Coherence Changes Gaze Behavior in Virtual Human Interactions

Richard Skarbez*   Gregory F. Welch†   Frederick P. Brooks, Jr.‡   Mary C. Whitton§

University of Central Florida, USA

University of North Carolina at Chapel Hill, USA

ABSTRACT

We discuss the design and results of an experiment investigating Plausibility Illusion in virtual human (VH) interactions, in particular, the coherence of conversation with a VH. This experiment was performed in combination with another experiment evaluating two display technologies. As that aspect of the study is not relevant to this poster, it will be mentioned only in the Materials section. Participants who interacted with a low-coherence VH looked around the room markedly more than participants interacting with a high-coherence VH, demonstrating that the level of coherence of VHs can have a detectable effect on user behavior and that head and gaze behavior can be used to evaluate the quality of a VH interaction.

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented and virtual realities; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Evaluation/methodology; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual Reality

1 EXPERIMENT

The experiment was a between-subjects design. Each participant performed one interview with a virtual human patient who had come to a medical facility complaining of stomach pain. The interviews lasted approximately ten minutes. For all participants, virtual human responses were generated by an experimenter in a Wizard-of-Oz (WoZ) setup. Totally freeform responses were not possible; the experimenter selected them from a searchable list of responses that had previously been recorded by a voice actor. The participant was not aware that responses were being chosen by a real person; it appeared that they were generated by voice recognition.

There were two experimental conditions. Specifically, the WoZ followed two different behavior patterns. In the high-coherence condition, the experimenter responded to the participant as quickly and as accurately as was possible. In the low-coherence condition, the experimenter responded according to a script with a variety of conversational errors. The different types of errors were derived from [5]. Since the exact conversation could not be predicted in advance, the error script was of the form, "On the fourth exchange, ignore the participant. On the ninth exchange, repeat the answer twice in a row," and so on. This ensured that all participants experienced a variety of errors at a predictable frequency.

We had initially intended for the low-coherence condition to have responses selected by the voice recognition software, and for the high-coherence condition to have responses selected by the experimenter in a WoZ setup. Piloting, however, revealed problems with this experimental setup, owing to the nature of the voice recognition software. Using it, some participants' voices were almost perfectly recognized by the system, leading to very few errors, while some other participants were almost unable to get the virtual patient to respond at all. In short, there was no way to have a standardized amount of unreasonable circumstances occur in the low-coherence condition using the voice recognition software.

Participants

Thirty-two medical school students (18 female, 14 male), with an average age of 25.8 $\pm$ 2.3 years, were recruited from the university medical school. They were compensated for their participation.

Materials

The virtual human models, scripts, and voice recordings were provided by the Virtual Experiences Research Group (VERG) at the University of Florida. Participants' eye and head positions were tracked using an Optitrack optical tracking system. Depending on the display condition, the virtual patient was rendered on a large 3DTV or was embodied in a physical-virtual avatar (PVA), as shown in Figure 1. The PVA was initially developed for use in another experiment [4]. Both displays were present in the room for all participants; whichever was not in use was covered with a black cloth. The displays were placed so that the virtual patient in both display conditions would subtend approximately the same visual angle from the participant's seated position, which was the same for all participants. This arrangement can be seen in Figure 2.

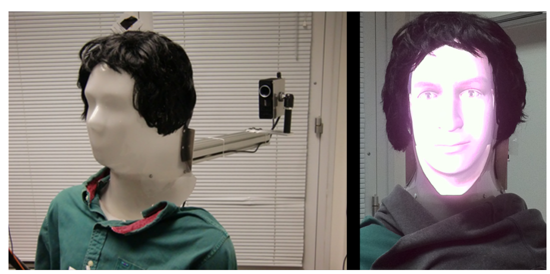

Figure 1: At left, the physical-virtual avatar (PVA) apparatus. At right, the participant's view of the PVA when in use. The face is animated, computer-generated imagery projected onto the inside of the plastic face. The projector can be seen in the center of Figure 1, Left.

Metrics

Participants filled out both pre- and post-experiment questionnaires, including demographic and medical training information, the Maastricht assessment of the simulated patient (MaSP), a series of questions asking the participant to compare the VH interaction to other types of interactions they may have had in everyday life, and a series of questions regarding the plausibility of the VH.

The MaSP is an instrument initially designed to assess the quality and authenticity of real simulated patients, i.e., human actors who portray a patient with a given condition [6]. The MaSP has been modified and used to evaluate the quality of virtual human simulated patients as well, as in [2] and [3].

We also recorded the positions of the eyes and torso for each participant during the interaction, so that we could measure large-

*e-mail: skarbez@gmail.com
†email:welch@ucf.edu
‡e-mail:brooks@cs.unc.edu
§e-mail:whitton@cs.unc.edu

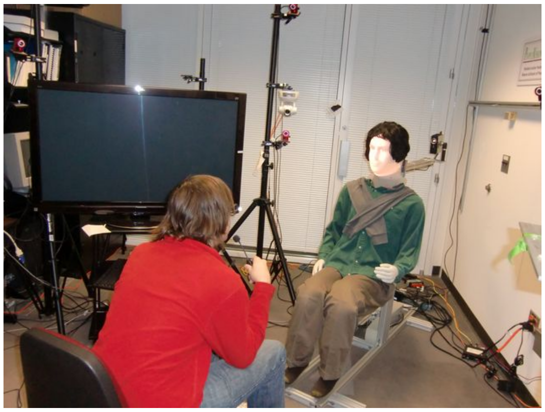

Figure 2: User performing an interview with the PVA. At left of image is the flat-panel stereo display used in the other display condition.

scale gaze behavior and postural response to the virtual patient. The eye position was tracked using glasses with retroreflective markers attached, which patients wore regardless of display condition. Note that we did not employ eye tracking; the tracked point was roughly the center of the two eyes, a point on the bridge of the nose.

2 Results

We performed Bayesian data analysis on the study data. In the Bayesian method of analysis, all variables are considered as part of a single overall model, where all the stochastic equations are evaluated simultaneously, rather than one at a time. Non-informational priors (normal functions with high variance) have been chosen so as not to bias the results.

Unlike traditional null-hypothesis testing, there is no single value such as a p-value that determines whether the result is "significant". Instead, we report the posterior probabilities and readers are free to interpret those probabilities for themselves. Posterior probabilities near 50% indicate that both outcomes are approximately equally likely, and so provide negligible evidence for the stated hypothesis. Greater probabilities provide stronger evidence for the stated hypothesis. (Probabilities can also be less than 50%, providing evidence in the corresponding way for the inverse hypothesis.) This manner of describing the results follows Bergström et al. [1].

For the majority of measures, there is negligible evidence to support an effect of the of the VH response coherence, of the display technology used, or of any covariants (age, semesters in medical school, number of patient interviews, number of standardized patient interviews, or self-rankings of anxiety, comfort, preparedness, or skill). Notably, the MaSP scores do not reveal a difference between the low-coherence and high-coherence conditions. However, there are a few results for which there is some evidence.

The posterior probability that participants in the low-coherence condition moved their heads more than participants in the high coherence condition, as measured by the standard deviation of head position, is 78.6%.

The posterior probability that participants in the low-coherence condition reported lower scores to the question, "How strongly did you sense that the patient was watching you?" than participants in the high-coherence condition is 77.2%.

3 Discussion

The results described above indicate that participants detected the experimental manipulation of coherence, at least at some level. Anecdotally, the experimenters noticed that participants in the low-coherence condition were more fidgety, and looked around the room (and not at the virtual patient) more than participants in the high-coherence condition. This observation is supported by the eye-position tracking data, in which the standard deviation of eye position is very likely to be smaller for high-coherence participants than low-coherence participants. Furthermore, the question, "How strongly did you sense that the patient was watching you?" is the only question that directly asked about the behavioral response of the virtual patient. In other words, it was the closest question available to, "Was the virtual patient paying attention?" We speculate that participants used this question as a means to say what they really noticed, which is that the virtual patient seemed less responsive to their statements in the low-coherence condition. Neither of these observations is definitive on its own, but in combination, we believe that this indicates that the lower coherence affected participants, it just did not have an effect on task completion or on the other post-test measures.

This result may have occurred because this group was highly motivated. Of the 32 participants, 30 stated in post-experiment interviews that, regardless of display or coherence condition, they would use the technology if it were available. The overwhelming feeling of participants was that they thought interviewing with virtual simulated patients was useful and they were excited about the potential of the technology. This feeling likely overcame any difficulties or concerns about the implementation. This is good news for the prospect of virtual human simulated patients in general, however, it may make the patient interview an unsuitable use case for differential evaluation of technology.

More generally, this result provides additional evidence that user motivation can supplement the technology, such that a more motivated, invested, or attentive user may feel presence or demonstrate realistic response in a situation where a less motivated user might not [7]. This is a boon for developers, because a user who is convinced that new technology can help them do real work is likely to devote more attentional resources, and this in turn will generate more presence, more realistic response, and more motivation in a virtuous circle. It is a challenge for researchers, though, as they face the problem of high user motivation obscuring experimental effects that might be more apparent with naïve users.

References

[1] I. Bergström, K. Kilteni, and M. Slater. First-person perspective virtual body posture influences stress: A virtual reality body ownership study. *PLoS ONE*, 11(2), 2016.

[2] K. Johnsen, A. Raij, A. O. Stevens, D. S. Lind, and B. Lok. The validity of a virtual human experience for interpersonal skills education. In *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1049–1058. ACM, 2007.

[3] A. B. Raij, K. Johnsen, R. F. Dickerson, B. Lok, M. S. Cohen, M. Duerson, R. R. Pauly, A. O. Stevens, P. Wagner, and D. S. Lind. Comparing interpersonal interactions with a virtual human to those with a real human. *IEEE Transactions on Visualization and Computer Graphics*, 13:443–457, 2007.

[4] D. Rivera-Gutierrez, G. Welch, P. Lincoln, M. Whitton, J. Cendan, D. A. Chesnutt, H. Fuchs, and B. Lok. Shader lamps virtual patients: The physical manifestation of virtual patients. *Studies in Health Technology and Informatics*, 173:372–378, 2012.

[5] R. Skarbez, A. Kotranza, J. Brooks, F.P., B. Lok, and M. Whitton. An initial exploration of conversational errors as a novel method for evaluating virtual human experiences. In *IEEE Virtual Reality (VR)*, pages 243–244, March 2011.

[6] L. A. Wind, J. Van Dalen, A. M. M. Muijtjens, and J.-J. Rethans. Assessing simulated patients in an educational setting: the MaSP (Maastricht Assessment of Simulated Patients). *Medical Education*, 38(1):39–44, 2004.

[7] W. Wirth, T. Hartmann, S. Böcking, P. Vorderer, C. Klimmt, H. Schramm, T. Saari, J. Laarni, N. Ravaja, F. R. Gouveia, F. Biocca, A. Sacau, L. Jäncke, T. Baumgartner, and P. Jäncke. A process model of the formation of spatial presence experiences. *Media Psychology*, 9(3):493–525, 2007.