006), 96–108.

2. See the unpublished 2003 manuscript by D. Traum, M. Fleischman, and E. Hovy, NL Generation for Virtual Humans in a Complex Social Environment (www.mit.edu/~mbf/AAAI-SS_03.pdf).

3. Initially, we performed some research in speech synthesis techniques (Johnson et al. 2002). More recently, we have used commercial synthesis systems such as Rhetorical.

4. At the same time, the fact that we use general-purpose processing mechanisms that are reusable to operate on the specialized knowledge bases makes it easier to switch between domains by substituting one knowledge base for another. This reusability made it possible to build the initial version of the SASO-ST system in about 90 days by reusing the MRE code base.

References

Alexander, O.; Rogers, M.; Lambeth, W.; Chiang, M.; and Debevec, P. 2009. Creating a Photoreal Digital Actor: The Digital Emily Project. Paper presented at the Sixth European Conference on Visual Media Production (CVMP), 12–13 November, London, UK.

American Psychiatric Association 2000. *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV-TR, 4th ed.). Washington, DC: American Psychiatric Association.

Artstein, R.; Gandhe, S.; Gerten, J.; Leuski, A.; and Traum, D. 2009. Semi-Formal Evaluation of Conversational Characters. In *Languages: From Formal to Natural—Essays Dedicated to Nissim Francez on the Occasion of His 65th Birthday,* Lecture Notes in Computer Science 5533, ed. O. Grumberg, M. Kaminski, S. Katz, and S. Wintner. Berlin: Springer.

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

Bhagat, R.; Leuski, A.; and Hovy, E. 2005. Statistical Shallow Semantic Parsing Despite Little Training Data. Poster paper presented at the Eleventh International Workshop on Parsing Technology, October 9–10, Vancouver, Canada.

Bickmore, T.; Pfeifer, L.; and Jack, B. 2009. Taking the Time to Care: Empowering Low Health Literacy Hospital Patients with Virtual Nurse Agents. Paper presented at the ACM Special Interest Group on Computer Human Interaction (SIGCHI) Conference on Human Factors in Computing Systems, April 4–9, Boston, MA.

Cassell, J.; Pelachaud, C.; Badler, N.; Steedman, M.; Achorn, B.; Becket, T.; Douville, B.; Prevost, S.; and Stone, M. 1994. Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. Paper presented at the ACM SIGGRAPH 1994, July 24–29, Orlando, FL.

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

Chu, C.-W., and Nevatia, R. 2008. Real-Time 3D Body Pose Tracking from Multiple 2D Images. In *Articulated Motion and Deformable Objects,* Lecture Notes in Computer Science 5098, 42–52. Berlin: Springer.

DeVault, D.; Traum, D.; and Artstein, R. 2008a. Making Grammar-Based Generation Easier to Deploy in Dialogue Systems. Paper presented at the Ninth SIGdial Workshop on Discourse and Dialogue, June 19–20, Columbus, OH.

DeVault, D.; Traum, D.; and Artstein, R. 2008b. Practical

Grammar-Based NLG from Examples. Paper presented at the Fifth International Natural Language Generation Conference, June 12–14, Salt Fork, OH.

Gratch, J., and Marsella, S. 2004. A Domain Independent Framework for Modeling Emotion. *Journal of Cognitive Systems Research* 5(4): 269–306.

Gratch, J., and Marsella, S. 2007. The Architectural Role of Emotion in Cognitive Systems. In *Integrated Models of Cognitive Systems*, ed. W. Gray. New York: Oxford University Press.

Hartholt, A.; Russ, T.; Traum, D.; Hovy, E.; and Robinson, S. 2008. A Common Ground for Virtual Humans: Using an Ontology in a Natural Language Oriented Virtual Human. Paper presented at the Sixth International Conference on Language Resources and Evaluation (LREC 2008), 28–30 May, Marrakech, Morocco.

Hayes-Roth, B.; Gent, R. V.; and Huber, D. 1997. Acting in Character. In *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents,* ed. R. Trapple and P. Petta, 92–112. Berlin: Springer-Verlag.

Hill, R. W.; Lane, H. C.; Core, M.; Forbell, E.; Kim, J.; Belanich, J.; Dixon, M.; and Hart, J. 2006. Pedagogically Structured Game-Based Training: Development of the ELECT BiLAT Simulation. Paper presented at the 25th Army Science Conference, November 27–30, Orlando, FL.

Johnson, W. L.; Narayanan, S.; Whitney, R.; Das, R.; Bulut, M.; and LaBore, C. 2002. Limited Domain Synthesis of Expressive Military Speech for Animated Characters. Paper presented at the 7th International Conference on Spoken Language Processing, September 16–20, Denver, CO.

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

Johnson, W. L.; Vilhjálmsson, H.; and Marsella, S. 2004. The DARWARS Tactical Language Training System. Paper presented at the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), December 6–9, Orlando, FL.

Kenny, P.; Hartholt, A.; Gratch, J.; Swartout, W.; Traum, D.; Marsella, S.; and Piepol, D. 2007. Building Interactive Virtual Humans for Training Environments. Paper presented at the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), November 26–29, Orlando, FL.

Kenny, P.; Parsons, T.; Gratch, J.; and Rizzo, A. 2008. Evaluation of Justina: A Virtual Patient with PTSD. Paper presented at the 8th International Conference on Intelligent Virtual Agents, September 1–3, Tokyo, Japan.

Kyriakakis, C. 1998. Fundamental and Technological Limitations of Immersive Audio Systems. *Proceedings of the IEEE* 86(5): 941–951.

Lee, C. M., and Narayanan, S. 2005. Towards Detecting Emotions in Spoken Dialogs. *IEEE Transactions on Speech and Audio Processing* 13(2): 293–302.

Lee, J., and Marsella, S. 2006. Nonverbal Behavior Generator for Embodied Conversational Agents. Paper presented at the 6th International Conference on Intelligent Virtual Agents, August 21–23, Marina del Rey, CA.

Leuski, A., and Traum, D. 2008. A Statistical Approach for Text Processing in Virtual Humans. Paper presented at the 26th Army Science Conference, December 1–4, Orlando, FL.

Marsella, S., and Gratch, J. 2009. EMA: A Model of Emotional Dynamics. *Journal of Cognitive Systems Research* 10(1): 70–90.

Marsella, S.; Johnson, W. L.; and LaBore, C. 2003. Interactive Pedagogical Drama for Health Interventions. Paper presented at the Conference on Artificial Intelligence in Education, July 20–24, Sydney, Australia.

Mateas, M., and Stern, A. 2003. Facade: An Experiment in Building a Fully-Realized Interactive Drama. Paper presented at the Game Developer's Conference, March 8–10, San Jose, California.

McCarthy, J.; Minsky, M. L.; Rochester, N.; and Shannon, C. E. 2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence (August 31, 1955). *AI Magazine* 27(4): 12–14.

Morency, L.-P.; de Kok, I.; and Gratch, J. 2008. Context-Based Recognition during Human Interactions: Automatic Feature Selection and Encoding Dictionary. Paper presented at the 10th International Conference on Multimodal Interfaces, October 20–22, Chania, Greece.

Morency, L.-P.; Sidner, C.; Lee, C.; and Darrell, T. 2005. Contextual Recognition of Head Gestures. Paper Presented at the 7th International Conference on Multimodal Interfaces, October 4–6, Torento, Italy.

Moses, J. 1971. Algebraic Simplification: A Guide for the Perplexed. In *Proceedings of the Second ACM Symposium on Symbolic and Algebraic Manipulation,* 282–304. New York: Association for Computing Machinery.

Moses, J. 1974. MACSYMA—The Fifth Year. *SIGSAM Bulletin* 8(3): 105–110.

Paiva, A.; Dias, J.; and Aylett, R. 2005. Learning by Feeling: Evoking Empathy with Synthetic Characters. *Applied Artificial Intelligence* 19(3–4): 235–266.

Pelachaud, C.; Badler, N. I.; and Steedman, M. 1996. Generating Facial Expressions for Speech. *Cognitive Science* 20(1): 1–46.

Pelachaud, C.; Carofiglio, V.; Carolis, B. D.; Rosis, F. D.; and Poggi, I. 2002. Embodied Contextual Agent in Information Delivering Application. Paper presented at the First International Joint Conference on Autonomous Agents and Multiagent Systems, July 15–19, Bologna, Italy.

Pellom, B. 2001. SONIC: The University of Colorado Continuous Speech Recognizer (No. TR-CSLR-2001-01). Boulder, Colorado: University of Colorado

Putnam, L., and Jones, T. 1982. Reciprocity in Negotiations: An Analysis of Bargaining Interaction. *Communication Monographs* 49(3): 171–191.

Rich, C., and Sidner, C. L. 2009. Robots and Avatars as Hosts, Advisors, Companions, and Jesters. *AI Magazine* 30(1): 29–41.

Rickel, J., and Johnson, W. L. 1999a. Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition, and Motor Control. *Applied Artificial Intelligence* 13(4–4): 343–382.

Rickel, J., and Johnson, W. L. 1999b. Virtual Humans for Team Training in Virtual Reality. Paper presented at the Ninth International Conference on Artificial Intelligence in Education, July, Le Mans, France.

Sillars, A. L.; Coletti, S. F.; Parry, D.; and Rogers, M. A.

1982. Coding Verbal Conflict Tactics: Nonverbal and Perceptual Correlates of the Avoidance-Distributive-Integrative Distinction. *Human Communication Research* 9(1): 83–95.

Smith, C. A., and Lazarus, R. 1990. Emotion and Adaptation. In *Handbook of Personality: Theory and Research,* ed. L. A. Pervin, 609–637. New York: Guilford Press.

Sternberg, R. J. 2000. *Practical Intelligence in Everyday Life*. New York: Cambridge University Press.

Swartout, W.; Gratch, J.; Hill, R.; Hovy, E.; Marsella, S.; Rickel, J.; and Traum, D. 2006. Toward Virtual Humans. *AI Magazine* 27(2): 96–108.

Swartout, W.; Hill, R.; Gratch, J.; Johnson, W. L.; Kyriakakis, C.; LaBore, C.; Lindheim, R.; Marsella, S.; Miraglia, D.; Moore, B.; Morie, J.; Rickel, J.; Thiebaux, M.; Tuch, L.; Whitney, R.; and Douglas, J. 2001. Toward the Holodeck: Integrating Graphics, Sound, Character and Story. Paper presented at the Fifth International Conference on Autonomous Agents, May 28–June 1, Montreal, Canada.

Thiebaux, M.; Marshall, A.; Marsella, S.; and Kallmann, M. 2008. SmartBody: Behavior Realization for Embodied Conversational Agents. Paper presented at the Seventh International Conference on Autonomous Agents and Multi-Agent Systems, May 12–16, Estoril, Portugal.

Traum, D.; Gratch, J.; Marsella, S.; Lee, J.; and Hartholt, A. 2008a. Multi-Party, Multi-Issue, Multi-Strategy Negotiation for Multi-Modal Virtual Agents. Paper presented at the 8th International Conference on Intelligent Virtual Agents, September 1–3, Tokyo, Japan.

Traum, D., and 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, July 15–19, Bologna, Italy.

Traum, D.; Rickel, J.; Gratch, J.; and Marsella, S. 2003. Negotiation over Tasks in Hybrid Human-Agent Teams for Simulation-Based Training. Paper presented at the Second International Conference on Autonomous Agents and Multiagent Systems, 14–18 July, Melbourne, Australia.

Traum, D.; Robinson, S.; and Stephan, J. 2004. Evaluation of Multi-Party Virtual Reality Dialogue Interaction. Paper presented at the Fourth International Conference on Language Resources and Evaluation (LREC 2004), 26–28 May, Lisbon, Portugal.

Traum, D.; Swartout, W.; Gratch, J.; and Marsella, S. 2008b. A Virtual Human Dialogue Model for Non-Team Interaction. In *Recent Trends in Discourse and Dialogue,* ed. L. Dybkjaer and W. Minker, 45–67. Berlin: Springer.

Traum, D.; Swartout, W.; Marsella, S.; and Gratch, J. 2005. Fight, Flight, or Negotiate Believable Strategies for Conversing under Crisis. Paper presented at the Intelligent Virtual Agents Conference (IVA 2005), September 12–14, Kos, Greece.

Wang, D., and Narayanan, S. 2002. A Confidence-Score Based Unsupervised MAP Adaptation for Speech Recognition. Paper presented at the 36th Asilomar Conference on Signals, Systems and Computers, November 3–6, Monterey, CA.

Zoll, C.; Enz, S.; Schaub, H.; Aylett, R.; and Paiva, A. 2006. Fighting Bullying with the Help of Autonomous Agents in a Virtual School Environment. Paper presented at the 7th International Conference on Cognitive Modelling (ICCM-06), April 5–8, Trieste, Italy.

William Swartout is director of technology for USC's Institute for Creative Technologies and a research professor of computer science at USC. He received his Ph.D. and M.S. in computer science from MIT and his bachelor's degree from Stanford University. Swartout has been involved in the research and development of artificial intelligence systems for more than 30 years. His particular research interests include virtual humans, explanation and text generation, knowledge acquisition, knowledge representation, intelligent computer-based education, and the development of new AI architectures. In July 2009, Swartout received the Robert Engelmore Award from the Association for the Advancement of Artificial Intelligence for seminal contributions to knowledge-based systems and explanation, groundbreaking research on virtual human technologies and their applications, and outstanding service to the artificial intelligence community. He is a Fellow of AAAI, has served on the Executive Council of AAAI, and is past chair of the Special Interest Group on Artificial Intelligence (SIGART) of the Association for Computing Machinery (ACM). He is a past member of the Air Force Scientific Advisory Board and currently serves on the Board on Army Science and Technology of the National Academies and the JFCOM Transformation Advisory Group.