Virtual Patients: Assessment of Synthesized Versus Recorded Speech

Robert Dickerson[1], Kyle Johnsen[1], Andrew Raij[1], Benjamin Lok[1],
Amy Stevens[2], Thomas Bernard[3], D. Scott Lind[3]
[1]*Department of Computer Information Science Engineering, University of Florida*
[2]*College of Medicine, University of Florida*
[3]*Medical College of Georgia*

Abstract. Virtual patients have great potential for training patient-doctor communication skills. There are two approaches to producing the virtual human speech: synthesized speech or recorded speech. The tradeoffs in flexibility, fidelity, and cost raise an interesting development decision: which speech approach is most appropriate for virtual patients? Two groups of medical students participated in a user study interviewing a virtual patient under each condition. We found no significant differences in the overall impression, speech intelligibility, and task performance. Our conclusion is that if the goal is to train students of which questions to ask, synthesized speech is just as effective as recorded speech. However, if the goal is to teach the student how to ask the correct questions, a high level of expressiveness in the virtual patient is needed. This in turn necessitates the higher cost – even with the lower flexibility – of recorded speech.

Keywords. Synthesized speech, Virtual Patients, Medical Education

1. Introduction

Virtual patients are receiving serious attention as a powerful tool for educating medical students. Through repeated experiences with virtual patients, medical students can be exposed to, and evaluated on, many more situations then through traditional methods. To better the training potential of virtual patients, we must study in-depth the components necessary to create compelling social experiences. Each of the core components: graphics rendering, speech recognition, speech processing, and speech synthesis play a part in the overall impression of a virtual human. Medical textbooks frequently detail the value of verbal cues as a critical source of patient information[1]. When designing virtual patient systems, developers are presented with two approaches to generate speech: pre-recorded or synthesized speech.

Recorded Speech – speech of the virtual patient is recorded using voice talent. While monetarily costly, time consuming, and inflexible, recorded speech has very high fidelity and can be extremely emotive.

Synthesized Speech - given a text string, software libraries generate audio output. While the fidelity can range from 'robotic' to passable, the low cost (time and money) and dynamic nature of synthesized speech makes it an attractive approach.

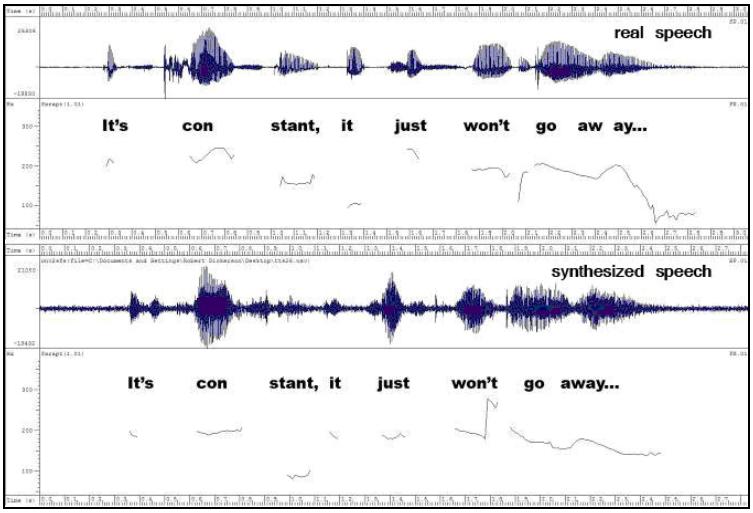

Figure 1. Comparison of intonation contours from a fundamental-frequency analysis for the VP's emotive expression "It's constant, it just won't go away", generated with Speech Filing System.

Spoken dialogue systems (such as telephony applications) typically use recorded speech whenever possible. When user interfaces employ synthetic speech, they typically use messages with simple structure to reduce the cognitive and memory demands on the user. Table 1 shows the tradeoffs from using either synthesize or recorded speech. Using recorded speech is time consuming since a voice talent is required to pre-record all the virtual standardized patient's responses. Synthesized speech makes it easy to update the script quickly and easily. Pre-recorded speech makes it difficult to use dynamic data, (such using names after introductions, or dynamically generating new responses). Systems using AI such as natural language generation would require the use of synthesized voice. However NLP seems unnecessary for closed domain scenarios such as the acute-abdominal pain, because of the predictability of the set of questions [2].

Table 1. Trade-off matrix for recorded speech vs. synthesized speech

Recorded speech	Synthesized Speech
Inflexible (non-dynamic)	Flexible (dynamic)
High fidelity (emotive)	Low fidelity (non-emotive)
Costly (time consuming)	Inexpensive (quick)

The tradeoffs in the three dimensions (flexibility, fidelity, and cost) raise an interesting development decision: which speech approach is most appropriate for virtual patients?

To investigate this question, we employ the high fidelity virtual interactive patient system (VIPS). In VIPS, medical students naturally interact with a virtual patient, DIANA (DIgital ANimated Avatar). DIANA is projected onto the wall of a mock examination room (she's 5'6") and talks and gestures with the student [Figure 2]. Modeled after a standardized patient experience, DIANA is scripted with an abdominal pain condition, and the student's goal in the 10-minute experience is to explore the

history of present illness (with no physical exam). Previous work has presented the system [2-4], the student-virtual patient interaction [5], and the students' responses [4].

Would experiencing DIANA with synthesized speech (Group SS) cause the student to perform differently than recorded speech (Group RS)? We conducted an experiment with medical students to explore how the virtual patient's speech type (RS or SS) impacted the experience.

The study result provides insight into the advantages and disadvantages of using synthesized speech and evaluates the necessary fidelity for communication skills training. Our approach was to empirically compare the two speech modes and weigh each against a human standardized patient. Could a lower fidelity system still be acceptable if it preserves the accuracy of simulation and training?

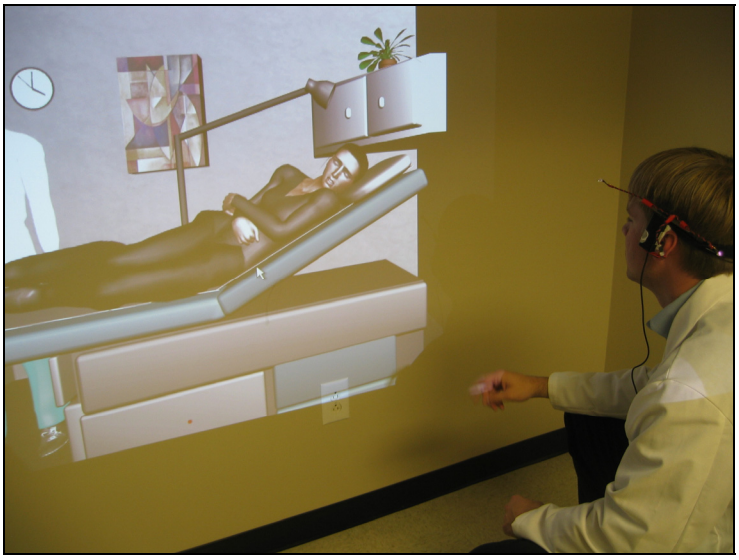

Figure 2. Medical student practices the AAP scenario with a virtual patient

2. TOOLS AND METHODS

2.1. Study Description

A user study was run with seventeen medical students at the Medical College of Georgia in their second or third year of study. All had several prior experiences with standardized patients. Participants were divided randomly into two groups with a system running with recorded speech (N=9) or synthesized speech (N=8).

Each participant filled out a background survey. Then they entered the exam room and spent 10 minutes interviewing with a virtual patient, took a history of present illness, and stated their differential diagnosis to a virtual instructor. After the interview, the participants completed a set of questionnaires then a recorded debriefing interview.

We used the "Crystal" 16K voice from the AT&T Labs Natural Voices SDK for synthesized speech. The recorded speech came from a female adult voice talent.

2.2. Measures

1. *Speech Quality Questionnaire*. The quality judgments were made by using an adapted questionnaire developed for evaluating telephone dialogue systems [6], targeting intelligibility, naturalness, pleasantness, comprehension, and overall acceptance of the voice.

2. *Maastricht Assessment of the Simulated Patient* [7]. The questionnaire is the standard method for assessing a standardized patient.

3. *Expert Evaluation*. Experts evaluated the tapes of the interactions and determined student task performance by identifying which core pieces of information, such as symptoms and signs, the student was able to elicit from DIANA including sections from chief complaint, history of present illness, sexual history, etc. Examples include: "I've been nauseous", "I have a fever", and "I am sexually active".

3. RESULTS

3.1. Learning objectives were met in both cases - no effect on task performance

Table 2. Expert Evaluation

	Synthesized	Real Speech	p
Evaluation Rating (score out of 12)	$\mu = 4.37$ $\sigma = 1.59$	$\mu = 5.00$ $\sigma = 1.85$	0.45

No differences were found in the task performance ratings assigned by experts [Table 2]. The ratings reflect the number of core questions asked during the interview. The SS condition presents lower fidelity audio than with RS, and may impact the effectiveness and believability of the simulation especially under more emotive scenarios. Synthesized speech allows the student to still meet educational objectives, and students scored DIANA was equally under each condition for teaching (RS μ 5.6, SS μ 5.6, $p=0.46$) and training (RS μ 5.1 μ SS 5.1, $p=0.49$) value.

3.2. No differences were identified in how participants' rating of the intelligibility, naturalness, pleasantness, comprehension, and overall acceptance of the voice

Based from questionnaire results, there was no reported difference in the intelligibility (RS μ 4.9, SS μ 4.6, $p<0.28$), naturalness (RS μ 4.3, SS μ 4.2, $p<0.47$), and clarity (RS μ 5.2, SS μ 5.0, $p<0.46$) of the voice.

In the post-experience debriefing, SS participants had varied impressions. One responded that *"[The VP's voice] was clear"* another said, *"I had expectation that she wasn't going to sound exactly like a real person. She sounded like a telephone operator."* Most RS participants were very satisfied with the quality of the voice. One said, *"I felt like they were really realistic as far as their voice intonation."*

3.3. Some SS participants noted the synthetic speech sounded unnatural at first, yet they adapted to it.

During interactions, there is often readjustment period that occurs when adapting to the speech recognition and being exposed to synthesized speech. Quickly the participants stopped paying attention to the lack of prosody, and accepted the flow of conversation that the interface presented them. The following are debriefing comments received by participants in the synthesized speech condition:

"For some people I can see how [the voice] would get confusing because [I heard the speech] just one word at a time, but it was ok for me because I had heard it many times before. Synthetic speech is clearly not a human being."

"If she's in that much pain she should be making more sounds."

"[VP's] voice could have been improved. I had difficulty hearing [the VP] at first".

In the questionnaire, the participants responded whether "this encounter is similar to other standardized patient encounters that I've experienced", there is some indication that recorded speech is more familiar to students than synthesized (RS: μ 2.8, SS: μ 2.0 $p<0.06$).

3.4. Lack of prosody was not detrimental for the basic skills teaching

The role of prosody (non-verbal cues) is used to identify grammatical structure, convey attitude and emotion, and convey personal or social identity [8]. However, these cues seemed to minimally impact this relatively simple scenario.

Stress and intonation can help identify grammatical structure. Stress is used to highlight or give emphasis to word, and can help with clarification. Intonation is used differentiate a question (yes/no, either or) from a simple declaration. The SS participants did not find SS limiting due to the simplicity of the VP's responses, the assumption that every response was a statement, and the simplicity of the conversation flow. Ambiguity did occur once in the scenario when the VP spontaneously asks the participant "can you help me!?" some SS participants were thrown off and had difficulty registering it as a question.

Speech can show attitude and emotion, personality and social identity, however much of this information is visually presented. There may be a synergy of graphics and audio, and DIANA's expressive animation might have filled in what the audio had missing. Prosody appears more important for speech-only systems.

4. Conclusion

The results indicate no significant difference in performance between Group SS and Group RS in many of the task performance measures, such as the asking the correct questions.

One important external validity measure is to identify the similarities and differences between experiencing a virtual patient and standardized patient – as clearly they are not equal. Upon closer inspection, there exist subtle – yet important – differences between virtual patients and standardized patients, primarily relating to *conversation flow* and the significant difference in level of *expressiveness*. Part of the lowered expressiveness is auditory, and thus SS's lower level of emotive expression impacts the overall experience. Recorded speech appears to be required to explore higher order communication skills. Our conclusions are as follows:

For lower level learning of communication skills, (knowledge on Bloom's Taxonomy of Learning), there appears to be little difference between RS and SS. Thus if the goal is to teach the student *which questions to ask*, SS provides a compelling dynamic approach with minimal loss of educational objectives.

However, if the goal is to teach the student *how to ask the correct questions*, (analysis and application) a high level of expressiveness in the virtual patient is needed. Essential information of the patient's condition could be lost from using synthesized speech. This in turn necessitates the higher cost – even with the lower flexibility – of recorded speech.

5. Future Work

In order to build more effective virtual patient applications we intend to explore other system design decisions, such as graphics, immersion, and speech understanding and how they affect overall system impressions.

6. References

1. Coulehan, J. and M. Block, *The Medical Interview: Mastering the Skills for Clinical Practice*. 1997.
2. Dickerson, R., K. Johnsen, A. Raij, B. Lok, J. Hernandez, A. Stevens. *Evaluating a Script-Based Approach for Simulating Patient-Doctor Interaction*. in *International Conference on Human-Computer Interface Advances for Modeling and Simulation (SIMCHI)*. 2005. New Orleans, LA.
3. Johnsen, K., et al., *Using Immersive Virtual Characters to Educate Medical Communication Skills.* Presence: Teleoperators and Virtual Environments, 2005.
4. Stevens, A., et al., *The Use of Virtual Patients to Teach Medical Students Communication Skills.* American Journal of Surgery, 2005.
5. Raij, A., et al. *Interpersonal Scenarios: Virtual ≈ Real?* in *VR 2006*. 2006.
6. Möller, S., *Quality of Telephone-Based Spoken Dialogue Systems*. 2005: Springer.
7. Wind, L., et al. *Assessing Simulated Patients in an Educational Setting: The Maastricht Assessment of Simulated Patients*. in *Medical Education*. 2004.
8. Cohen, M., J. Giangola, and J. Balogh, *Voice User Interface Design*. 2004: Addison-Wesley.