Virtual Human Patients for Training of Clinical Interview and Communication Skills

TD Parsons, P Kenny, AA Rizzo

Institute for Creative Technologies, University of Southern California,
Los Angeles, California, USA,

tparsons@usc.edu, *arizzo@usc.edu, http://vrpsych.ict.usc.edu/*

ABSTRACT

Although schools commonly make use of standardized patients to teach interview skills, the diversity of the scenarios standardized patients can characterize is limited by availability of human actors. Virtual Human Agent technology has evolved to a point where researchers may begin developing mental health applications that make use of virtual reality patients. The work presented here is a preliminary attempt at what we believe to be a large application area. Herein we describe an ongoing study of our virtual patients. We present an approach that allows novice mental health clinicians to conduct an interview with virtual character that emulates 1) an adolescent male with conduct disorder; and 2) an adolescent female who has recently been physically traumatized.

1. INTRODUCTION

Although there are a number of perspectives on what constitutes trauma exposure in children and adolescents, there is a general consensus amongst clinicians and researchers that this is a substantial social problem. The effects of trauma exposure manifest themselves in a wide range of symptoms: anxiety, post-traumatic stress disorder, fear, and various behavior problems. Trauma exposure is associated with increased risk of psychological problems in adulthood. Effective interview skills are a core competency for the clinicians who will be working with children and adolescents exposed to trauma.

Developing effective interviewing skills for the clinicians, residents and psychotherapists who will be working with children and adolescents exposed to trauma is a necessary skill. A clinician needs to ask various questions relating to the trauma and its effect to properly assess the patient's condition. Current therapeutic training systems resort to using real persons (hired actors or resident students) acting as standardized patients to portray patients with a given mental health problem in what is called an Objective Structured Clinical Examination (OSCE). The problem portrayed by the actor could be physical or psychological. Although schools commonly make use of standardized patients to teach interview skills, the diversity of the scenarios standardized patients can characterize is limited by availability of human actors and their skills. This is an even greater problem when the actor needs to be an adolescent. The potential of using computer generated virtual humans as standardized virtual patients (VPs) for use in clinical assessments, interviewing and diagnosis training is becoming recognized as the technology advances (Bernard et al., 2006; Bickmore, Pfeifer, & Paasche-Orlow, 2007). These VPs are embodied interactive agents who are designed to simulate a particular clinical presentation of a patient with a high degree of consistency and realism (Kenny et al., 2007). VPs have commonly been used to teach bedside competencies of bioethics, basic patient communication, interactive conversations, history taking, and clinical decision making (Bickmore, & Giorgino, 2006).VPs can provide valid, reliable, and applicable representations of live patients (Triola et al., 2006). Research into the use of VPs in psychotherapy training is in its nascent stages (Johnson et al., 2007; Parsons et al., 2008). Since virtual humans and virtual environments can allow for precise presentation and control of conversations and interactions, they can provide ecologically valid assessments that combine the control and rigor of laboratory measures with a verisimilitude that reflects real life situations.

The current project aims to improve child and adolescent psychiatry residents, and medical students' interview skills and diagnostic acumen through practice with a female adolescent virtual human with post-

traumatic stress disorder (PTSD). This interaction with a virtual patient provides a context where immediate feedback can be provided regarding trainees' interviewing skills in terms of psychiatric knowledge, sensitivity, and effectiveness. Use of a natural language-capable virtual character is beneficial in providing trainees with exposure to psychiatric diagnoses (e.g. PTSD), prevalent in their live patient populations, and believed to be under-diagnosed due to difficulty in eliciting pertinent information. Virtual reality patient paradigms, therefore, will provide a unique and important format in which to teach and refine trainees' interview skills and psychiatric knowledge. In order to be effective, virtual humans must be able to interact in a 3D virtual world, must have the ability to react to dialogues with human-like emotions, and be able to converse in a realistic manner. The combination of these capabilities allows them to serve as unique training tools whose special knowledge and reactions can be continually fed back to trainees. The goal of this virtual patient was to focus on a character with PTSD, our previous effort was on a character with Conduct Disorder. The eventual goal is to build a library of characters with a variety of psychiatric diagnoses to train residents and students at multiple levels.

2. Method

2.1 Participants

Participants were asked to take part in a study of novice clinicians interacting with a VP system. They were not told what kind of condition the VP had if any. Two recruitment methods were used: poster advertisements on the university medical campus; and email advertisement and classroom recruitment to students and staff. A total of 15 people (6 females, 9 males; mean age = 29.80, SD 3.67) took part in the study. Ethnicity distribution was as follows: Caucasian = 67%; Indian = 13%; and Asian = 20%. The subject pool was made up of three groups: 1) Medical students (N=7); 2) Psychiatry Residents (N=4); 3) Psychiatry Fellows (N=4). For participation in the study, students were able to forgo certain medical round time with the time spent in the interview and questionnaires.

2.2 Measures

Virtual Patient Pre-Questionnaire. 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.

Virtual Patient Post-questionnaire. This scale was exactly the same as the Virtual Patient Pre-questionnaire and will be used in the future for norming of a pre-post assessment of learning across multiple interactions with the VP. In the future we will also include social presence and rapport scales and include a control set that will just go thru a fixed script with the interview.

Justina Pre-questionnaire. We developed this scale 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 5 questions regarding the technology and how well they thought they might perform with the agent.

Justina Post-questionnaire. We developed this scale to survey the user's perceptions related to their experience of the virtual environment in general and experience interacting with the virtual character in particular the patient in terms of it's condition, verbal and non-verbal behavior and how well the system understood them and if they could express what they wanted to the patient. Additionally there were questions on the interaction and if they found it frustrating or satisfying. There were 25 questions for this form.

2.2 Procedures

One of the challenges of building complex interactive VPs that can act as simulated patients has been in enabling the characters to act and carry on a dialog like a real patient with the specific mental issues present for that condition in the domain of interest. Additional issues involve the breadth and depth of expertise required in the psychological domain to generate the relevant material for the character and dialog. The current domain of PTSD requires the system to respond appropriately based on certain criteria for PTSD as described in the DSM manual (309.81; American Psychiatric Association, 2000). According to the

most recent revision to the American Psychiatric Association's DSM Disorders, PTSD is divided into six major categories (see DSM for a full description and subcategories):

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 include 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. Rather than assessing for all of the specific criteria, we focused upon the major clusters of symptoms following a traumatic event. Next, we developed two additional categories that we felt would aid in assessing user questions and VP responses that are not included in the DSM:

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 or something that does not fit into the other categories.

Table 1 is an example of questions and responses from Justina for each category. Once the responses were establish, we used a voice actor to record the voice for Justina to be used by the system.

Table 1. Question / Response Categorization.

Category	User Question	Justina Response
1(A) Trauma	So, what happened to you that night?	Something really bad happened.
2(B) Re-experience	Do you still think about what happened?	Sometimes I feel like the attack is happening all over again
3(C) Avoidance	Do you go out with your friends?	I just stay away from everyone now.
4(D) Arousal	Do you feel jumpy?	I feel like I have to watch my back all the time.
5(E) Duration	How long has this been going on?	A few months
6(F) Life Effect	Are you upset?	Sometimes I don't do anything but stay in my room and cry.
7(G) Communication	Hi Justina, I'm Doctor..	Hello
8(H) Other	Button Press	I don't get what you mean.

For the PTSD domain we built an adolescent girl character called Justina, see Figure 1. Justina has been the victim of an assault and shows signs of PTSD. The technology used for the system is based on the virtual human technology developed at USC (Kenny et al., 2007; Swartout et al., 2006).

Figure 1: Justina Virtual Patient

The data in the system was logged at various points to be processed later. Figure 2 is a diagram of how the user interacts with the VP system and the data logging and annotation pipeline. The user speech is recorded from what s/he says; this lets us transcribe what the speech engine processes. Next the speech recognition client sends the recognized text to the statistical question/response system. Once an appropriate response is selected a behavior is generated for the character based on the response, the resulting behavior animation is shown in the graphical engine.

A transcript of the entire dialog session is recorded along with the system logs. This data allows us to reconstruct what happened in the system if needed. Cameras recorded participant's facial expressions and system interaction with the patient to be analyzed at a later time. The set of responses Justina would say were classified into one of the DSM categories from above. This allowed us to assess the responses of the system to questions asked by the subjects.

Figure 2: Testing setup and interaction

The subject testing was divided into three phases, a pre-test and pre-questionnaire, the interview and a post-questionnaire. The pre-test and pre-questionnaire were performed in a separate room from the interview and took about 10 minutes.

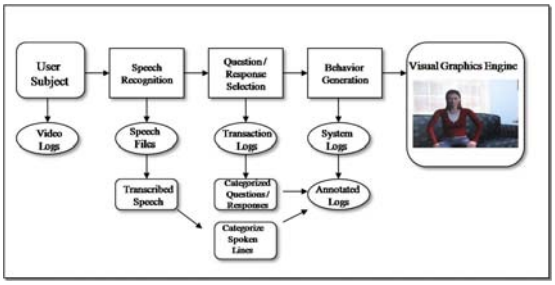

Figure 2: Interaction and Data Logging Pipeline

For the interview the participants were asked to conduct a 15 minute interaction with the VP and assess any history or initial diagnosis of a condition of the patient. The participants were asked to talk normally as they would to a standardized patient actor, but were informed that the system uses speech

recognition and was a research prototype. They were free to ask any kind of question and the system would try to respond appropriately. At the end of the 15 minute exchange they would be sent to another room to take the post-questionnaire. Assessment of the system was completed by the data gathered from the log files of the participants as they communicated with the VP in addition to the questionnaires. The log files allowed us to evaluate the amount and types of questions that the subjects were asking, along with a measure to see if the system was responding to the questions. After the subject testing sessions the set of questions were manually classified into one of the DSM categories.

3. Results

Assessment of the system was completed by the data gathered from the log files of the participants as they communicated with the VP in addition to the questionnaires. The log files allowed us to evaluate the amount and types of questions that the subjects were asking, along with a measure to see if the system was responding to the questions. For a 15 minute interview the participant asked on average, 68.6 questions with the minimum being 45 and the maximum being 91. Figure 4 is a graph showing the distribution of the questions that were classified into the 8 categories for each of the participants. It is interesting to note that most of the questions asked were either general questions (Category #G, 362 questions) or questions about the Trauma (Category #A, 200 questions), followed by category #C, 126 and #B, 123. The larger number of questions asked in #G was partially due to clarification questions, however we did not break down the category further to try to classify this. It is also interesting to note that the distribution of questions in each category for each participant were roughly equivalent. Which means in general people asked the same kinds of questions, maybe due to the fact that they have all had the same training.

There are several areas in the system that can be problematic due to technological issues which would cause the system to either generate an inappropriate response for a question an out of domain question, a question that the system did not know about. This was a particular issue, based on the questionnaires, is where the subjects got frustrated. There was one response that Justina would utter if she did not understand the question, due to out of domain or speech recognition deficiencies. This response was the "I don't get what you mean" utterance. This response was said in total 418 times across all subjects, comparing that to the total questions asked, 1029, the ratio was one in every 2.5 questions asked would yield an inappropriate response. While there is no standard for a good the question/response ratio, it at least gives us a measure as to how well the system was performing. Future analysis on the speech recognition word error rate and accuracy and decomposition of the question vs. the response choice would yield more data as to where the problems are arising. This analysis is part of the future work. If the system was performing bad for a subject, as it was for Subject #2, then this particular response was more prevalent. It is clear from the transcriptions that the domain we built was not sufficient to capture all of the questions people were asking, the results from this study will be added to the domain for future testing. The way people asked questions varied largely; there were many different styles and personality factors that influenced the length and type of question. And there are many novice assumptions by the subjects in how well this technology performs. Natural language and speech recognition is still a hard problem.

From the post questionnaires on a 7 point likert scale, the average people rated the believability of the system to be 4.5 and people were able to understand the patient, 5.1. People rated the system at 5.3 as frustrating to talk to, due to speech recognition problems, out of domain questions or inappropriate responses. However most of the participants left favorable comments that they thought this technology will be useful, they enjoyed the experience and trying different ways to talk to the character and trying to get an emotional response for a difficult question. When the patient responded back appropriately to a question they found that very satisfying.

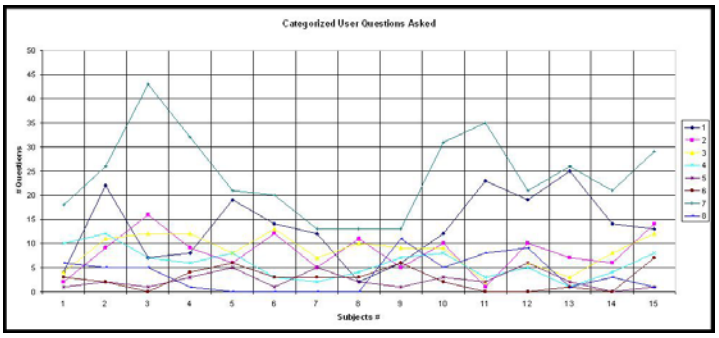

Figure 4: Number of questions for each category for each subject

For the assessment of student questions phase of the analysis, we aimed at investigating the relationship between a number of psychological variables and the resulting VP Responses. A summary of relations between each 1) DSM PTSD Category cluster of user questions; and 2) each (corresponding) cluster of responses from the VP representing the same DSM PTSD Category.

Following standard convention, an effect size of 0.20 was regarded as a small effect, 0.50 as a moderate effect, and 0.80 as a large effect. Moderate effects existed between User Questions and VP Response pairs for Category A (r = 0.45), Category B (r = 0.55), Category C (r = 0.35), and Category G (r = 0.56), but only small effects were found for Category D (r = 0.13) and Category F (r = 0.13). After controlling for the effects of the Tellegen Absorption Scale, increased effects were found for Category A (r = 0.48), Category C (r = 0.37), Category D (r = 0.15), and Category F (r = 0.24).

We also assessed impact of psychological characteristics such as absorption and immersiveness upon the "believability" of the VP and Student interaction. To assess this relation we created a composite variable that included scores from the TAS and the ITQ (Trait Composite). Strong effects existed between the Trait Composite and the Presence Questionnaire (r = 0.78), and moderate effects existed between the Trait Composite and the Justina Post-questionnaire (r = 0.40).

6. CONCLUSIONS

The primary goal in this study was evaluative: can a virtual standardized patient generate responses that elicit user questions relevant for PTSD categorization? Findings suggest that the interactions between novice clinicians and the VP resulted in a compatible dialectic in terms of rapport (Category G), discussion of the traumatic event (Category A), and the experience of intrusive recollections (Category B). Further, there appears to be a pretty good amount of discussion related to the issue of avoidance (Category C). These results comport well with what one may expect from the VP (Justina) 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) and was attempting to avoid (Category B) experienced that my lead to intrusive recollections (Category C). However, the interaction is not very strong when one turns to the issue of hyper-arousal (Category D) and impact on social life (Category F). While the issue of impact on social life (Category F) may simply reflect that we wanted to limit each question/response relation to only one category (hence, it may have been assigned to avoidance instead of social functioning), the lack of questions and responses related to hyper-arousal and duration of the illness (Category E) reflects a potential limitation in the system lexicon. These areas are not necessarily negatives for the system as a whole. Instead, they should be viewed as potential deficits in the systems lexicon.

A secondary goal was to investigate the impact of psychological variables upon the VP Question/Response composites and the genreal believability of the system. After controlling for the effects of these psychological variables, increased effects were found for discussion of the traumatic event (Category A), avoidance (Category C), hyper-arousal (Category D), and impact on social life (Category F). Further, the impact of psychological characteristics revealed strong effects upon presence and believability. These findings are consistent with other findings suggesting that hypnotizability, as defined by the applied measures, appears moderate user reaction. Future studies should make use of physiological data correlated with measures of immersion to augment and quantify the effects of virtual human scenarios.

Herein we described an ongoing study of our Virtual Patient System. We presented an approach that allows novice mental health clinicians to conduct an interview with a virtual character that emulates an adolescent female with trauma exposure. The work presented here builds on previous initial pilot testing of virtual patients and is a more rigorous attempt to understand how to build and use virtual humans as virtual patients and the many issues involved in building domains, speech and language models and working with domain experts. The lessons learned here can be applied across any domain that needs to build large integrated systems for virtual humans. We believe this is a large and needed application area, but it's a small enough domain that we can perform some serious evaluations on using virtual humans in real settings.

We will continue to perform more rigorous subject testing with both professional medical students and with non experts to evaluate how well the different populations perform in the types of questions asked. Additionally further studies in comparing to real OSCE's with real actors to the virtual patient will be performed. Additional incorporation of rapport [7,12] using the facial gestures analysis with the system will further enhance the virtual patient interaction to produce more results in this domain.

Additional analyses that need to be performed with the data include: investigate the domains questions and responses to assess how many were on-topic and how many off topic; How well did the speech recognition perform based on word error rate; How did the speech recognition, graphics and non-verbal impact the subjects, their interview experience, presence and immersion in the system?; Can we automate the process of extracting data from large corpus of speech data in this domain to build topic areas?; Can we automate the process of classifying the subjects questions into the DSM categories from the speech or transcriptions of the speech? Define further sub-categories for interactive conversions, such as; opening, closing, empathy, topic area, follow-up, query, clarification, self-disclosure to name a few and annotate the transcriptions with these categories. This will help us to build better tools to build domains and characters.

People have many interviewing and personality styles, some people are more direct, while others more empathetic. The system needs to be able to recognize these and accommodate it responses. Develop methods to recognize frustration levels, such as picking it up from the speech, as people interact with the system and adjust. Areas to improve the system include; the non-verbal behavior generation rules of the character and how to better tie that into the patient's condition; More autonomous behaviors for the character including; assertiveness, initiative, history tracking, topic tracking, and even have pre-programmed times that the patient will start to reveal information to enhance the training. By incorporating learning objectives into the interview session and investigating wither the virtual patient system has a learning impact is something that is valuable and will be the focus of future subject testing.

It is our belief that with more questions covered in the domain the accuracy of the system will go up along with the depth of the conversions which will further enhance the virtual patient system. In order to be effective virtual humans must be able to interact in a 3D virtual world, must have the ability to react to dialogues with human-like emotions, and be able to converse in a realistic manner with behaviors and facial expressions. The combination of these capabilities allows them to serve as unique training and learning tools whose special knowledge and reactions can be continually fed back to trainees. Our initial goal of this study was to focus on a VP with PTSD, but a similar strategy could be applied to teaching a broad variety of psychiatric diagnoses to trainees at every level from medical students, to psychiatry residents, to child and adolescent psychiatry residents.

8. REFERENCES

Bernard T, Stevens A, Wagner P, Bernard N, Schumacher L, Johnsen K, Dickerson R, Raij A, Lok B, Duerson M, Cohen M, Lind DS (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., and Paasche-Orlow, M. (2007). Health Document Explanation by Virtual Agents. *Proceedings of the Intelligent Virtual Agents Conference, Paris.*

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

Johnsen, K., Raij, A., Stevens, A., D. Lind and B. Lok (2007). The Validity of a Virtual Human Experience for Interpersonal Skills Education. In *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., & Rizzo, A.A. (2007). Virtual Patients for Clinical Therapist Skills Training. *Lecture Notes in Artificial Intelligence 4722,* 197-210.

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

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.

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

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.

ICDVRAT 2008 & ArtAbilitation 2008

Publication Agreement and Assignment of Copyright

Agreement: We are pleased to have the privilege of publishing your article in the forthcoming 7th International Conference on Disability, Virtual Reality and Associated Technologies 2008 and ArtAbilitation 2008 (collectively "ICDVRAT/AA"). By submission of your paper, you hereby grant to the ICDVRAT/AA all your right, title, and interest including copyright in and to the paper as it appears in the Proceedings of the ICDVRAT/AA ("the Paper"). Management of the copyright for all papers will be maintained by ICDVRAT.

Rights Reserved by Author(s): You hereby retain and reserve for yourself a non-exclusive license: 1.) to photocopy the Paper for your use in your own teaching activities; and 2.) to publish the Paper, or permit it to be published, as part of any book you may write, or in any anthology of which you are an editor, in which your Paper is included or which expands or elaborates on the Paper, unless the anthology is drawn primarily from ICDVRAT/AA. As a condition of reserving this right, you agree that ICDVRAT/AA will be given first publication credit, and proper copyright notice will be displayed on the work (both on the work as a whole and, where applicable, on the Article as well) whenever such publication occurs.

Rights of ICDVRAT: This agreement means that ICDVRAT/AA will have the following exclusive rights among others: 1.) to license abstracts, quotations, extracts, reprints and/or translations of the work for publication; 2.) to license reprints of the Paper to third persons for educational photocopying; 3.) to license others to create abstracts of the Paper; 4.) to license secondary publishers to reproduce the Paper in print, microform, or any computer readable form including electronic on-line databases. This includes licensing the Paper for inclusion in an anthology from ICDVRAT/AA 2008.

Warranties: You warrant that the Paper has not been published before in any form, that you have made no license or other transfer to anyone with respect to your copyright in it, and that you are its sole author(s), and generally that you have the right to make the grants you make to ICDVRAT/AA. Any exceptions are to be noted below. You also warrant that the Paper does not libel anyone, invade anyone's privacy, infringe anyone's copyright, or otherwise violate any statutory or common law right of anyone. You agree to indemnify ICDVRAT/AA against any claim or action alleging facts which, if true, constitute a breach of any of the foregoing warranties.

Concerning U.S. Government Employees: Some of the foregoing grants and warranties will not apply if the Paper was written by U.S. Government employees acting within the scope of their employment. U.S. Government employees may reserve the right to reproduce the Paper for U.S. Government purposes by making a request at the time of submission of the Paper. If no copyright can be asserted in this work and it should be considered in the public domain, the ICDVRAT/AA should be notified at the time of submission of the Paper.

In Conclusion: This is the entire agreement between you and ICDVRAT/AA and it may only be modified in writing. It will bind and benefit our respective successors in interest, including assignees, and our licenses, provided that you may not assign this agreement without our prior written consent. It will terminate if we do not publish your article in ICDVRAT/AA 2008.