Building Interactive Virtual Humans for Training Environments

Patrick Kenny, Arno Hartholt, Jonathan Gratch, William Swartout,
David Traum, Stacy Marsella, Diane Piepol
Institute for Creative Technologies /
University of Southern California

Marina Del Rey, CA

{kenny, hartholt, gratch, swartout, traum, piepol}@ict.usc.edu
marsella@isi.usc.edu

ABSTRACT

There is a great need in the Joint Forces to have human to human interpersonal training for skills such as negotiation, leadership, interviewing and cultural training. Virtual environments can be incredible training tools if used properly and used for the correct training application. Virtual environments have already been very successful in training Warfighters how to operate vehicles and weapons systems. At the Institute for Creative Technologies (ICT) we have been exploring a new question: can virtual environments be used to train Warfighters in interpersonal skills such as negotiation, tactical questioning and leadership that are so critical for success in the contemporary operating environment? Using embodied conversational agents to create this type of training system has been one of the goals of the Virtual Humans project at the institute. ICT has a great deal of experience building complex, integrated and immersive training systems that address the human factor needs for training experiences.

This paper will address the research, technology and value of developing virtual humans for training environments. This research includes speech recognition, natural language understanding & generation, dialogue management, cognitive agents, emotion modeling, question response managers, speech generation and non-verbal behavior. Also addressed will be the diverse set of training environments we have developed for the system, from single computer laptops to multi-computer immersive displays to real and virtual integrated environments.

This paper will also discuss the problems, issues and solutions we encountered while building these systems. The paper will recount subject testing we have performed in these environments and results we have obtained from users. Finally the future of this type of Virtual Humans technology and training applications will be discussed.

ABOUT THE AUTHORS

Patrick G. Kenny has over fifteen years of industry experience working in the field of software development and artificial intelligence. He is currently the System Integration Scientist leading the effort to design and integrate the next generation virtual human architecture with cognitive and simulation architectures for the Virtual Human immersive and interactive training environment. Mr. Kenny previously worked at the University of Michigan Artificial Intelligence Lab, researching and developing robotics, mission planning and cognitive models for unmanned ground robotic vehicles. Mr. Kenny is also a founder of Soar Technology Inc. a company specializing in AI and cognitive models. Mr. Kenny's research interests are in creating highly realistic high fidelity interactive virtual humans, personality models for virtual humans, robotics and gaming. Mr. Kenny has a BS from The University of Minnesota and is a member of SIGGRAPH, AAAI and IGDA.

Arno Hartholt is of Dutch origin and received his M.S. in Computer Science from the University of Twente, the Netherlands. He participated in several research activities at the University of Twente involving human-computer interaction and embodied conversational agents. Mr. Hartholt has worked as a software engineer at various

companies, including innovative start ups like Verne Business Excellence and Sqills IT Revolutions, and the multinational financial corporation Fortis, from which he received a prestigious scholarship. Mr. Hartholt has a wide variety of interests including virtual humans, cognitive modeling, natural language processing, knowledge representation, knowledge sharing, human-computer interaction, game technology and project management.

Dr. Jonathan Gratch is an Associate Director for Virtual Humans Research at ICT and a Research Associate Professor of Computer Science at the University of Southern California. He completed his Ph.D. in Computer Science at the University of Illinois in Urban-Champaign in 1995. His research focuses on virtual humans and cognitive modeling. He studies the relationship between cognition and emotion, the cognitive processes underlying emotional responses, and the influence of emotion on decision making and physical behavior. He has worked on a number of applications of virtual agents, including considerable experience in the research and development of automated and semi-automated forces in military training simulations (STOW, CFOR and ASTT efforts). He is sitting member of the organizing committee for the International Conference on Intelligent Virtual Agents, a member of the European Unions Network of Excellence on Emotion and Human-computer interaction. He is a member of the American Association for Artificial Intelligence (AAAI) and the International Society for Research on Emotion. Dr. Gratch is the author of over 100 technical articles.

Dr. William Swartout is Director of Technology at ICT 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. Dr. Swartout has been involved in the research and development of AI systems for over 30 years. His particular research interests include virtual humans, explanation and text generation, knowledge acquisition, knowledge representation, knowledge sharing, education, intelligent agents and the development of new AI architectures. Dr. Swartout is a Fellow of the American Association for Artificial Intelligence (AAAI), has served on the Board of Councilors of AAAI and is past chair of the Special Interest Group on Artificial Intelligence of the Association for Computing Machinery (ACM). He is a member of the Air Force Scientific Advisory Board, the Board on Army Science and Technology of the National Academies, and the JFCOM Transformation Advisory Group.

Dr. Stacy Marsella is a project leader at the University of Southern California's Information Sciences Institute (USC/ISI) and a research associate professor of Computer Science at the University of Southern California. Dr. Marsella received his Ph.D. from Rutgers University in 1993. His research interests include multi-agent systems, computational models of emotion, modeling social interaction and group behavior as well as the use of simulation in education. Dr. Marsella has developed numerous agent-based systems across a range of applications, including military training applications, exploratory social simulations for modeling psychological operations, language training and health interventions. Dr. Marsella has numerous publications spanning research in artificial intelligence, cognitive science, psychology, political science and the arts.

Dr. David Traum is a Research Scientist at ICT and a research assistant professor of Computer Science at the University of Southern California. He completed his Ph.D. in Computer Science at the University of Rochester in 1994. His research focuses on collaboration and dialogue communication between agents, including both human and artificial agents. Of primary interest is the interaction between the individual cognitive functioning and the social fabric, and the relationship between task-related and communicative actions. He has engaged in theoretical, implementational, and empirical approaches to the problem, studying human-human natural language and multi-modal dialogue, as well as building a number of dialogue systems to communicate with human users. Dr. Traum is author of over 100 technical articles, has served on many conference program committees, and is currently the president of SIGDIAL, the international special interest group in discourse and dialogue.

Diane Piepol is considered among the pioneers in the Los Angeles based digital graphics community. As a producer, digital educator and digital effects artist, her credits encompass work on R&D projects, feature films, broadcast television and commercials. Piepol joined the Institute for Creative Technologies team in the summer of 2000. She has served as the project director of ICT's *FlatWorld* immersive, mixed reality display system since March 2003. Previous to this, as producer with Dr. Paul Debevec's Graphics Lab, she was instrumental in the *LightStage 2* work for Sony Pictures Imageworks' *Spiderman 2* resulting in the first feature film credit for the ICT. Within the computer graphics production community, Ms. Piepol has line produced animation segments for James Cameron's prototype web channel called *Earthship.tv*. She produced the 90 minute, juried festival of computer animation, *The Electronic Theater*, for the *SIGGRAPH 26th International Conference on Computer Graphics*. Her

feature film credits include: *Eraser, True Lies, Waiting to Exhale, Tank Girl, Super Mario. Bros.*and others. Her client list includes: Silicon Graphics Inc., Dreamworks Interactive, Mass Illusion, VIFX, Discreet Logic, Digital Domain, CBS Animation, RGA/LA and Robert Abel & Associates. Ms. Piepol has taught computer graphics at the UCLA Department of Design and has served two elected terms as Chair of the Los Angeles ACM SIGGRAPH Professional Chapter. Diane is also a founding member of the Visual Effects Society (VES) and a current member of the VES Board of Directors.

Building Interactive Virtual Humans for Training Environments

Patrick Kenny, Arno Hartholt, Jonathan Gratch, William Swartout,
David Traum, Stacy Marsella, Diane Piepol
Institute for Creative Technologies /
University of Southern California

Marina Del Rey, CA

{kenny, hartholt, gratch, swartout, traum, piepol}@ict.usc.edu
marsella@isi.usc.edu

INTRODUCTION

In order to maintain the edge and transform the joint forces, training systems of the future will need to simulate all aspects of a virtual world, from the physics of vehicles to realistic human behavior. The Virtual Humans project at the Institute for Creative Technologies (ICT) is concentrating on building high fidelity embodied agents that are integrated into these environments. These agents would provide a social and human focus to training and serve as guides, mentors, competitors and teammates or other roles that support interactive face-to-face interaction and provide a powerful mechanism for training interpersonal skills and experiential learning. Existing virtual worlds, such as military simulations and computer games, often incorporate virtual humans with varying degrees of intelligence that provide training for physical skills, team training or strategy and tactics. However, these characters' ability to interact with human users is usually limited to shooting engagements. There has been a growing need in recent years to train leadership, negotiation, cultural awareness and interviewing skills. The goal of the Virtual Humans project is to fill this gap in these training environments.

These interpersonal skills require a vast knowledge of the various aspects of human behavior that are hard to formalize and appropriately display. To effectively perform this task requires building virtual humans that have the capability to interact with trainees on this interpersonal level. By incorporating this set of human behavior with virtual characters, virtual worlds can be made applicable to a wide range of training tasks that currently require labor-intensive live exercises, role playing, or are taught non-experientially (e.g., in a classroom setting). This potential depends on our success in creating engaging characters that convey three main characteristics:

- **Believable;** they must provide a sufficient illusion of human-like behavior so that the human user will be drawn into the scenario.
- **Responsive**; they must respond to the human user and to the events surrounding them, which will be fundamentally influenced by the user's actions and contain a rich inner dynamic that unfolds in response to the scenario.
- **Interpretable**; the user must be able to interpret their responses to situations, including their dynamic cognitive and emotional state, using the same verbal and nonverbal cues that people use to understand one another.

Thus, the virtual humans cannot simply create an illusion of life through cleverly designed randomness in their behavior; their inner behavior must respond appropriately to a dynamically unfolding scenario, and their outward behavior must convey that inner behavior accurately and clearly. Building virtual humans requires fundamental advances in AI, graphics and animation. These intelligent agents must perceive and respond to events in the virtual world. They must be able to construct and revise plans in coordination with humans and other agents. They must have and express realistic emotions and they must be able to carry on spoken dialogues with humans and other agents, including all the nonverbal communication that accompanies human speech (e.g., eye contact and gaze aversion, facial expressions, and gestures). While there has been work on all these individual components, no previous effort has tried to integrate all of these capabilities into a single agent and to deal with the complex interplay among them.

This paper will describe the Integrated Virtual Humans project and associated research at ICT, the technology used, the applications built and lessons learned in the hopes that more of these systems can be built and deployed.

RELATED WORK

The ICT Integrated Virtual Humans effort is widely considered the most advanced research project of its kind in the world, but the scope of building a complete virtual human is too vast for any one research group. ICT's virtual human research is a *multidisciplinary* effort, joining traditional artificial intelligence problems (Anderson and Lebiere, 1998; Laird, 2001) with a range of issues from computer graphics (Lee and Waters, 1995; Perlin, 1995; Becheiraz and Thalmann, 1996; Rousseau and Hayes-Roth, 1996; Kalra and Magnenat-Thalmann, 1998; Brand, 1999) to the social sciences (Frijda, 1987; Wiggins, 1996). ICT's virtual human research is also a *multi-institutional* effort, involving cooperation across USC, but also involving joint projects and the development of shared tools and standards with institutions across the world including MIT, University of Colorado, UPenn, University of Paris, University of Twente, Reykjavik University, and the European Union's HUMAINE Network of Excellence on Emotion and Human-Computer Interaction.

There are also a small but growing number of independent efforts to develop virtual humans approximating the scope considered at ICT. These have been applied to a variety of applications including training, tutoring, marketing, and entertainment. Current major research efforts include the work of Justine Cassell's group at Northwestern, Elisabeth André at the University of Augsburg, Ron Cole at the University of Colorado, and Ipke Wachsmuth, University of Bielefeld. ICT's effort is generally acknowledged as the most comprehensive in terms of the breadth and depth of integrated capabilities. Few of these efforts directly address issues and applications of military relevance. One effort that is similar in goals is the Solider Virtual System at the University of Iowa, however their goals are to model and simulate the biomechanics and internal structures.

Our work is closely related to other research on embodied conversational agents (Gratch, 2002, Cassell, Bickmore et al., 2000). Cassell and her colleagues have built several sophisticated systems that support face-to-face conversations between a pair of virtual humans (Cassell, Pelachaud et al., 1994) and between a human user and a virtual human. Their most recent agent, Rea (Cassell, Bickmore et al., 2000), acts as a real estate agent, conversing with human users about available apartments and homes. Although several other recent systems have applied artificial intelligence to team training (Bindiganavale, Schuler et al., 2000), none of them provide embodied virtual humans that can collaborate with human users in a three-dimensional virtual world. The PuppetMaster [Marsella et al, 1998] serves as an automated assistant to a human instructor for large-scale simulation-based training. AETS [Zachary et al, 1998] monitors a team of human students as they run through a mission simulation using the actual tactical workstations aboard a ship, rather than a virtual mock-up. AETS employs detailed cognitive models of each team member to track and remediate their performance

Spoken dialogue between virtual humans and human users is crucial to our goals. Unfortunately, the most sophisticated embodied conversational agents fall far short of real human spoken dialogue. The animated pedagogical agents of Lester (Lester, Stone et al., 1999) and his colleagues require the user to communicate with the agents through menus. Rea (Cassell, Bickmore et al., 2000) supports spoken dialogue, but does not have any sophisticated natural language understanding capabilities, so it is limited to understanding a small set of utterances that have been manually added to the speech recognition grammar and directly mapped to concepts the agent understands. The virtual humans developed by (Bindiganavale, Schuler et al., 2000) include sophisticated natural language understanding, but they have no capabilities for dialogue with users; they only accept instructions.

Another key area related to our goals is computational models of emotion. A person's emotional state influences their decision making, actions, memory, attention, and body language, all of which may subsequently impact their emotional state (Berkowitz, 2000). To model the behavior of teammates in stressful situations, as well as create virtual humans that can induce stress in the human user by reacting emotionally, our virtual humans include a believable model of emotions. Several researchers have experimented with emotions in animated agents (Ball and Breese, 2000; Poggi and Pelachaud, 2000) but these models of emotion fall far short of cutting edge work. In contrast, state of the art models of emotion, such as Gratch and Marsella's EMotion and Adaptation (EMA) model (Marsella and Gratch, 2003; Gratch and Marsella, 2004), have not previously been integrated with animated agents. A unique aspect of the work we are doing at the ICT is that EMA has a computational model of human coping. As a consequence, our virtual humans can both "feel" emotions as well as intentionally use them as a signal to manipulate others and their own beliefs. For example, it is possible for our virtual humans to feel guilty about some event but cope with the stress of that guilt by shifting blame to another agent and express anger at that other. The problem of synthesizing realistic behaviors for conversational characters has been addressed by

several researchers (Kopp et al, 2004, Cassell et al, 2001). Generally, the main approach taken has been to develop an animation architecture and then to fine tune it in order to meet with the behavioral requirements at hand. All of these works produce compelling results, however not addressing synchronization issues when different methods have to be combined in order to achieve composed behavior. In contrast, ICT uses a sophisticated generator of non-verbal behavior based on the agent's speech output (Lee at al, 2006) and a procedural animation system called SmartBody (Thiebaux et al, 2007), described later in the paper.

One last area to take into consideration is the advances being done in the game and film entertainment industry. The goals of the entertainment industry are similar to the goals of ICT's effort in creating believable virtual humans. Though these industries are not known for their research, they are known for trying new and creative ideas, especially in trying to create very realistic looking characters. Reviewing ideas and technology that they develop and incorporating those that make sense into our effort will keep us current with the latest trends and help ICT become a leader in creating virtual humans.

VIRTUAL HUMAN RESEARCH

Imagine a simulated military exercise where the characters you interact with are almost human – they converse with you in English, they understand the world they are in and can reason about what to do, and they exhibit emotions. Such a simulation could open up whole new horizons for training and simulation. Because virtual humans are intended to mimic a broad range of human behaviors and characters for these domains, they must integrate a diverse set of graphics, AI technologies, and domain knowledge.

The goal of ICT's Virtual Humans project is two fold, to perform advanced research in areas that lead to a fully realistic virtual human and to research technologies to enable virtual humans for training environments. These two goals complement each other; as the research gets mature it is transitioned into training applications. One set of questions we endeavor to answer is how realistic do the virtual humans need to be to be effective. How believable do they need to be? How much verbal and non-verbal behavior is needed? How many human capabilities are needed for training environments, and which capability should be used in each training application? These are not easy questions to answer by any means as they involve large integrated efforts and testing, but by building a fundamental base architecture we hope to answer some of these pressing issues.

Conceptually, the virtual humans should include three layers that make up the mind the agent thinks with, the body the agent acts with, and the world the agent interacts in, as seen in Figure 1.

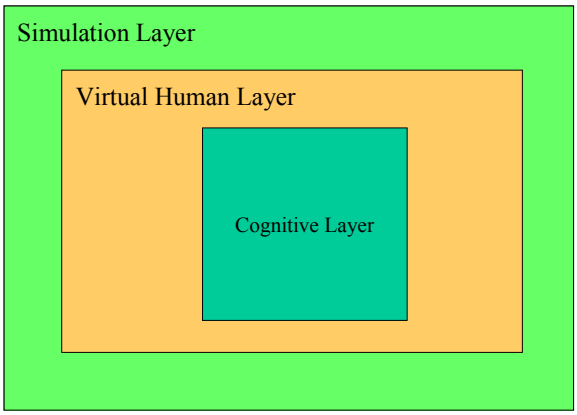

Figure 1: 3-Layered Virtual Human

These layers are represented as follows:

Cognitive Layer: The inner layer is where the cognitive component exists. There is usually one cognitive level per virtual human. This is the mind of the virtual human that makes decisions based on input, goals, and desired behavior. This layer does not necessarily need to include a full theory of cognition. To a varying degree of usefulness, question response systems, finite state machines, and even scripted agents can prove effective.

Virtual Human Layer: At this layer is the set of components that make up the virtual human, including input and output processing. Input could include vision, speech, and even smell. Output would include verbal speech, body gestures, and actions the character performs, for example walking. In our architecture this layer can be used by one or more of the virtual humans. For example, speech from a human can go to all agents in the system. The virtual human layer can be thought of as the body. This layer is closely connected to the cognitive layer in terms of information sharing, input/output and communication.

Simulation Layer: This is everything else that has to do with the environment that the virtual humans exist in. This would include the game engine that creates the world that the characters are displayed in and interact with, a world or social simulator, background characters, any scenario management interfaces, and any form of after action review. All input from the real world, like gesture recognition, object or human

positions, microphones or cameras would feed into the simulation layer.

Research Areas

Virtual humans follow the same paradigm as Belief-Desire-Intention (BDI) style agents with a sense-think-act cycle as seen in Figure 2. The cognitive and virtual human layers correspond to the right side of this diagram, while the simulation and real world correspond to the left side. The virtual human research attempts to integrate all of these layers and attempts to answer some fundamental questions about how they should be built, and how they can be effectively used in training environments.

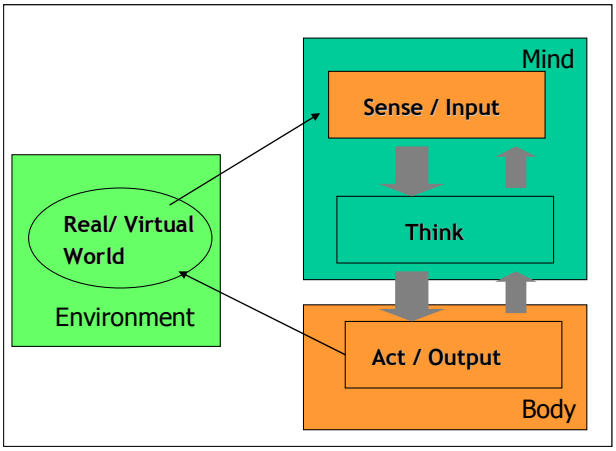

Figure 2: Sense-Think-Act Cycle

For example, users of the system would expect to have the character recognize their voice and respond in kind. They would expect them to respond appropriately with gestures and responses. They would expect them to display emotion and interact with the environment. There are focused research efforts that address each of these areas, but few large-scale projects bring all of the pieces together. Only when each of these pieces of technology are integrated together will larger system interaction issues be understood. These interactions include the interfaces, shared knowledge, and world state required by the components along with the speed and effectiveness of inputs and outputs. Once these components are integrated in a distributed fashion, then the system design can be evaluated and modified, which in turn might affect changes to the individual components. In addition, it is only within the context of a rich functioning system that one can seriously test integrated theories of human behavior (Swartout et al., 2004). More specifically, research in each of these areas includes:

Cognitive Layer (Mind):

The mind reasons on both internal triggers as well as input it receives from its body's senses. We believe that by applying knowledge in helping to understand this input will improve the system. One of the main research areas therefore is to seek out which information can and should be shared between modules and in what fashion. In addition, we want to research how the complex internal models of our virtual humans can best be visualized, both as stand alone tools as well as augmented within the virtual world. Below, we describe each research area:

Cognition and emotion modeling – Our research in this area emphasizes the close connection between cognition and emotion posited by current psychological and neuroscience findings. We also place a great emphasis on models of social (as opposed to individual) cognitive and affective processes, the detection, analysis, and integration of computer vision techniques for recognizing human emotional behavior and a growing maturation and sophistication in our methodologies for validating computational models of human behavior. We are modeling beliefs about self and others (Theory of Mind) and are working to extend our models to answer questions like how emotion arises from reason, how emotion impacts reason and physical behavior, how emotional displays of a virtual human influences the cognitive process of the trainee and, vice versa, how the emotional display of a trainee should influence the cognitive process of a virtual human.

Natural language (NL) processing and dialogue management – The overall theme in this area of research is to extend the agent's ability to understand and generate natural language and to achieve a tighter integration between all NL related components and the rest of the system. For example, we are investigating the use of hybrid NL approaches that combine statistical methods with symbolic processing. One of the goals we have is extending the agent's capability to converse about the environment it is inhabiting by making use of multi-modal input like speech features (e.g., question and emotion detection, etc), vision features (e.g., gestures, trainee position), world state information (e.g., objects and their location) and dialogue state information from the agent (e.g., current topics, goals). We are also researching how to better link verbal and non-verbal behavior, including developing extended functional markup of the reasons for textual choices to nonverbal generator, providing more detailed information about the course of performance to the text generator and dialogue manager, and investigating methods for making low-level text vs. gesture decisions. In addition, we want to

integrate the emotion and dialogue reasoning by extending the appraisal system to the dialogue state as well. Finally, we are extending our work on modeling how social relationships are established and maintained through conversation, in collaboration with researchers at the USC Marshall School of Business.

Knowledge representation (KR) – The integration of various data sources into our ontology is fairly recent and much remains to be done. Research in this area focuses on extending the framework so that any new additions will automatically trigger the appropriate internal learning mechanisms that allow the components to be able to learn from data to update themselves as appropriate. We are also researching ways to extend the depth, sophistication, and coverage of our KR capabilities, including semantic modeling of specified new phenomena for improved NLP and planning (e.g., modeling multi-sentence structures and modeling emotions that are linked to prosodic and lexical cues). In addition, we want to enable rapid growth of scenarios and porting to new scenarios by building and/or assembling libraries of background resources and creating sophisticated tools. One of our efforts is to look into building a generic engine to convert input text into symbolic representations in order to build training data.

Virtual Human Layer (Body):

The body contains the senses that provide data from the environment and that output actions as generated by the mind. Our research focuses on increasing the amount of available information and especially on how to best use and combine this information.

Speech recognition - The goal is to enable training capabilities that live-action simulations provide but in virtual environments. This in turn implies that these virtual environments, and the interactions within, are similar to those in live environments. We therefore focus on robust, large vocabulary speech recognition capabilities for multi-person, multilingual scenarios with noisy backgrounds and extracting rich information, beyond what is conveyed by just words, such as intonation, affect and higher level linguistic information such as speech acts. In addition, we are extending the core capabilities in processing the speech information in the larger context of human communication, notably the interplay with non-verbal information conveyed by face.

Non-verbal behavior sensing and recognition – Although current virtual humans can understand natural language, they have no ability to sense a user's gestures, posture, or facial expressions, thereby ignoring a significant source of information in face-to-face interactions and making it much more difficult for virtual humans to interact with users in a natural way. Correcting this limitation is a central theme in this area of research that will open up new communication channels between people and virtual humans and could significantly improve interaction. We collaborate with USC's Vision Lab, directed at creating virtual human characters and computers able to sense visual gestures and facial expressions from human participants.

Non-verbal behavior generation – Virtual humans are not responsive to the (social) environment in the ways that people are. A virtual human will walk up and down the stairs all day long without getting tired or becoming irritated. We know that in people there is a variety of mirroring and entrainment behaviors whereby for example if a person smiles, then others will smile. It is largely theorized that such responses are likely not rooted in high cognition functions in people but nevertheless have a major impact on social interaction (Kendon, 1967). More importantly, they are absent altogether in virtual humans. Further, these social, physical, and physiological responses are not uni-directional in humans. They in turn influence higher level cognitive and affective processes. In particular, they influence emotion states. However, this is not the case in virtual human designs.

A key goal of the virtual human embodiment research therefore is to realize this responsiveness, to realize low-level reactive capabilities in the virtual human's body, and to have those responses influence cognition and dialogue. Note that there are fundamental research questions here concerning how the virtual human's reactivity integrates with high level cognitions. Many of these issues have been of fundamental concern to AI planning and scheduling research and more fundamentally work in robotics. Understanding the relation between higher level cognition and lower level instinctual or associatively learned behaviors and body processes has also been a central area of study in the neural sciences. In addition, we are continuing our research on modeling key aspects of non-verbal behavior including gaze, facial expressions, gestures, and postures.

Simulation Layer (Environment):

A real life human is amongst all a social being and in order to successfully create a virtual human, the creation of its environment is vital. Our research focuses on creating this social environment in which our virtual humans can live.

Virtual worlds and simulation – Interactive virtual worlds populated with many virtual characters that

model interpersonal experiences, social dynamics and are story driven provide a powerful medium for experiential learning. Although our virtual worlds are inhabited by a number of background characters, their behavior is extremely limited and scripted. We are researching how to leverage some of the in-house developed light weight agents as background characters. The goal is to enrich our worlds with engaging believable characters that convey a rich inner dynamic that plays out in the emergent properties of a virtual world.

Real world integration – Our aim is to blur the dividing line between the real and the virtual world. The above mentioned recognition of gestures, the use of natural language and knowledge representation will allow real humans to point objects in the world and use them in their conversation with our virtual humans.

VIRTUAL HUMAN ARCHITECTURE

Interactive virtual training environments with virtual human characters can succeed only when various disparate technologies are integrated together. The integration effort on the Virtual Humans project is a primary distinguishing characteristic from other efforts in this area. The various technologies include everything that a person would expect to encounter while interacting with a real human in a real environment. The major components in the Virtual Human Architecture are shown in Figure 3 and described here in terms of the 3 layers mentioned above. The system is a set of modular distributed components that communicate with message passing.

Cognitive Layer

Intelligent Agent – This is the major reasoning engine of the agent, based on the Soar Cognitive Architecture. Soar is a symbolic reasoning system that includes concepts such as long term and short term memory, goal directed behavior and a decision procedure for selecting rules. This component uses a task model and planner to reason about what actions to take. It also includes a Dialogue Manager (DM), which reasons about the trainee's utterances (as processed by the Natural Language Understanding component) and interprets them in the context of past utterances and the utterances of other agents. The DM is also responsible for generating communication goals and their associated semantic representation. The agent also performs emotional modeling, appraising current events and the current situation and comparing them with the agent's beliefs and goals to determine emotional state and appropriate coping behavior.

Virtual Human Layer

Speech Recognition – This is based on the SONIC speech recognition engine from the University of Colorado, Boulder (Pellom, 2001). We customized the engine's acoustic and language models for the domain of interest (Sethy et al., 2005). A human user talks in plain English to the system using a close-capture microphone. The user's speech is converted into text by an automatic speech recognition system and sends it to the Natural Language Understanding (NLU) component.

Natural Language Understanding (NLU) – The NLU parses the text string produced by the speech recognition component and forms a semantic representation by matching it to semantic frames that are part of a framebank generated from an ontology for the domain. In addition to the core semantics, this frame also includes information like speech act and modality. The resultant frame is sent to be processed by the Dialogue Manager. We also have a system that can replace the agent and the NLU to perform response selection based on a statistical text classification approach (Leuski, 2006) that is used in many of the virtual human applications.

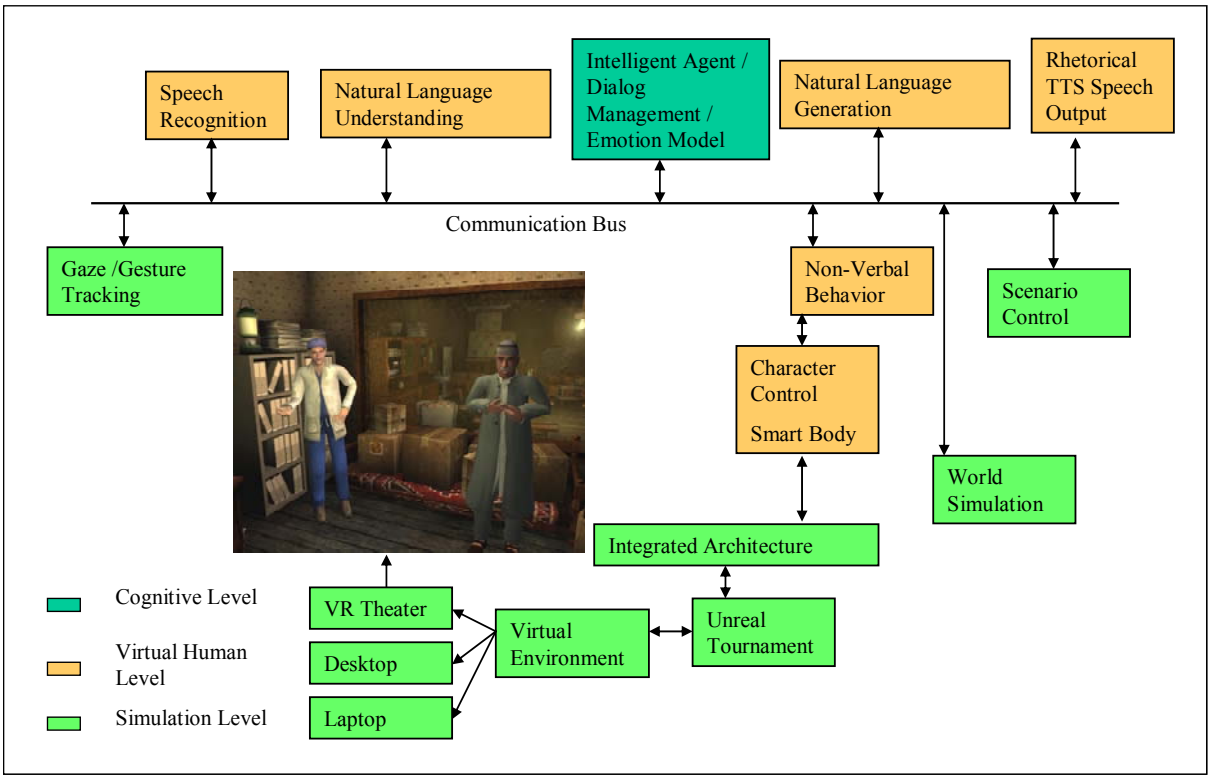

Figure 3: Virtual Human Integrated Architecture

Natural Language Generation (NLG) – The NLG component maps an internal semantic representation generated by the DM into a surface string. This can either be very much like the process as the NLU, but in reverse order, or based upon a domain dependent grammar. The resultant string is sent to the text to speech converter and the Non-Verbal Behavior component.

Non-Verbal Behavior Generator (NVBG) – Gestures and postures play a key role in realizing expressive, interpretable behavior in general and communicative intent specifically. For example, they qualify information in the speech such as a shrug when saying, "I don't know." They also emphasize important words by using, for example, a beat gesture (e.g., short chopping movement) synchronized with the word. The timing of gestures to the speech is critical, and small changes can alter an observer's interpretation of the utterance of the speaker. Without gestures, or with poorly timed gestures, a character will look unnatural. The NVBG (Lee et al., 2006) applies rules based on theoretical foundations of movement space to select the appropriate gesture animations, postures, facial expressions, and lip synch timing for the virtual character. Once the NVBG selects the appropriate behavior for the input text, it then packages this up into a Behavioral Markup Language (BML) (Kopp et al., 2006) structure and sends it to the procedural animation system, SmartBody.

SmartBody – SmartBody (Thiebaux et al, 2007) takes as input the BML message that contains the set of behaviors that need to be executed for the head, facial expressions, gaze, body movements, arm gestures, speech and lip syncing and synchronizes all of this together. It is capable of both using generated speech or pre-recorded speech. SmartBody controls the character in the game engine and also specifies which sound to play for the characters speech output. It is hooked up to a visualization engine, in this case Epic's Unreal Tournament game engine. Smartbody is also capable of having controllers that perform specific actions based on rules or timing information, such as head nods. The controllers are seamlessly blended in with the input animations specified in the BML. A motex, which is a looping animation file, can be played for the character to give it a bit of sway, finger tapping, or some repetitive movement.

Text to Speech – The voice generation capability is done by a commercial product called Rhetorical. The software performs speech synthesis from the text generated by the NLG.

Simulation Layer

Unreal Tournament and Integrated Architecture – This is the underlying graphics engine used for the virtual human project and associated applications at ICT. The engine is built on a common interface called the Integrated Architecture that allows the system to be run on desktops as well as the 160 degree Virtual Reality Theater at ICT. The engine is currently a bit out dated compared to the next generation game engines and plans are to upgrade and support both an open source engine and a new commercial engine.

Gaze / Gesture Tracking / Input System – This is the interface to hook up external gesture, head, and facial tracking software to provide real world input into the system. Currently some projects support the Watson Head tracker vision library from MIT, Intersense 3D space tracking and head/gesture detection software built as part of the virtual human project for use in the VR-Theater at ICT.

Tools – There are various tools in the system that gather data for system testing, logging, after action review or to simulate components, such as fake speech. One of the main tools we use to provide a common knowledge framework is Stanford's Protégé (Knublauch et al, 2004). Protégé allows us to build a specific domain ontology as an extension of our general world ontology, using OWL. In-house developed plug-ins and stand-alone applications interface Protégé with our intelligent agent and natural language components. As an ongoing effort, we continue to add components that make use of our ontology.

Architecture Principles

The Virtual Human Architecture is based on years of research in developing these systems. As in any large software engineering effort the system has seen many revisions as we add new components or learn new insight we adjust the architecture. We try to constrain the architecture to various principles:

- **Distributed** – This allows the various research groups to more easily perform work in their area; they can upgrade their component without affecting the rest of the system.
- **Multi-layered** – The system should be based on a set of layers that share information but may perform actions or decisions at various independent speeds. The lower layer should be more reactive while the higher layer should be more deliberative. Information sharing can be within a level or across several layers.
- **Cognitively and psychologically plausible** – The system should be based on sound theories and not just a set of integrated components. The main theory we hold is the idea of symbolic processing, i.e. things in the world can be represented as symbols and can be manipulation as such in the mind of the agent.

There are many approaches to creating a human-like architecture, and this is still an active area of research. One of the basic principles we strive towards is the integrative approach. Only when you integrate all the components together and place the virtual humans in an environment do you learn about the interactions of the components in the system and how to more effective leverage their capabilities in associated components.

VIRTUAL HUMAN APPLICATIONS

The virtual human technology has been applied to build both research prototypes and training applications. The virtual humans range from the more complex cognitive agents to question response agents. Because of the distributed nature of the architecture we are able to replace components without large integration efforts thus reducing the time it takes to build an application. Additionally, various projects have integrated certain components, for example Smartbody, into their application without using the full virtual human suite. Each application type has tradeoffs that will be discussed in the next section after a review of the applications.

MRE: *Mission Rehearsal Exercise*. This was the first immersive training research prototype that was developed at ICT that included several virtual humans in an integrated system. As a trainee captain your job was to lead and diffuse a situation where a military vehicle hit a boy (Rickel et al., 2001).

SASO-ST: *Stability and Support Operations – Simulation and Training*. A research prototype demonstrating advanced virtual human technologies in a new negotiation domain. Trainees are to communicate in real-time with an embodied virtual human doctor to negotiate and convince Dr. Perez to move the clinic out of harms way (Swartout et al., 2006). See Figure 4.

Figure 4: SASO-ST Negotiation Training in ICT's VR Theater

SASO-EN: *Stability and Support Operations – Extended Negotiations*. Based on SASO-ST, a new virtual human was added. In this research prototype, trainees are to conduct multilateral negotiations with a doctor and a village elder to move the clinic to another part of the town. See picture in Figure 3.

ELECT-Bilat: *Enhanced Learning Environment with Creative Technologies* is a game based simulation for soldiers to practice and conduct bilateral engagements in a cultural context that includes virtual humans that verbally respond to the selected questions. Uses a menu based system instead of speech recognition. See Figure 5.

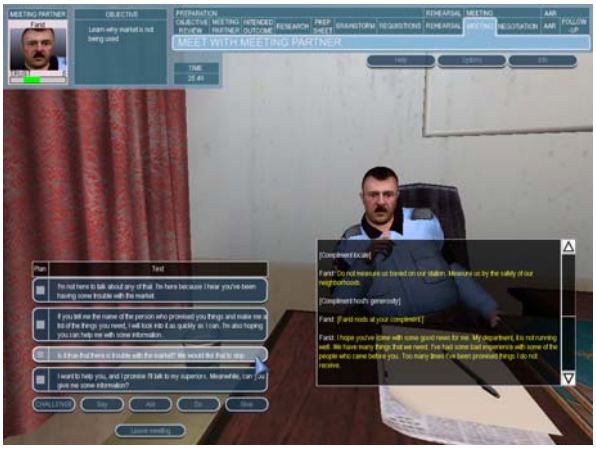

Figure 5: ELECT-Bilat

C3IT: *Cultural and Cognitive Combat Immersive Trainer.* Depicts a new class of immersive training, incorporating Mixed Reality simulation environments and virtual humans. Soldiers are placed in critical decision-making situations that are highly realistic and which require cultural awareness in order to make the best judgments. See Figure 6 and 7.

Figure 6: Tactical Question / Answering Agent in C3IT Immersive Environment

Tactical-Questioning: This virtual human uses a response selection system for verbal behavior. Trainees are to Interview a suspect about a bombing incident in a cultural aware and volatile situation. See Figure 6.

Sgt. Blackwell: This technology demonstration is a virtual human that uses a sophisticated question and response system, but does not contain a cognitive model, has been widely demonstrated, even at the Army Science Conference 2006 in front of a live audience of 200 people by Dr. John Parmentola, who considers Sgt. Blackwell a close friend of his.

Virtual Patient: This prototype application applies the virtual human technology to create a patient with conduct disorder for clinician interview training (Kenny et al., 2007).

Figure 7: C3IT Application Setting with Virtual Humans

EVALUATIONS AND DISCUSSIONS

To assess the value of the virtual humans work for training applications subject testing is constantly performed with the systems to evaluate it on several criteria. These criteria include; performance, interactivity, believability and feedback from users or trainees. One area of interest is on the use of speech as a natural interface. During subject testing we gather data on the speech interface, words the system could not recognize, utterances that were not recognized or understood by the agent and if the verbal and non-verbal behaviors were understood or seemed appropriate. We also ask testers to generally rate their interaction experience with the virtual human and system as a whole, and anything they would improve. Subject testing is performed in-house with cadets, interns, and ICT personnel. Other applications have been tested outside of ICT.

A prototype of the *C3IT* system was demonstrated in November 2006 at the Army Science Conference in Orlando, and in December at Ft. Benning USAIS to approximately 80 Soldiers and instructors. Soldiers acted as demonstrator trainees in those presentations. The demo includes a scenario called *Liar's Market* which involves an investigative questioning after an IED explosion in a marketplace. The Soldier demonstrators engaged with market vendor witnesses and two suspect characters projected at human scale on digital flat displays. Feedback surveys collected from Warfighters and trainers at Ft. Benning are due to be collected into a report in second quarter GFY07. The general feedback from them has been positive, with remarks like "much needed tool for soldier's first rotation to Iraq / Afghanistan" and "outstanding initiative […] unlimited potential to support valued training requirements." Criticism was mostly found in the limited repertoire of the virtual human and the relatively clean and game-like environment, a clear signal that we need to continue to improve our visualization. An often heard remark was that the addition of a translator would both make the scenario more realistic and more useful as soldiers hardly have a chance to train associated skills due to the lack of available translators. Interesting feedback was provided by suggesting using the application on laptops that could be deployed in the field. The inclusion of an after action review was valued by many participants. See Figure 6 and 7.

The ELECT-Bilat application that uses virtual humans as part of the negotiation engagement, has been used at Ft. Leavenworth at the school for command prep by over 20-30 colonel level soldiers. The general feedback has been good, and interaction with the characters adds to the engagement of the system. ARI is performing a more formal evaluation in terms of the learning objectives, pre and post evaluation of the system which is due out in GFY07. See Figure 5.

As for the SASO negotiating scenarios, we are continuously conducting subject testing with a mix of cadets and civilians and the overall was positive. Subjects enjoy the challenge of trying to negotiate to get the clinic moved. Cadets find it a great benefit applying their tactical training in the virtual environment. Although we notice the limitations the current state of the art virtual humans still posses, we see that trainees learn from their mistakes; often, a trainee will fail the negotiation a first time, but will be able to convince the doctor in a second or third try.

Table 1 gives an overview of average rating per metric out of 30 test subjects that tried the SASO-ST scenario. On a scale of 1 to 7, the pace of the conversation and the naturalness of interaction score a disappointing 3. Other metrics, like the ability to understand the virtual doctor, the satisfaction fo the experience and the overall success of the system in simulating a real-life experience are rated with a more positive 4 or 5. These numbers show room for improvement and provide a base line against which we can test future iterations or our system, including the SASO-EN scenario.

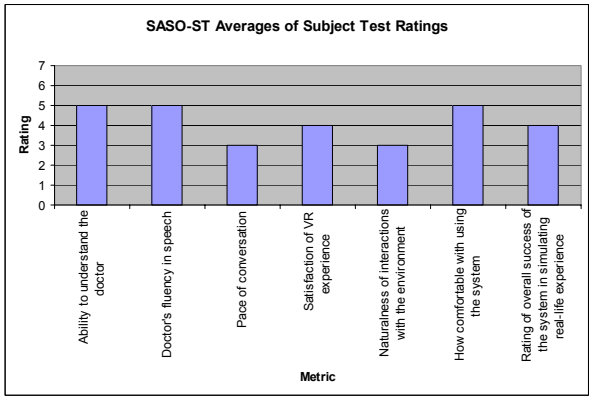

Table 1: SASO-ST Average of Subject Test Ratings

As mentioned earlier there are various trade offs and issues with the technology that need to be addressed. These virtual human systems can be built and fielded today, however it requires several experts to make them useful. The set of authoring tools to build domains required for wide adoption is still a few years away. Other issues include:

Artwork: The artwork of the characters, the environments, animations, any background music and interface screens all need to be built. This is an important part that many people under estimate the time and skill required for this effort. As game engines and technology change it's important to keep up to date with what the entertainment industry is doing.

Dialogue: It takes time to gather the large corpus of dialogue needed for an application; what will people say to the virtual human, how should it respond? This is currently done though role playing exercises, subject testing and wizard of Oz testing or general common sense. The question response systems we use are quite easy to get up and running, however it takes lots of testing to get it to respond to all the different queries people will ask. Use of advanced machine learning techniques would be of great help here.

Procedural system: We believe that systems of the future will need to be more intelligent and do work itself instead of having programmers, designers, or artists build every little piece. The goal of procedural systems, like Smartbody and the Non-verbal behavior component is to move the work into the system and let it choose for itself using rules based on psychologically plausible theories, the action selections for the behaviors of the character.

Domain and Agent building: Building the whole domain set of knowledge into the system is still a hard process. And there are few end user tools for this task. It still requires programmers. Leveraging and integrating existing tool sets could be valuable.

LESSONS LEARNED

As we continue to research, build and apply this technologies we are constantly learning new things to improve them. One major question is the ability for our systems to be effectively used as a training tool. While there have been many subject tests, there have been no formal evaluations to see if using virtual training environments are better than live role playing exercises. However, both informal feedback as well as formal questionnaires give us a general sense of how our technology is valued. Most people see the value of these training systems, either as a great potential or as an addition to existing training methods. This brings up the question how advance virtual humans need to be in order for human trainees to be able to engage with and learn from them. The ultimate goal is to create virtual humans that allow a trainee to have his own personal trainer and the ability to experience an endless variety of real as life training sessions as needed. This is definitely a story of the future, though, and might take decades to achieve. That is not to say we will have to wait that long to reap the benefits from our work. While advancing the state of the art in virtual human training environments, we are able to use existing technology to replace some forms of training or use them as an addition to existing programs. For example, we have found that many trainees find our training environments way more engaging than the currently used online courses.

As trainees' level of comfort and ease of use of this new technology becomes more widespread, so does the realism level of the character in terms of the dialogue, look and interactivity. One major factor in getting more wide use on these is the ability to design and author scenarios, to add training data to the system, and to more easily control the environment for after action review or asymmetrical training. As we are working with leading edge technology, setting up and creating a new scenario is no easy task. This prevents us from iterating quickly on the larger scope. It is a major

challenge to streamline our creation process in such a way that we can leverage the latest technology and models, and quickly try out new theories while not committing ourselves to a particular format.

CONCLUSIONS AND FUTURE WORK

This paper described the virtual human effort that is on-going at ICT. As is seen in this paper, virtual humans require large amounts of research in many areas and integrating all of this together is a grand effort. Developing a distributed architecture that is modular, and supports loose-coupling of components is valuable because all of the details and issues are never known at the beginning.

The virtual human effort has been transitioned into several applications in the last few years and as our understanding and tool development increases, the time and effort that is needed to create new scenarios is slowly declining. One of our goals is to get end users to develop scenarios in a matter of weeks not months as is currently the case.

The ultimate goal is to create fully realistic interactive characters that can remember you, know what training you require and contain a vast array of knowledge, tactics and training procedures. Although this goal might be decades away from achieving, the near terms goal will be achieved by integrating many component technologies. The project seeks to create functioning virtual human that will advance the state of the art in immersive training by facilitating face-to-face interactions between users and synthetic autonomous characters.

Creating virtual humans that are believable in their appearance, language and behavior, responsive to the user and simulation, and interpretable by the trainee will ultimately create compelling training environments that will train and transform the next generation Warfighter for novel interpersonal engagements..

ACKNOWLEDGEMENTS

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.

REFERENCES

Anderson, J. R. and C. Lebiere (1998). The Atomic Components of Thought. Mahwah, NJ, Lawrence Erlbaum Associates.

Ball, G. and J. Breese (2000). Emotion and Personality in a Conversational Character. Embodied Conversational Agents. J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Cambridge, MA, MIT Press: 189-219.

Becheiraz, P. and D. Thalmann (1996). A Model of Nonverbal Communication and Interpersonal Relationships Between Virtual Actors. Computer Animation, IEEE Press.

Berkowitz, L. (2000). Causes and Consequences of Feelings, Cambridge University Press.

Bindiganavale, R., W. Schuler, et al. (2000). Dynamically Altering Agent Behaviors Using Natural Language Instructions. Fourth International Conference on Autonomous Agents, Barcelona, Spain, ACM Press.

Brand, M. (1999). Voice puppetry. ACM SIGGRAPH, ACM Press/Addison-Wesley Publishing Co.

Cassell, J., T. Bickmore, et al. (2000). Human conversation as a system framework: Designing embodied conversational agents. Embodied Conversational Agents. J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Boston, MIT Press: 29-63.

Cassell, J., C. Pelachaud, et al. (1994). Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. ACM SIGGRAPH, Reading, MA, Addison-Wesley.

Cassell, J., Vilhj_almsson, H.H., Bickmore, T.W.: Beat: the behavior expression animation toolkit. In: Proceedings of SIGGRAPH. (2001) 477-486

Frijda, N. (1987). "Emotion, cognitive structure, and action tendency." Cognition and Emotion 1: 115-143.

Gratch, J., Rickel, J., André, E., Badler, N., Cassell, J., Petajan, E.: Creating Interactive Virtual Humans: Some Assembly Required, IEEE Intelligent Systems, July/August, 54-63, (2002)

Gratch, J. and S. Marsella (2004). "A domain independent framework for modeling emotion." Journal of Cognitive Systems Research 5(4): 269-306.

Gratch, J. and S. Marsella (2004). Evaluating the modeling and use of emotion in virtual humans. 3rd International Joint Conference on Autonomous Agents and Multiagent Systems, New York.

Kalra, P. and N. Magnenat-Thalmann (1998). "Real-time Animation of Realistic Virtual Humans." IEEE Computer Graphics and Applications 18(5): 42-55.

Kendon, A. Some functions of gaze direction in social interaction. Acta Psychologica, 26: 22-63, 1967.

Kenny, P., Parsons, T., Gratch J., Leuski, A., Rizzo A.: Virtual Patients for Clinical Therapist Skills Training. 7th International Conference on Intelligent Virtual Agents, pp 197-210, Paris France. (2007).

Knublauch, H., Fergerson, R. W., Noy, N. F., Musen, M. A. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. Third International Semantic Web Conference, Hiroshima, Japan, 2004.

Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds 15(1) (2004) 39-52.

Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thorisson, K., Vilhjalmsson, H: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language,. 6th International Conference on Intelligent Virtual Agents (Marina del Rey, CA, August 21-23 2006).

Laird, J. E. (2001). It Knows What You're Going To Do: Adding Anticipation to a Quakebot. Proceedings of the Fifth International Conference on Autonomous Agents, Montreal, Canada, ACM Press.

Lee, Y. and K. Waters (1995). "Realistic Modeling for Facial Animation." SIGGRAPH.

Lee, J., Marsella, S: Nonverbal Behavior Generator for Embodied Conversational Agents. 6th International Conference on Intelligent Virtual Agents, pp 243-255, Marina del Rey, CA. (2006).

Lester, J. C., B. A. Stone, et al. (1999). "Lifelike Pedagogical Agents for Mixed-Initiative Problem Solving in Constructivist Learning Environments." User Modeling and User-Adapted Instruction 9(1-2): 1-44.

Leuski, A., Patel, R., Traum, D., Kennedy B.: Building effective question answering characters. (2006) In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, Sydney, Australia.

Marsella, S. and J. Gratch (2003). Modeling coping behaviors in virtual humans: Don't worry, be happy. Second International Joint Conference on Autonomous Agents and Multi-agent Systems, Melbourne, Australia.

Marsella, S. & Johnson, W.L. (1998) Intelligent Tutoring Systems, Springer-Verlag. Lecture Notes in Computer Science

Pellom, B.: Sonic: The University of Colorado continuous speech recognizer. Technical Report TR-CSLR-2001-01, University of Colorado, Boulder, CO (2001)

Perlin, K. (1995). "Real Time Responsive Animation with Personality." IEEE Trans. on Visualization and Computer Graphics 1(1): 5-15.

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

Rickel, J., Gratch, J., Hill, R., Marsella, S., Swartout, W.: Steve Goes to Bosnia: Towards a New Generation of Virtual Humans for Interactive Experiences. In AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment, Stanford University, CA, March (2001)

Rousseau, D. and B. Hayes-Roth (1996). Personality in Synthetic Agents. Stanford, CA, Knowledge Systems Laboratory, Stanford University.

Sethy, A., Georgiou, P., Narayanan, S.: Building topic specific language models from webdata using competitive models. In: Proceedings of EUROSPEECH, Lisbon, Portugal (2005)

Swartout, W., Gratch, J., Hill, R., Hovy, E., Marsella, S., Rickel J., Traum, D.:Toward Virtual Humans, AI Magazine, v.27(1). (2006)

Thiebaux, M., Marshall, A., Marsella, S., Fast, E., Hill, A., Kallmann, M., Kenny, P., Lee, J., SmartBody: Behavior Realization for Embodied Conversational Agents, Submitted to IVA07, (2007), Paris, France.

Wiggins, J. S. (1996). The Five-Factor Model of Personality: Theoretical Perspectives. New York, The Guilford Press.

Zachary, W., Cannon-Bowers, J., Bilazarian, P., Krecker, D.,Lardieri, P., & Burns, J. (1999). The Advanced Embedded Training System (AETS): An intelligent embedded tutoring system for tactical team training. International Journal of Artificial Intelligence in Education, 10, 257-277.