Virtual Patients for Virtual Sick Call Medical Training

Patrick G. Kenny, Thomas D. Parsons
Institute for Creative Technologies / USC
Marina Del Rey, CA
{kenny,tparsons}@ict.usc.edu

Pat Garrity
U.S. Army Simulation and Training Technology Center
Orlando, FL
Pat.garrity@us.army.mil

ABSTRACT

Training military clinicians and physicians to treat Soldiers directly impacts their mental and physical health and may even affect their survival. Developing skills such as: patient interviewing, interpersonal interaction and diagnosis can be difficult and is severely lacking in hands-on-training due to the cost and availability of trained standardized patients. A solution to this problem is in using computer generated virtual patient avatars that exhibit the mental and physiologically accurate symptoms of their particular illness; such physical indicators as sweating, blushing and breathing due to discomfort and matching conversational dialog for the disorder. These avatars are highly interactive with speech recognition, natural language understanding, non-verbal behavior, facial expressions and conversational skills.

This paper will discuss the research, technology and the value of developing virtual patients. Previous work will be stated along with issues behind creating virtual characters and scenarios for the joint forces. It will then focus on subject testing that is being conducted with a Navy scenario at the Navy Independent Duty Corpsman (IDC) School at the Navy Medical Center in San Diego. The protocol involves pre and post tests with a 15 minute interview of the virtual patient. Analysis of the data will yield results in user interactions with the patient and discuss how the system can be used for training for future deployment of these systems for medical professionals.

The Virtual Sick Call Project under the Joint Medical Simulation Technology Integrated Product Team (JMST IPT) seeks to push the state of the art in developing high fidelity virtual patients that will enable the caregiver to improve interpersonal skills for scenarios that require not only medical experience, but the ability to relate at an interpersonal level, with interviewing and diagnosis skills as patients can be hiding symptoms of post-traumatic stress disorder, suicide and domestic violence.

ABOUT THE AUTHORS

Mr. Patrick Kenny has been researching and developing in the Artificial Intelligence field for the past fifteen years. He is currently runs the Virtual Patient Simulation Lab and is the principal investigator creating virtual human patients for research and training in medical applications at The Institute for Creative Technologies, a world known research lab part of The University of Southern California. He previously was the virtual human group lead integrating the virtual human architecture with cognitive and simulation systems. Mr. Kenny previously worked at The University of Michigan AI Lab researching and developing cognitive robotics for unmanned ground vehicles. While in Michigan Mr. Kenny founded a spin-off company, Soar Technology, from the AI Lab to support and develop cognitive agents in simulation. His research interests are in creating realistic high fidelity virtual humans with personality, dramatic agents and cognitive architectures integrating neuroscience and AI. Mr. Kenny has a BS from The University of Minnesota and an MS from The University of Michigan. Mr. Kenny has organized workshops on believable characters and virtual human toolkits. He is the author of over 30 publications has been on several panels and has served on various paper view committees.

Thomas D. Parsons, PhD is a Clinical and Experimental Neuropsychologist, Assistant Research Professor, and Research Scientist at the University of Southern California's Institute for Creative Technologies. He directs the NEUROSIM Laboratory, helping to facilitate human-computer interface research. His work with human-computer interfaces began with invasive brain-computer interfaces and the cognitive sequelae of deep brain stimulation. The

long-range goal of Dr. Parsons‘s laboratory is to develop noninvasive brain-computer interfaces and psychophysiologically adaptive virtual environments that may be used for neuropsychological assessment, stress inoculation, virtual reality exposure therapy, cognitive training, and rehabilitation. This goal is being pursued with a combination of theoretical and experimental approaches at several levels of investigation ranging from the biophysical level to the systems level. In addition to his patents (with eHarmony.com), he has over 100 publications in peer-reviewed journals and other fora.

Pat Garrity is a Chief Engineer at the Simulation and Training Technology Center (STTC) Human Dimension, Simulation and Training Directorate, Army Research Laboratory (ARL). He currently works in Dismounted Soldier Simulation Technologies conducting R&D in the area of dismounted soldier training & simulation where he was the Army's Science & Technology Manager for the Embedded Training for Dismounted Soldiers program. His current interests include Human-In-The-Loop (HITL) networked simulators, virtual and augmented reality, and immersive dismounted training applications. He earned his B.S. in Computer Engineering from the University of South Florida in 1985 and his M.S. in Simulation Systems from the University of Central Florida in 1994.

Virtual Patients for Virtual Sick Call Medical Training

Patrick G. Kenny, Thomas D. Parsons
Institute for Creative Technologies / USC
Marina Del Rey, CA
{kenny,tparsons}@ict.usc.edu

Pat Garrity
U.S. Army Simulation and Training Technology
Orlando, FL
Pat.garrity@us.army.mil

INTRODUCTION

There is a growing need to augment the current medical training environment with innovative and different learning techniques. The phrase "death by PowerPoint" is an all too common occurrence in class room settings and can lead to boredom and loss of focus. It is important that military clinicians and physicians get hands-on training to reinforce and learn new skills.
Current medical training resorts to using real people (usually hired actors, instructors or staff) acting as simulated patients with given medical conditions, which could be physical or psychological. These clients are often referred to as Standardized Patients (Triola, 2006). However the availability of actors and the kinds of conditions they can portray in a relatively standard way is an area of concern. It is hard to hire patients such as children or the elderly. The actors have to learn all of the clinical cases and there is high turn-over rate so the actors need constant training, which leads to low standardization. Additionally it is costly to hire that many actors on a continual basis.

A solution to address these problems is with new and advanced simulation technologies such as virtual reality, medical simulation, medical dummies and Virtual Patients (VP). One such technology coming out of the research stage is to use advanced interactive virtual characters that can be used as virtual standardized patients. These patients can be endowed with physical as well as mental attributes that will enable the student to interact with them and perform physical and mental examinations.

Virtual patients provide a powerful mechanism to practice interpersonal skills such as patient interviewing, rapport, question asking, and diagnosis from basic to advanced methods. They are available 24/7 and can perform data collection and real-time analysis of performance. The use of virtual patient technology is not meant to replace human standardized patients but augment the live actor program with virtual characters that can portray a multitude of conditions that might be difficult for actors to represent or repeat with success. Additionally, the use of virtual patients provide the ability to have a variety of characters available from elderly to young persons in different genders and cultures that students can practice with is a benefit.

There are many challenges in developing and applying VPs that can act as simulated patients. The characters need to act and carry on a dialog like a real patient with the specific mental or physical issues. There are challenges in adding realism and believability to the characters which requires considerable advancements in a number of Artificial Intelligence (AI) domains: speech recognition; natural language processing; non-verbal behavior animations; 3D computer graphics; cognition and physical process models. Further, there are a number of challenges (e.g., knowledge generation; behavior models, Artwork) that are inherent in attempts to develop virtual characters that act and look like human patients with specific mental or physical health problems. A related challenge is the application of these characters in a learning environment that aims to enhance appropriate skills. Finally this technology needs to be validated and standardized so it can be used across many disciplines to assess skill levels of students.

THE PROBLEM STATEMENT

This paper describes a research effort to develop virtual patients that can be used to train clinicians and mental health providers in patient interviewing, assessment and diagnosis skills. Specifically, this paper describes subject testing that was performed at the Independent Duty Corpsman school at the Navy Medical College in San Diego with a Navy virtual patient character suffering from Post-Traumatic Stress Disorder (PTSD). The purpose of this study was evaluative; can clinicians, medical students and instructors interact with a virtual patient system and ask appropriate questions to elicit a proper response from the system to perform an initial intake interview and possible diagnosis? Data was acquired from the users through questionnaires and an interview with the virtual patient. The data analyzed included demographics, the

questions the subjects asked to the VP and the responses the system gave.

BACKGROUND AND RELATED WORK

Virtual patients are artificially intelligent, virtual agents that have computer generated avatar bodies controlled by software that models human thought, human behavior and human action and can interact with users through speech and gesture in virtual environments (Gratch, et al., 2002, Kenny et al., 2007b). Advanced virtual humans are able to engage in rich conversations (Traum et al., 2008), recognize nonverbal cues (Morency et al., 2008), analyze social and emotional factors (Gratch & Marsella, 2004) and synthesize human conversation and nonverbal expressions which is synchronized together to produce realistic actions, facial expressions, lip syncing and gaze. (Thiebaux et al., 2008). Additionally, they can contain underlying physiological models that simulate blood pressure, heart rate, breathing, and blushing (DeMelo et al., 2009, 2010). Building virtual humans requires fundamental advances in AI, speech recognition, natural language understanding and generation, dialog management, cognitive modeling and reasoning, virtual human architectures and computer graphics and animations. All these technologies need to be integrated together into a single system that can work in unison, be expandable, flexible and plug-and-play with different components

Virtual patients fulfill the role of standardized patients by simulating a particular clinical presentation with a high degree of consistency and realism and offer a promising alternative (Deladismaet et al., 2008; Green et al., 2004; Hayes-Roth et al., 2004, 2009; Hubal et al., 2000; Kenny et al 2007a, 2008; Lok et al., 2006; Parson et al., 2008). There is a growing field of research that applies VPs to training and assessment of bioethics, basic patient communication, interactive conversations, history taking, and clinical assessments (Bickmore & Giorgino, 2006; Bickmore et al., 2007; Lok et al., 2006; Parsons et al., 2008; Johnsen et al., 2007). Results suggest that VPs can provide valid and reliable representations of live patients (Kenny et al., 2007; Triola et al., 2006; Andrew et al., 2007). Additionally, VPs enable a precise presentation and control of dynamic perceptual stimuli; along with conversational dialog and interactions, they have the potential to provide ecologically valid assessments that combine the veridical control and rigor of laboratory measures approaching a verisimilitude that reflects real life situations (Parsons et al., 2009; Andrew et al., 2007). Some groups have developed complex cognitive models of patients (Sergei et al., 2009).

VIRTUAL PATIENT TECHNOLOGY

Building virtual humans to be used as patients requires a large integrated effort with many components. These patients are 3D computer generated characters that act, think and look like real humans. The users can interact with them through multi-modal interfaces such as speech and vision. The components together form the virtual patient avatars that exhibit the mental and physiologically accurate symptoms of their particular illness; such physical indicators as sweating, blushing and breathing due to discomfort and matching conversational dialog for the disorder. These avatars are highly interactive with speech recognition, natural language understanding, non-verbal behavior, facial expressions and conversational skills.

The Virtual Sick Call Project under the Joint Medical Simulation Technology Integrated Product Team (JMST IPT) seeks to push the state of the art in researching and developing high fidelity virtual patients that will enable the caregiver to improve interpersonal skills for scenarios that require not only medical experience, but the ability to relate at an interpersonal level, with interviewing and diagnosis skills as patients can be hiding symptoms, for example, of post-traumatic stress disorder, suicide and domestic violence.

The general architecture supports a wide range of virtual humans from simple question / answering to more complex ones that contain cognitive and emotional models with goal oriented behavior(Kenny, 2007). The architecture is a modular distributed system with many components that communicate by message passing. Because the architecture is modular it is easy to add, replace or combine components as needed.

Research Issues in Virtual Patients

There are many research areas in developing virtual patient technology which are best done in an iterative process with subject testing to help inform the development and identify problem areas. The main research areas can be broken down into the following categories:

Speech Recognition

A human user talks to the system using a head-mounted, close-capture, Universal Serial Bus (USB) microphone. The user's speech is converted into text by an automatic speech recognition system. The Virtual Sick Call project makes use of the SONIC speech recognition engine from the University of Colorado, Boulder customized with acoustic and language models for the domain of interest (Pellom, 2001, Narayanan, 2004). In general a language model is tuned to the

domain word lexicon which helps improve performance; however, the speech recognition system is speaker independent and does not need to be trained beforehand. User's voice data is recorded during each testing sessions which allows collection of words not acknowledged by the speech recognizer and further enhance the lexicon. The speech recognition engine processes the audio data and produces a message with the surface text of the user's utterance that is sent to the question / response selection module.

Natural Language Question / Response Dialog

Dialog is an important part of interacting with VP characters since a clinical interview relies heavily on the types of questions asked and the responses from the patient. The current VP system uses a statistical classifier system called NPCEditor, for Non-Player-Character Editor, a term used in computer games to represent characters in the game that are not controlled by another human player (Leuski & Traum, 2010). The dialog and discourse involves two parts, the human speech input and the classifier of the input to responses for the output. The question / response selection module receives a surface text message from the speech recognition module, analyzes the text, and selects the most appropriate response. This response selection process is based on a statistical text classification approach developed by the natural language group at (Leuski et al, 2006). The approach requires a domain designer to provide some sample questions for each response. There is no limit to the number of answers or questions, but it is advised to have at least four to ten questions for each answer. When a user question is sent from the speech recognition module, the system uses the mapping between the answers and sample questions as a "dictionary" to "translate" the question into a representation of a "perfect" answer. It then compares that representation to all known text answers and selects the best match. This approach has been shown to outperform traditional state-of-the-art text classification techniques (Leuski et al, 2006).

Responses can be divided into several categories based on the type of action desired. Sometimes the system combines the text from answers of different categories to produce the final response. The only category required is on-topic responses, the others are optional, but make the system more interactive and realistic. The category types are as follows:

- On-topic – These are answers or responses that are relevant to the domain of the conversation. These are the answers the system has to produce when asked a relevant question. Each on-topic answer should have a few sample questions and single sample question can be linked to several answers. The text classifier generally returns a ranked list of answers and the system makes the final selection based on the rank of the answer and whether the answer has been used recently. That way if the user repeats his questions, he may get a different response from the system.
- Off-topic – These are answers for questions that do not have domain-relevant answers. They can be direct, e.g., "I do not know the answer", or evasive, e.g., "I will not tell you" or "Better ask somebody else". When the system cannot find a good on-topic answer for a question, it selects one of the off-topic lines.
- Repeat –If the classifier selects an answer tagged with this category, the system does not return that answer but replays the most recent response. Sample questions may include lines like "What was that?" or "Can you say that again?" Normally, there is at most one answer of this category in the domain answer set.
- Alternative – If the classifier selects an answer tagged with this category, the system attempts to find an alternative answer to the most recent question. It takes the ranked list of answers for the last question and selects the next available answer. Sample questions may include lines like "Do you have anything to add?" Normally, there is at most one answer tagged with this category in the answer set.
- Pre-repeat – Sometimes the system has to repeat an answer. For example, it happens when a user repeats a question and there is only one good response available. The system returns the same answer again but indicates that it is repeating itself by playing a pre-repeat-tagged line before the answer, e.g., "I told you already." There is no need to assign sample questions to these answer lines.
- Delayed – These are the lines from the system that prompt the user to ask about a domain related thing, e.g., "Why don't you ask me about…" Such a response is triggered if the user asks too many off-topic questions. The system would return an off-topic answer followed by a delayed-tagged answer. That way the system attempts to bring the conversation back into the known domain. This category has no sample questions assigned.

Once the output response is selected, it is packaged up into an FML (Functional Markup Language) message structure. FML allows the addition of elements such as affect, emphasis, turn management, or coping strategies. For the VP, the response selection module does not add any additional information besides the text.

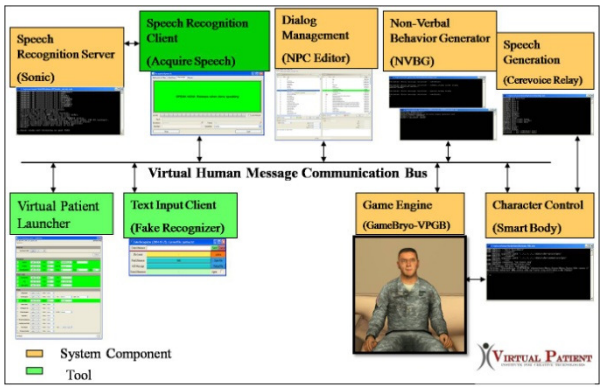

Figure 1: System Architecture Components

Non-Verbal Behavior

The FML message is sent to the Non-Verbal Behavior Generator (NVBG) which applies a set of rules to select gestures, postures and gazes for the virtual character (Lee and Marsella, 2006). Since the VP in this application was sitting down, the animations mainly consisted of arm movements, wave offs and head shakes or nods. The VP character does not do any posture shifts or go into standing posture, although this will be added for future scenarios. Once the NVBG selects the appropriate behavior for the input text, it then packages this up into a Behavioral Markup Language (BML) structure which adds timing information, animation information and lip synching information into the message and sends it to a procedural animation system for character action control.

Animation Response

The last part of the process is the execution and display of the characters' multimodal behavior accomplished through a procedural animation system called Smartbody (Thiebaux & Marsella, 2008). Smartbody inputs the BML message from the NVBG which contains the set of behaviors that need to be executed, synchronizing together the head, facial expressions, gaze, body movements, arm gestures, speech and lip syncing. These behaviors need to be in sync with the output speech to look realistic. Smartbody is capable of using generated or pre-recorded speech. The VP uses generated speech for this scenario. The graphics output is performed in the GameBryo commercial game engine. An added component integrated into the system is a breathing controller. The breathing controller synchronizes with Smartboday and performs breathing actions based on a simple breathing model. The model is capable of performing breathing patterns such as deep breathing, laughter, and sighs (DeMello, 2009, 2010). The controllers are seamlessly blended with the input animations specified in the BML. A motex, which is a looping animation file, is played for the character to give it a bit of sway or motion, or in the VP case, finger tapping for more realistic idle behavior that matches the scenario.

THE NAVY SCENARIO

A sick call in the military is when someone reports to the medical doctors instead of duty because of some medical problem. Military medical doctors, especially shipboard ones, have to manage multitudes of people along with many different cases ranging from common colds to trauma and mental issues such as suicide and PTSD. These medical personnel need a vast amount of training. Most students only have High school or AA schooling yet are required to learn all the skills a normal doctor does in a short period of time. They need as much help as they can get to learn and recognize the many cases they will have to deal with.

This current VP project aims to improve the interview skills and diagnostic acumen of psychiatry residents, military leaders and medical students. To this end a scenario was developed to be used as part of the subject testing with Navy medical personal at the Independent Duty Corpsmen (IDC) School at the Surface Warfare Medicine Institute at the Navy School of Health Sciences in San Diego. The IDC training program is an intense year long training regimen that strives to educate, train and develop medical personnel in support of force readiness and teaches all medical needs from basic first aid to tactical combat casualty and patient care including physical and mental clinical training.

The PTSD Scenario

War is one of the most challenging environments that people may experience. The cognitive, emotional and physical demands of combat environments place enormous stress on even the best-prepared military personnel. The current combat theatre, with its ubiquitous battlefronts, ambiguous enemy identification, and repeated extended deployments has resulted in a significant number of returning veterans with mental disorders. In a recent Rand (2008) report, eighteen percent of the 1.64 million returning veterans were found to have signs and symptoms of PTSD, major depression or Traumatic Brain Injury, resulting in them having a higher risk for developing familial and social problems, depression, alcohol and drug abuse, and suicide. Clinicians and social workers need to develop skills for identifying potential stress related disorders. It is common for returning veterans to lie about their problems, in most cases lie about not having them when they in fact do. Usually this is because the veteran is so eager to get home and not have to go through observation and delay the process.

Clinicians will need to be able to ask appropriate questions to recognize this.

The scenario is a Navy one and consists of a Petty Officer 2nd Class Samuel Sarax, See Figure 2, who came in because he has become progressively more isolated and withdrawn during the last few months and often appears tense and his friends are worried about him. He has seen combat before but only recently started to suffer issues due to a deadly friendly helicopter crash that he had to help clean up. Sarax does in fact have PTSD, however the subjects did not know this before testing. The dialog in the domain consisted of a set of 166 responses that the VP could say, the non-verbal behavior was meant to match with someone that is dissociative.
This Navy PTSD scenario has been adopted from a previous one in which a virtual patient was experiencing PTSD from a sexual assault (Kenny 2008). These two similar scenarios will enable an eventual comparison between the subjects of the questions, responses and performance.

Figure 2: Sarax Virtual Patient

The PTSD Domain
The domain of PTSD requires the system to respond appropriately based on certain criteria for PTSD as described in the Diagnostic and Statistical Manual of mental disorders (DSM) (309.81; DSM American Psychiatric Association, 2000). PTSD is divided into six major categories as described in the DSM-IV:

A. Past experience of a traumatic event and the response to the event.
B. Re-experiencing of the event with dreams, flashbacks and exposure to cues.
C. Persistent avoidance of trauma-related stimuli: thoughts, feelings, activities or places, and general numbing such as low affect and no sense of a future.
D. Persistent symptoms of anxiety or increased arousal such as hyper vigilance or jumpy, irritability, sleep difficulties or can't concentrate.
E. Duration of the disturbance, how long have they been experiencing this.
F. Effects on their life such as clinically significant distress or impairment in social or educational functioning or changes in mental states.

Diagnostic criteria for PTSD includes a history of exposure to a traumatic event in category A and meeting two criteria and symptoms from each B, C, and D. The duration of E is usually greater than one month and the effects on F can vary based on severity of the trauma. Effective interviewing skills are a core competency for the clinicians, residents and developing psychotherapists who will be working with children and adolescents exposed to trauma. Rather than assessing for all of the specific criteria, the focus was on the major clusters of symptoms following a traumatic event. Two additional categories that would aid in assessing user questions and VP responses that are not included in the DSM were developed:

G. A general category meant to cover questions regarding establishing rapport, establishing relations, clarifications, opening and closing dialog.
H. Another category to cover accidental mouse presses with no text, the user is required to press the mouse button while talking.

A clinician needs to ask questions in each of these categories to properly assess the patient's condition. Table 1 is and example of the types of questions that a typical subject asked in each of the DSM categories and the kinds of responses that the virtual patient would convey.
An example interchange between a clinician and the VP for the Navy scenario is as follows:

Clinician: Hello Petty Officer Sarax, I'm Doctor Cowne, ah, good to meet you. What brings you in today?
Sarax: No.
Clinician: (deep breath) um..(tic) You don't want to talk, or, what, can you tell me what brought you in today?
Sarax: When we got to the scene of the Helicopter crash, the smell of blood and death made me nauseous even before I saw the wreckage, but it is the images of the slain Seals that will never leave my mind.
Clinician: Ok so it sounds like you've been through a very traumatic incident and, um, you came in because you need some help coping with that.
Sarax: Yes.

Clinician: Ok, well I'm sorry to hear you went through something like that. Um, can you tell me a little more about it?
Sarax: What would you like me to tell you?
Clinician: Well how long ago did this happen?
Sarax: A few weeks ago.
Clinician: Are you having, flash backs about this?
Sarax: Yes.
Clinician: Is this keeping you awake at night, or affecting your sleep?
Sarax: Yes.
Clinician: How are you sleeping?
Sarax: I have nightmares almost every night. I come upon the wreckage of some vehicle and I see the dead eyes of corpses all looking at me in horror. I can't stand it.
In this interchange the clinician covered questions in the DSM categories of: G,A,B,D,E

Table 1. Question / Response Categories for PTSD

DSM	Question / Response Categories	
	User Questions	**Patient Response**
(A) Trauma	Do you have any other shipmates that went through this with you?	I have witnessed some terrible things and I think it is really starting to affect me.
(B) Re-experience Event	Do you still think about what happened?	I feel like this is happening all over again
(C) Avoidance	Are you avoiding certain things?	I just stay away from everyone now.
(D) Arousal	Do you feel jumpy?	I feel like I have to watch my back all the time.
(E) Duration	How long has this been going on?	A few months
(F) Life Effect	Have you been able too, keep going to work?	My friends think I stay in my quarters too much and I look sad..
(G) Rapport	What kind of music do you like?	I like classic rock.
(H) Other	Button Press	I don't get what you mean.

SUBJECT TESTING

Subject testing of the VP system was conducted with the Sarax Character and IDC students and instructors. The testing was not an evaluation of their skills, but an evaluation of the technology and its application to medical simulation. See Figure 3. There were 14 subjects, 2 female and 12 males. Ave age was 32.6, schooling varied from AA, High school to MD and BS/BA. There were 6 instructors and staff, 7 students and 1 resident.

Protocol
The protocol consisted of the subjects filling out a set of pre-questionnaires, followed by a 15 minute interview with the VP, followed by a set of post questionnaires. System data and log files, speech data and the discourse interaction were also gathered for evaluation. See Figure 4. The following questionnaires were administrated.

Measures:
NEO-FFI (NEO). The NEO-FFI is a Five Factor Personality Inventory of statements that the user marks to agree or disagree in a 5-point scale. This measures personality based on the Big 5 personality dimensions of Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness. There are 60 questions in the short form.

Sarax Pre-Test (SPQ1). This scale was developed to establish basic competence for interaction with a virtual character that is intended to be presented as one with PTSD, although no mention of PTSD is on the test. There are 10 questions.

VP Pre-questionnaire (VPQ1). This scale was developed to gather basic demographics and ask questions related to the user's openness to the environment and virtual reality user's perception of the technology and how well they think the performance will be. There were five questions regarding the technology and how well they thought they might perform with the agent.

Tellegen Absorption Scale (TAS). The TAS questionnaire aims to measure the subject's openness to absorbing and self-altering experiences. The TAS is a 34-item measure of absorption.

Immersive tendencies questionnaire (ITQ). The ITQ measures individual differences in the tendencies of persons to experience "presence" in an immersive Virtual Environment (VE). The majority of the items relate to a person's involvement in common activities. While some items measure immersive tendencies directly, others assess respondents' current fitness or alertness, and others emphasize the user's ability to focus or redirect his or her attention. The ITQ is comprised of 18 items, each is rated on a 7-point scale.

After the 15 min interview the following were administered.

VP Post-questionnaire. (VPQ2). This scale was developed to survey the user's perceptions related to their experience of the VE in general and the interaction with the virtual character, particularly the patient in terms of its condition, verbal and non-verbal behavior, how well the system understood them and if they could express their needs to the patient. Additionally, there were questions on the interaction and if they found it frustrating or satisfying. There are 25 questions for this form.

Sarax Post-Test. (SPQ2) This was the same as the pre test with the same questions but in a different order. It was meant to measure what they learned from the interaction and if their scores improved.

Presence Questionnaire (PQ). The PQ is a common measure of presence in immersive virtual reality. Presence has been described as comprising three particular characteristics: sense of being within the VE; extent that the VE becomes the dominant reality for users; and extent to which users view the VE as a place they experienced rather than simply images they observed.

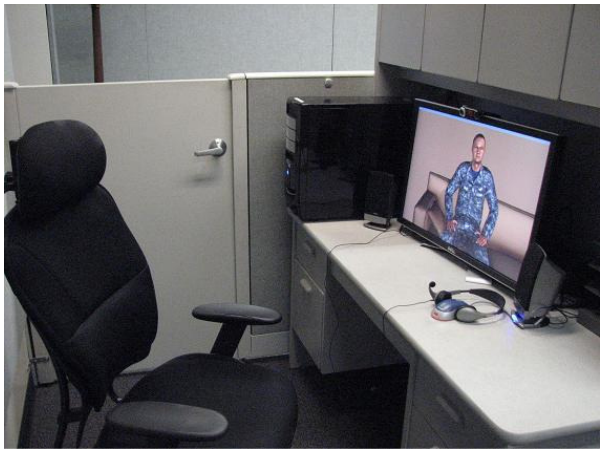

Figure 3: Subject Testing Setup

Procedures:

The data in the system was logged at various points to be processed later. Figure 4 is a diagram of how the user interacts with the VP system and the data logging and annotation pipeline. There are four areas where the data is logged. 1) The user speech data was recorded from what s/he says; this allowed transcription of what the speech engine processed. 2) A transcript of the entire dialog of the questions asked and responses given for the session was recorded. 3) System logs were stored to allow reconstruction, if needed, of what happened in the system during the test run, this included both the verbal and non-verbal behavior of the character with associated user questions. 4) Cameras recorded participant's facial expressions and system interaction with the patient to be analyzed at a later time for subjects non-verbal behavior patterns.

After the testing the set of questions from the user and responses from the VP were classified into one of the DSM categories mentioned above. This allowed for the study the responses and the questions asked by the subjects to see if they covered all the DSM categories.

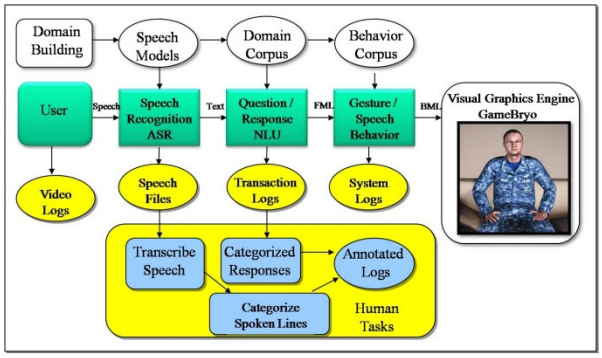

Figure 4: System Flow and Data Logs

DATA ANALYSIS AND RESULTS

Assessment of the system was completed with the data gathered. The focus was on effective interview skills—a core competency for clinicians and clinicians in training. The keys aspects of the interview looked at were: interpersonal interaction; attention to the VP's vocal communications, as well as verbal and non-verbal behavior. Specifically, the goal was to assess whether the clinician established and maintained rapport, as well as ask questions related to the reason for referral. Additionally, assessment as to whether the user made attempts to gather information about the VP's problems. Finally, investigation to see if the user would attempt detailed inquiry to gain specific and detailed information from the VP, separating relevant from irrelevant information.

VP Question/Response Composite

Question/response composites were developed to reflect the shared relation existing between the responses of the VP and of DSM IV TR-specific Questions (from users) that are necessary for differential diagnosis. The question/response composites drawn from user questions and VP responses were referred to as VP Question/Response composites or (VP_QR'). Again, the primary goal in this study was evaluative and the VP_QR' scores were calculated to assess whether a virtual standardized patient could generate responses that elicit user questions relevant for diagnostic categorization. For the VP_QR' scores, we first calculated effect sizes via least squares procedures and separate composite measures were created for each observation. The

resulting weights were used in conjunction with the original variable values to calculate each observation's score. The VP_QR' scores were standardized.

To assess whether the responses of the VP could elicit a number of DSM IV TR-specific questions (from users) that are necessary for differential diagnosis, our data analysis involved the development of a reference distribution representing effect sizes relevant to each cluster of questions and response pairs (from the users) making up a particular DSM PTSD Category. Each (corresponding) cluster of responses from the VP representing the same DSM PTSD Category.

The present focus is on effect sizes indicating strength of correlation, that is, effect sizes that describe the strength of association between question and response pairs for a given diagnostic category (see Table 2). Given our small sample size, we wanted a more conservative estimate of effect. Hence, an effect size (herein we use "r" as a standard of effect size) of 0.20 was regarded as a small effect, 0.50 as a moderate effect, and 0.80 as a large effect. Strong effects existed between User Questions and VP Response pairs for Category A (Trauma: r = 0.85), and Category G (Rapport: r = 0.75). Moderate effects existed between User Questions and VP Response pairs for Category E (Duration: r = 0.47), and Category F (Life Effect: r = 0.53). Small effects were found for Category C (Avoidance: r = 0.27). Of note, there were only negligible effect findings related to Category B (Re-experiencing: r=.13) and Category D (Arousal: r=0.02).

Table 2: Comparison of Questions and Responses

DSM	Questions Mean (SD)	Responses Mean (SD)	Effect Size
(A)	8.21 (5.82)	2.86 (1.79)	0.85
(B)	1.36 (1.28)	5.36 (2.82)	0.13
(C)	1.86 (1.46)	12.43 (6.97)	0.27
(D)	3.71 (2.33)	4.43 (2.62)	0.02
(E)	1.64 (1.34)	2.86 (1.51)	0.47
(F)	9.50 (4.07)	4.43 (1.50)	0.53
(G)	24.64 (11.35)	21.50 (6.69)	0.75

Note: For all analyses N=14. SD=Standard Deviation. (A) Trauma; (B) Re-experience; (C) Avoidance; (D) Arousal; (E) Duration; (F) Life Effect; (G) Rapport.

The assessment of the system was completed with the data gathered from the log files in addition to the questionnaires to evaluate the number and types of questions being asked. Figure 5 is a graph showing that for the 15 minute interview the 14 subjects asked on average, 54.14 questions, lighter color, and responses, 53.86, by the system, darker color for each of the 8 categories from Table 1.

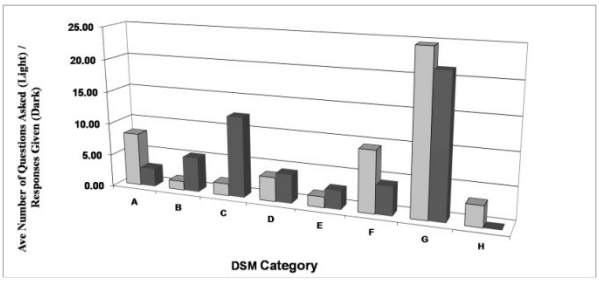

Figure 5: IDC Questions / Responses,

Questions – Light, Responses - Dark

Table 3 are values from the post questionnaires measuring user opinion on a 7 point Likert scale of the believability, understandability (Users were able to understand the patient) and frustration with the system. Some people rated the system at somewhat frustrating to talk to, due to speech recognition problems of out of domain questions or inappropriate responses or low recognition rates. However, most participants left favorable comments, (e.g. they thought the technology will be useful; they enjoyed the experience and trying different ways to talk to the character; and when the patient responded back appropriately to a question they found that very satisfying).

Table 3. Example of a Single Column Table

	7-Point Likert Scale, 1-Low, 7-High			
	Believable	**Understandable**	**Frustration**	
Value	4.4	4.1	5.3	

CONCLUSIONS

In this paper, there was a discussion of methods advanced technologies (i.e., Virtual Patients) can move beyond traditional approaches to training clinicians in assessment, diagnosis, interviewing and interpersonal communication. The traditional approaches rely upon a combination of classroom learning and role-playing

with human patients. Much of this work is done with actors recruited and trained to exhibit characteristics of an actual patient, thereby affording novice clinicians a realistic opportunity to practice and be evaluated in a mock clinical environment. Although a valuable training approach, there are limitations with the use of human patients that can be mitigated through simulation technology. For example, human patients are expensive and cost several thousand dollars per student. Further, given the fact that there are only a few sites providing standardized patient assessments as part of the U.S. Medical Licensing Examination, the current model provides limited availability.

In addition to issues of availability of trained actors, there is the issue of standardization. Despite the expense of standardized patient programs they are typically unskilled actors. As a result of common turnover, administrators face considerable challenges for offering psychometrically reliable and valid interactions with the training clinicians. The limited scope that the actors are able to portray tends to be an inadequate array of developmentally, socially, and culturally appropriate scenarios. For example, when a clinician has a pediatric focus and needs access to children, it is difficult for the clinician to pretend that the actor is a child. Finally, many clinical cases (e.g., traumatic brain injury) have associated physical symptoms and behaviors (e.g., dilated pupils, spasms, and uncoordinated movements) that simply cannot be accurately portrayed by human actors.

Findings suggest that the interactions between clinicians and the VP resulted in a compatible dialectic in terms of rapport (Category G), discussion of the traumatic event (Category A). Further, there appears to be a pretty good amount of discussion related to the experience of a negative impact on the VP's life (Category F) and the duration of symptoms (Category E). These results comport well with what one may expect from the VP system. Much of the focus was upon developing a lexicon that, at minimum, emphasized a VP that had recently experienced a traumatic event (Category A), the negative experience secondary to the trauma (Category F), the duration (Category E), and an establishment of rapport (Category G). However, the interaction is not very strong when one turns to the important DSM diagnostic information. For example, only small effects were found for avoidance (Category C). Further, there were almost no effects found for the very important diagnostic categories of re-experiencing (Category B); and hyper-arousal (Category D).

Although the low effect size for re-experiencing (Category B; r=0.13); and hyper-arousal (Category D: r=0.02) may represent a potential limitation in the system lexicon, it is important to compare the means and standard deviations of the re-experiencing (Category B: mean=1.36; SD=1.28); and hyper-arousal (Category D: mean= 3.71; SD=2.33) questions with the means and standard deviations of the re-experiencing (Category B: mean=5.36; SD=2.82); and hyper-arousal (Category D: mean=4.43; SD=2.62) responses. A general review of these descriptions seem to reveal that while the VP system was giving adequate responses for re-experiencing (Category B); and hyper-arousal (Category D), the users were only asking a limited number of questions related to re-experiencing (Category B); and hyper-arousal (Category D). That being said, it is important to note that this is a small sample size and a more fully developed comparison of the differences between these questions and response pairs should be performed in later studies.

In summary, effective interview skills are a core competency for training clinicians. Although schools commonly make use of standardized patients to teach interview skills, the diversity of the scenarios standardized patients can characterize is limited. Virtual standardized patient technology has evolved to a point where researchers may begin developing mental health applications that make use of virtual human patients for training.

FUTURE WORK AND RECOMMENDATAIONS

This research presented an approach that allowed novice mental health clinicians to conduct an interview with a virtual character in a military setting that emulated a male Navy personnel with trauma exposure. The work presented here builds on previous initial pilot testing of VPs and is a more rigorous attempt to understand how to build and use virtual humans as VPs along with the issues involved in building military domains and dialog, the speech and language models and working with domain experts.

The field of creating conversational VPs is an emerging area that requires more advanced speech and language understanding and dialog systems that are tied to the underlying physical and mental models. Additionally, developing tools to assist in dialog and scenario building is of a great need. Other multi-modal input, such as cameras and bio-physiological monitoring will allow for powerful sensing of the user and give the characters additional data to use in the interaction that will increase the realism. Adding more physiological characteristics and models to the character that emulate accurate blood flow, breathing, sweating and emotions will enhance the realism. Data from the additional questionnaires such as immersion and presence and personality will be used to evaluate the users performance based on these attributes.

Previous results suggested that people with high immersion tend to do better with this technology and user personality traits also play a role (Kenny, 2008).

LESSONS LEARNED

The experiment yielded the following lessons learned:

- The system needs to be able to handle a deeper conversation path by answering follow on questions more appropriately.
- Speech recognition remains a problem but can be improved by training users first and also by adding collected data into the models in an iterative process.
- Confusing or wrong systems response can sometimes led to confusing follow-on questions.
- The character suffers from no initiative or assertiveness and only answers questions and does not ask them.
- Designing domains is still complicated and is not a science, there are no standards.
- A Rubric for evaluating VP's along several dimensions of fidelity would be a good asset.
- Automate the classification of the user speech into the DSM categories would allow a real-time assessment level based against other peoples performances.

ACKNOWLEDGEMENTS

This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM) STTC, and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. We also wish to thank Mr. George Frausto at the IDC School.

REFERENCES

Andrew, R., Johnsen, K., Dickerson, R., Lok, B., Cohen, M., Stevens, A., Bernard, T., Oxendine, C., Wagner, P., Lind, S. (2006). Comparing Interpersonal Interactions with a Virtual Human to those with a Real Human. IEEE Transactions on Visualization and Computer Graphics.

Bernard, T., Stevens, A., Wagner, P., Bernard, N., Schumacher, L., Johnsen, K., Dickerson, R., Raij, A., Lok, B., Duerson, M., Cohen, M., Lind, D.S. (2006). A Multi-Institutional Pilot Study to Evaluate the Use of Virtual Patients to Teach Health Professions Students History-Taking and Communication Skills. Proceedings of the Society of Medical Simulation Meeting.

Bickmore, T., Pfeifer, L., Paasche-Orlow, M. (2007). Health Document Explanation by Virtual Agents. Proceedings of the Intelligent Virtual Agents Conference, Paris.

Bickmore, T., Giorgino, T. (2006). Health Dialog Systems for Patients and Consumers. Journal of Biomedical Informatics, 39(5): 556-571.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., and Yan, H. (1998). An Architecture for Embodied Conversational Characters. Proceedings of the First Workshop on Embodied Conversational Characters, October 12-15, Tahoe City, California.

De Melo C., Gratch J. (2009). Expression of Emotions using Wrinkles, Blushing, Sweating and Tears. Intelligent Virtual Agents Conference, Amsterdam, Sep 14-16

De Melo C., Kenny P., Gratch J. (2010). The Influence of Autonomic Signals on Perception of Emotions in Embodied Agents. Special Issue of the Journal of Applied Artificial Intelligence.

Deladisma AM., Johnsen K., Raij A., Rossen B., Kotranza A., Kalapurakal M., Szlam S., Bittner JG 4th, Swinison D., Lok B., Lind DS. (2008). Medical student satisfaction using a virtual patient system to learn history-taking communication skills. Stud Health Technology Inform.;132:101-5.

DSM, American Psychiatric Association (2000). (DSM-IV-TR) Diagnostic and statistical manual of mental disorders, 4th edition, text revision. Washington, DC: American Psychiatric Press, Inc.

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

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

Green, N., Lawton, W., and Davis, B. (2004). An Assistive Conversation Skills Training System for Caregivers of Persons with Alzheimer's Disease. In Proceedings of the AAAI 2004 Fall Symposium on Dialogue Systems for Health Communication.

Hayes-Roth B., Amano K., Saker R., and Sephton T. (2004). Training Brief Intervention with a Virtual Coach and Patients. In BK. Wiederhold, G. Riva Annual Review of CyberTherapy and Telemedicine, 2:85-96.

Hayes-Roth B., Saker R., Amano K. (2009). Automating Brief Intervention Training with Individualized Coaching and Role-Play Practice. Methods Med Informatics, in rev.

Hubal, R.C., Kizakevich, P.N., Guinn, C.I., Merino, K.D., & West, S.L. (2000). The Virtual Standardized Patient-Simulated Patient-Practitioner Dialogue for Patient Interview Training. In J.D. Westwood, H.M.

Hoffman, G.T. Mogel, R.A. Robb, & D. Stredney (Eds.), Envisioning Healing: Interactive Technology and the Patient-Practitioner Dialogue. IOS Press: Amsterdam.

Hubal, R., Fishbein, D., Paschall, M., (2004). Lessons Learned using Responsive Virtual Humans for Assessing interaction Skills. Proceedings of I/ITSEC.

Johnsen, K., Raij, A., Stevens, A., Lind, D., Lok, B. (2007). The Validity of a Virtual Human Experience for Interpersonal Skills Education. Proceedings of the SIGCHI conference on Human Factors in Computing Systems, ACM Press, New York, NY, 1049-1058.

Kenny, P., Parsons, T.D., Gratch, J., Leuski, A., and Rizzo, A.A. (2007a). Virtual Patients for Clinical Therapist Skills Training. Intelligent Virtual Agent Conference, LNAI 4722, 197-210

Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsella, S., Piepol, D. (2007b). Building Interactive Virtual Humans for Training Environments. Proceedings of I/ITSEC. Nominated for Best Paper.

Kenny, P., Parsons, T.D., Gratch, J., & Rizzo, A.A. (2008). Evaluation of Justina: A Virtual Patient with PTSD. Lecture Notes in Artificial Intelligence, 5208, 394–408.

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

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

Leuski A., Traum, D. (2010). NPCEditor: A tool for building question-answering characters. In Proceedings of The Seventh International Conference on Language Resources and Evaluation.

Lok, B., Rick F., Andrew R., Kyle J., Robert D., Jade C., Stevens A., Lind, D.S., (2006). Applying Virtual Reality in Medical Communication Education: Current Findings and Potential Teaching and Learning Benefits of Immersive Virtual Patients. Virtual Reality, 10:185-195.

Morency, L.-P., & de Kok, I. (2008). Context-based Recognition during Human Interactions: Automatic Feature Selection and Encoding Dictionary. 10th International Conference on Multimodal Interfaces, Chania, Greece, IEEE.

Narayanan, S., and Alwan, A., editors. (2004). Text to Speech Synthesis: New paradigms and advances. Prentice Hall.

Parsons, T.D., Kenny, P., Ntuen, C., Pataki, C.S., Pato, M., Rizzo, A.A., St-George, C., Sugar, J. (2008). Objective Structured Clinical Interview Training using a Virtual Human Patient. Studies in Health Technology and Informatics, 132, 357-362.

Parsons, T.D., Bowerly, T., Buckwalter, J.G., & Rizzo, A.A. (2007). A controlled clinical comparison of attention performance in children with ADHD in a virtual reality classroom compared to standard neuropsychological methods. Child Neuropsychology.

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

RAND Report MG720. (2008). Invisible Wounds of War: Psychological and Cognitive Injuries, Their Consequences, and Services to Assist Recovery http://rand.org/pubs/monographs/2008/RAND_MG720.pdf

Rizzo, A.A., Pair, J., Graap, K., Treskunov, A. & Parsons, T.D. (2006). User-Centered Design Driven Development of a VR Therapy Application for Iraq War Combat-Related Post Traumatic Stress Disorder. Proceedings of the 2006 International Conference on Disability, Virtual Reality and Associated Technology, 113-122.

Stevens, A., Hernandex, J., Johnsen, K., R. Dickerson, Raij, A., Harrison, C., DiPietro, M., Allen, B., Ferdig, R., Foti, S. (2005). The use of virtual patients to teach medical students communication skills. The American Journal of Surgery, Volume 191, Issue 6, Pages 806-811

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

Thiebaux, M., Marshall, A., Marsella, S., and Kallmann, M. (2008). SmartBody: Behavior Realization for Embodied Conversational Agents Proceedings of Autonomous Agents and Multi-Agent Systems (AAMAS). Portugal.

Traum, D., J. Gratch, et al. (2008). Multi-party, Multi-issue, Multi-strategy Negotiation for Multi-modal Virtual Agents. 8th International Conference on Intelligent Virtual Agents. Tokyo, Japan, Springer

Traum, D., Roque, A., Leuski, A., Georgiou, P., Gerten, J., Martinovski, B., Narayanan, S., Robinson, S., and Vaswani, A. (2007). Hassan: A virtual human for tactical questioning. In Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 71-74, Antwerp, Belgium.

Triola, M., Feldman, H., Kalet, A.L., Zabar, S., Kachur, E.K., Gillespie, C. (2006). A randomized trial of teaching clinical skills using virtual and live standardized patients. Journal of General Internal Medicine, 21, 424-429.