Initial Evaluation of a Virtual Pediatric Patient System

Toni Bloodworth, Lauren Cairco, Jerome McClendon, Larry F. Hodges, Sabarish Babu, Nancy K. Meehan, Arlene Johnson
Clemson University
Clemson, SC 29634
{tbloodw, Lcairco, jmcclen, lfh, sbabu, nmeehan, aejohns}@clemson.edu

Amy C. Ulinski
University of Wyoming
Laramie, WY 307-766 USA
amyulinski@uwyo.edu

ABSTRACT

The most commonly used technique for teaching student nurses patient interviewing skills is reenacting written scenarios with classmates. Unfortunately, this is far from simulating the real world experiences that they will soon encounter. The Virtual Pediatric Patient System is designed to help baccalaureate nursing students prepare for real patient interactions by allowing them to practice interviewing skills with virtual characters. In this paper we describe our system and report on a usability evaluation conducted with experienced nursing faculty.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences – *health.* I.6.3 [**Computer Methodologies**]: Simulation and Modeling – *applications.* I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism – *virtual reality.* K.3.1 [**Computers and Education**]: Computer Uses in Education.

General Terms

Design, Experimentation, Human Factors.

Keywords

Virtual Patient, Nurse Education, Speech Interaction, Simulations.

1. MOTIVATION AND BACKGROUND

Nursing students have limited opportunities for interaction with real patients, especially pediatric patients [1]. Experiential learning through simulation provides students with a standardized experience in which they problem solve, develop clinical decision making skills, and use critical thinking [1]. In order to provide appropriate educational opportunities, we are developing the Virtual Pediatric Patient System (VPPS), which allows nursing students to interact with virtual characters acting as patients. This system has the potential to provide a consistent experience through repetitive practice and immediate feedback. In our initial prototype our scenario involves a mother and child. Children are a vulnerable population and it is essential that nurses practice patient interviewing skills prior to interacting with real patients.

1.1 Simulation in Nursing Education

The Institute of Medicine has recommended the use of simulation training as an educational technique that may result in improved health care [2]. Studies have also shown that simulation training is an effective strategy to help promote safe clinical practices [3] and that simulation training positively impacts the development of self-efficacy and judgment skills [4]. Five advantages to using simulation in nursing education have been identified: 1) providing opportunity for interactive learning without risk to patients, 2) boosting students' self confidence and reducing anxiety in the practice setting, 3) allowing nursing students to practice clinical decision making and critical thinking in a controlled environment, 4) allowing skills and procedures to be repeated until proficient, and 5) providing immediate feedback [5].

1.2 Virtual Reality in Health Care Education

Virtual patients have been used to teach medical students communication skills and students have rated the virtual patient experience as being as effective as a standardized patient (a paid actor) [6]. Medical students have used virtual patients to help practice patient interviewing skills and results have indicated that using life-size virtual characters and speech recognition is useful in their education [7]. Adult virtual patients are fairly common, but virtual pediatric patients are rare. The use of virtual pediatric patients was first addressed in [8], where they were used for training and assessment for medical students.

1.3 Communicating with Children/Parents

The dynamic of the nurse-family-child relationship is complex due to the many factors that affect this relationship. During an assessment, the nurse has to obtain information from the parent(s) and child, and observe interactions between them [9]. Studies by pediatric experts have shown that the nurse-family-child relationship is heavily dependent upon effective communication, which is a skill that is developed through interaction with different kinds of pediatric patients and families [10]. Nursing students must be aware of the interactions that may affect their communication skills, therefore affecting the nurse-family-child relationship. A positive nurse-family-child relationship promotes the health of the child [10].

2. SYSTEM DESCRIPTION

The Virtual Pediatric Patient System is a simulation of a mother (Mrs. Jones) and her daughter (Sarah) who have come to a medical clinic because Sarah has an earache. The user sees an animated image of the mother and child on a large screen (52") display along with their environment (Fig. 1). The user wears a microphone headset and speaks directly to the characters to interact with them. The correct response is then retrieved from a database and executed. Our scenario is based on a written script provided by faculty members in the School of Nursing.

The software implementation can be described as three linked software modules: speech recognition, question matching, and scenario rendering. The overall program is written in C++ with MFC (Microsoft Foundation Class) for supporting interaction between modules. The first module, speech recognition, is implemented using Dragon Naturally Speaking. This module

Conference'10, Month 1–2, 2010, City, State, Country.

enables a user to create a voice profile for recognition accuracy. We tested two vocabulary options: limited and general. We created the limited vocabulary from words specific to the application. The general vocabulary is provided by Dragon Naturally Speaking containing all the words in the dictionary.

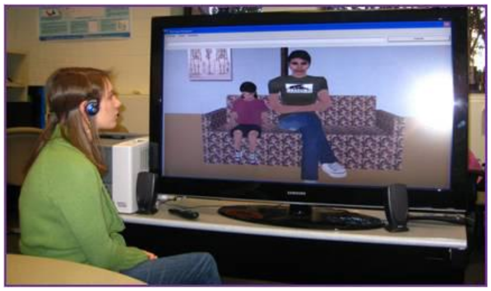


Figure 1. An example of a participant using the system.

The second module matches speech input to a question in a database. Our corpus of questions and responses are stored in a SQLite database. We use a variant of the Answers First Algorithm to match the spoken phrase captured by the speech recognition engine to a question stored in the database [12].

A limitation of the Answers First Algorithm is that a match may not be retrieved if sentences are semantically similar but not syntactically similar. Another problem is that a false match may be retrieved if sentences are syntactically similar but not semantically similar. We can improve matching accuracy by determining and storing many alternative phrasings of a question in the database [13], but doing this manually is very time consuming. To reduce the amount of time and effort spent generating alternative phrasings of a question, we implemented a sentence generation algorithm to automate the process using the Natural Language Toolkit (NLTK) [14] and Wordnet [15]. First, we took each question from our script and broke it into an array of individual words. Each word was then tagged with its part of speech using the Brill Tagger implemented in NLTK. After that, we removed all stopwords from the sentence, including stopwords defined in the NLTK stopwords list as well as proper names included in our script. Next, we applied the Porter Stemming Algorithm to remove common morphological and inflexion endings of words so that synonyms could be found for them in Wordnet-for example, removing the “ing" from “hurting" and mapping the word to its intermediate form “hurt" [16]. After the words were stemmed, we used the parts-of-speech tags and the adapted Michael Lesk Algorithm [17] to find the correct sense, or definition, of the word in Wordnet. We then used Wordnet to find the lemmas and hypernyms from the synset (the set of all possible related words) provided for that word. Finally, variations of the original sentence were generated by substituting all possible lemmas and hypernyms for each word in the sentence into the same location in the sentence as the original word, and generating every possible combination of those synonyms, yielding a set of semantically similar questions (*question set*) that is inserted in the database and mapped to the appropriate response.

To match user questions to the correct question set, our question matching algorithm splits the recognized speech into bigrams, which are pairs of words that appear next to each other. In the original Answers First algorithm [13] the question set with the highest number of matching bigrams was chosen as the correct answer. However, this approach may return an incorrect match when a large question set has a small number of matching bigrams per question, which would yield a high score although each question matched poorly. To avoid this problem, we determine the matching question set by choosing the question set that has the greatest average number of matching bigrams per question in the question set that contained at least one matching bigram, as shown in equation (1). The corresponding response, stored as a series of actions and sentences to be performed by the characters, is then retrieved from the database.

$$\begin{pmatrix} Our\ Score\ for \\ a\ Question\ Set \end{pmatrix} = \frac{\sum_{\substack{Question\,\in \\ Question\ Set}} \begin{matrix} Matching \\ Bigrams \end{matrix}}{\sum_{\substack{Question\,\in \\ Question\ Set}} \begin{Bmatrix} 1\ if\ any\ bigram\ matches \\ 0\ otherwise \end{Bmatrix}} \qquad (1)$$

The final module, scenario rendering, is implemented using DI-Guy [18] for virtual human animation and virtual environment rendering, and Microsoft SAPI SDK for text-to-speech. Once the responses are retrieved from the database, they are sent to this module so that they may be appropriately spoken and displayed.

3. STUDY DESCRIPTION

After completing our initial prototype, we conducted a usability study where we asked five experienced nursing faculty members to use and evaluate our system. Our goals were to identify problems with the system design, check the stability of the software, evaluate the visual and behavioral fidelity of the simulation, and obtain suggestions of improvement.

3.1 Procedures

Each participant began by filling out a consent form and pre-questionnaire. Next, participants completed speech recognition training which consisted of using the short or medium length training module provided by Dragon Naturally Speaking. The participant then sat in a chair in front of the large screen (52") television, which displayed the virtual patients. The participant was then asked to conduct a patient interview with the virtual patients as they normally would. The participants interacted with the virtual patient until they felt they were done with the interview. Each participant then filled out a post-questionnaire and completed a debriefing interview.

3.2 Measures

The pre-questionnaire collected data on demographics, occupational status, and computer experience. The post-questionnaire consisted of the System Usability Scale (SUS) [19], questions about quality of speech interaction, and modified Slater co-presence and presence questionnaires [20]. Through a debriefing interview, we obtained specific information about the overall performance of the system. We audio recorded everything the participant said for transcription, and our system logged data related to speech recognition, animation, the underlying conversational model, and every query into the database.

4. RESULTS AND DISCUSSION

Five experienced nursing faculty members evaluated the system. The nurses were Caucasian females between the ages of 36 and 54. Nurses reported a high level of health care experience as well as a high level of daily computer use (both means above 6, where 7 was frequent experience), but low levels of virtual human experience (mean=2.6 out of 7, sd=1.14, where 7 was frequent use) and daily virtual reality use (mean=1.4 out of 7, sd=0.89, where 7 was frequent use). Participants 1, 3, 4, and 5 interacted with the system once, while participant 2 interacted with the system twice. We gathered questionnaire data for participant 2 once, but we gathered interaction data for participant 2 twice-these instances are referred to as 2a and 2b. Participant 2 interacted with the system using two different settings.

Participants 1, 2a, and 2b used short speech recognition training, while participants 3, 4, and 5 used medium length speech recognition training. Participants 1, 2a, 3, and 4 used the system setting for limited vocabulary while participants 2b and 5 used the system setting for full vocabulary, in order to measure whether our limited vocabulary was effective in increasing recognition accuracy.

4.1 Verbal Interaction

To evaluate our system's overall performance with respect to the verbal aspect of the interview, we transcribed recordings of the interview so that each line of recognized speech and the system's response was paired with the corresponding line of transcribed speech. We categorized each spoken question-matched answer pair into one of six categories (Table. 1). Spoken questions that were semantically similar to questions in our database were categorized as either: correct match, which indicates that the virtual patients answered the question correctly; incorrect match but reasonable response, which indicates that the patient gave an incorrect response that made sense in the context of the nurse's question (for example, nodding in response to a yes/no question instead of giving the intended verbal response); or incorrect response, which indicates that the patient gave a response that did not make sense. Questions that were not semantically similar to any question in the database were classified as: reasonable response; “don't know" response, where the patient responded with a phrase indicating that they did not understand the question; or unreasonable response.

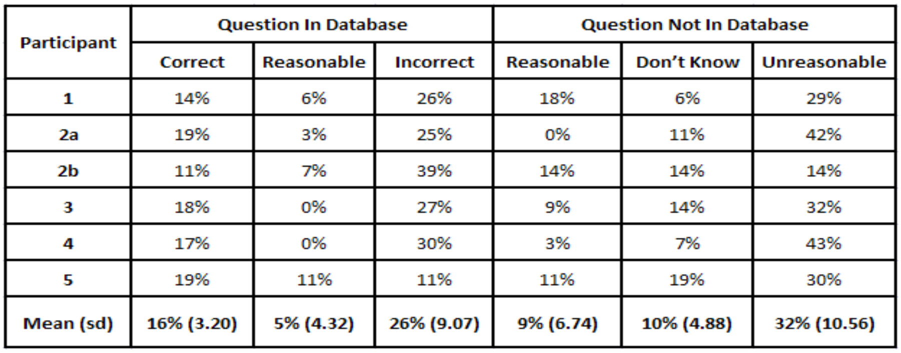
During interaction with the system, each nurse asked between 22 and 36 questions (mean=29.5, sd=5.05). 40% of questions were answered reasonably (correct match, reasonable responses, or “don't know" responses). There are several reasons that we believe the reasonable response rate was low. One important observation is that 51% of the questions that the nurses asked were not represented in the database, so the system had no way to respond to over half of the questions asked. Out of the questions asked that we had answers for in the database, only 16% of them were answered correctly. This could have resulted from lack of accuracy in speech recognition and insufficiency of our conversation model.

4.1.1 Speech Recognition and Conversational Model

By comparing transcribed speech to recognized speech, we found that overall speech recognition correctly recognized 66% of the words nurses used. Nurses with the setting for limited vocabulary had recognition accuracy of 60%, while the recognition accuracy for nurses with the setting for a full vocabulary was 86%. Training length had a negligible effect on recognition accuracy (66.17% for short training and 67.39% for medium training).

We originally chose a limited vocabulary in hopes of improving accuracy, assuming that most words that a nurse used would be included in the automatically generated sentences. Our analysis shows this is true-82% of the words that nurses used were in our limited vocabulary. However, looking at the spoken lines in comparison to the recognized lines, the words that nurses used that were not present in the vocabulary were mapped into incorrect words. These mismatches caused the recognition rate to be much lower for participants in the limited vocabulary than in the full vocabulary. A converse problem was observed using the full vocabulary: words were often mapped to homonyms that made no sense in context of our application. We feel that the benefits of a higher recognition rate through a full vocabulary outweigh the disadvantages, and in future revisions we plan on using a full vocabulary. Out of the 82% of words that nurses used that were in our database, 72% of the words were in our original script, while the additional 10% were added through our sentence generation algorithm. In order to make a manageable and accurate word set we used a subset of the synsets, but in future iterations we may gain better vocabulary coverage by using a larger subset. In addition to better question matching in the database, this larger vocabulary could make limited vocabulary matching in speech recognition more accurate.

Since our question matching algorithm is bigram based, we also evaluated speech recognition and sentence generation in terms of bigram matching. Speech recognition correctly recognized both consecutive words in a spoken bigram 53% of the time. Training length only had a small effect on bigram recognition accuracy (51.58% for short training and 54.92% for medium training). 31% of the bigrams that speech recognition recorded (whether correctly recognized or not) were present in our database. These percentages are a sharp drop from the word-based matching statistics. We chose a bigram-based matching model to help provide context for our word matches. However, this observation suggests that in the next iteration of this software we should consider matching our questions using word-by-word matches instead of bigram matches. Additionally, we noticed that many questions asked with a correct keyword were not answered correctly due to bigram matching. For example, a nurse asked “Is Sarah allergic to any medicine?", but because the bigrams “Sarah allergic" and “allergic to" were not in the database, the question did not find a match, although “allergic" would have been a clear keyword choice for the matching sentence in our database. This is also a limitation caused by our sentence generation algorithm. Since our current algorithm only replaces synonyms of words in the same position in the sentence of the original word, all of our question variants are syntactically identical which makes the



Participant	Question In Database			Question Not In Database		
	Correct	Reasonable	Incorrect	Reasonable	Don't Know	Unreasonable
1	14%	6%	26%	18%	6%	29%
2a	19%	3%	25%	0%	11%	42%
2b	11%	7%	39%	14%	14%	14%
3	18%	0%	27%	9%	14%	32%
4	17%	0%	30%	3%	7%	43%
5	19%	11%	11%	11%	19%	30%
Mean (sd)	16% (3.20)	5% (4.32)	26% (9.07)	9% (6.74)	10% (4.88)	32% (10.56)

Table 1. Categorized questions by participant.

bigrams syntactically similar as well.

4.1.2 *Nurse Feedback*

All nurses expressed frustration with the verbal interaction and commented that they felt that they could not interact naturally with the patient because they had to rephrase questions unnaturally. Most nurses also commented that they felt that the system was unresponsive. In most cases the database response time was less than one second, but because many of the character's responses were nonverbal signs leading up to a verbal answer, it seems that nurses did not realize that the system was responding to their question. Many nurses asked questions in succession without giving the system enough time to process their input and respond. One limitation of our question matching algorithm is that the length of the recognized speech string affects the processing time for finding an answer, so as nurses tried to accommodate for the system's unresponsiveness by stringing several questions together, the processing time increased.

Despite speech recognition performance, all of the nurses still expressed that they would prefer to interact with the system through speech because that is the way a real patient interview works. On a scale from 1 (not at all) to 7 (very much), all nurses ranked the usefulness of talking to the system as 4 or above (mean=5.6, sd=1.14).

5. FUTURE WORK

Despite the problems encountered during interaction with our system, nurses were enthusiastic to see the continued development of this project. After revising our system, we will conduct a second usability study with experienced nurses to gather additional suggestions for modification. When our system becomes sufficiently realistic and usable, we will conduct a large scale user study with a class of nursing students, comparing our system to their current training methods. A long term goal is to extend our system to include other scenarios representing a variety of patient genders, ages, ethnic backgrounds, and physical characteristics. Additionally, our system framework gives us the potential to create interview scenarios that can be used for other fields.

6. FUTURE WORK

This research was supported, in part, by NSF Research Experience for Undergraduates Sit Grant CNS-0850695, the NSF Graduate Research Fellowship Program, and a grant from the College of Health Education and Human Development at Clemson University.

7. REFERENCES

[1] Rothgeb, M.K.: Creating a Nursing Simulation Laboratory: A Literature Review. J. Nurs. Ed. 47(11). 489--494 (2008)

[2] Kohn, L.T., Corrigan, J.M., Donaldson, M.S.: To Err is Human: Building a Safer Health System. From: Committee on Quality of Health Care in America, Institute of Medicine. National Academy Press, Washington (2000)

[3] McKeon, M., Norris, T., Cardwell, B., Britt, T.: Developing Patient-Centered Care Competencies Among Prelicensure Nursing Students Using Simulation. J. Nurs. Ed. 48(12). 711--715 (2009)

[4] Bambini, D., Washburn, J., Perkins, R.: Outcomes of Clinical Simulation for Novice Nursing Students: Communication, Confidence, Clinical Judgement. In: Nursing Education Perspectives. 30(2). 79--82 (2009)

[5] Durham, C.F. and Alden, K.R.: Enhancing Patient Safety in Nursing Education Through Patient Simulation. In: Hughes, R. Patient Safety and Quality: An Evidence-Based Handbook for Nurses, Chapter 51. AHRQ publication No. 08-0043 (2008)

[6] Stevens, A., Hernandez, J., Johnsen, K., Dickerson, R., Raij, A., Harrison, C. et al.: The Use of Virtual Patients to Teach Medical Students History Taking and Communication Skills. In: The American Journal of Surgery. 191. 806--811 (2006)

[7] Johnson, K., Dickerson, R., Raij, A., Lok, B., Jackson, J., Shin, M., Hernandez, J., Stevens, A., Lind, D.S.: Experiences in Using Immersive Virtual Characters to Educate Medical Communication Skills. In: Proceedings of the IEEE Virtual Reality Conference. 324. 179--186 (2005)

[8] Hubal, R.C., Deterding, R.R., Frank, G.A., Schwetske, H.F., Kizakevich, P.N.: Lessons Learned in Modeling Virtual Pediatric Patients. In: Studies in Health Technology and Informatics. 94. 127--130 (2003)

[9] Hockenberry, M.J. and Barrera, P.: Communication and Physical Developmental Assessment of the Child. In: M.J. Hockenberry and D. Wilson (eds.). Wong's Nursing Care of Infants and Children, pp. 141--2004. Mosby Elsevier, Missouri (2007)

[10] Ball, J.W., Bindler, R.C., Cohen, K.J.: Child and Family Communication. In: Ball, Bindler, and Cowen (eds.). Child Health Nursing (4th ed.), pp. 170--187. Pearson, New Jersey (2009)

[11] Wilson, D.M., Martin, A.M., Gilbert, J.E.: `How may I help you?'- Spoken Queries for Technical Assistance. In: Proceedings of the 48th Annual Southeast Reginal Conference, (43). New York (2010)

[12] Wilson, D.M. and Gilbert, J.E.: ITECH: An Interactive Technical Assistant. Dissertation. Auburn University, AL (2006)

[13] Natural Language Tooklit, http://nltk.sourceforge.net

[14] Wordnet: A Lexical Database for English, http://wordnet.princeton.edu/

[15] Dao, T.N. and Simpson, T.: Measuring Similarity Between Sentences, http://opensvn.csie.org/WordNetDotNet

[16] Banerjee, S. and Pedersen, T.: An adapted Lesk Algorithm for Word Sense Disambiguation Using Wordnet. In: Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing, pp 136--145. Springer-Verlag, United Kingson (2002)

[17] DI-Guy Human Simulation Software, http://www.diguy.com/

[18] Brooke, J.: SUS - A Quick and Dirty Usability Scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (eds.). Usability Evaluation in Industry, pp. 189--194. Taylor and Francis, London (1996)

[19] Mortensen, J., Vinayagamoorthy, V., Slater, M. Steed, A., Lok, B., Whitton, M.C.: Collaboration in Tele-Immersive Environments. In: Eight Eurographics Workshop on Virtual environments, pp. 99--101. (2002)

[20] Measuring Usability with the System Usability Scale, www.measuringusability.com/sus.php