A Demonstration of the Perception System in SimSensei, a Virtual Human Application for Healthcare Interviews

Giota Stratou, Louis-Philippe Morency, David DeVault, Arno Hartholt, Edward Fast,
Margaux Lhommet, Gale Lucas, Fabrizio Morbini, Kallirroi Georgila, Stefan Scherer,
Jonathan Gratch, Stacy Marsella, David Traum and Albert Rizzo
University of Southern California, Institute for Creative Technologies
Los Angeles, California

***Abstract*—We present the SimSensei system, a fully automatic virtual agent that conducts interviews to assess indicators of psychological distress. With this demo, we focus our attention on the perception part of the system, a multimodal framework which captures and analyzes user state behavior for both behavioral understanding and interactional purposes. We will demonstrate real-time user state sensing as a part of the SimSensei architecture and discuss how this technology enabled automatic analysis of behaviors related to psychological distress.**

Keywords*—*Perceptive Virtual Agent; Healthcare Virtual Agent; SimSensei

I. Introduction

Virtual humans that can develop intimacy with people are now becoming reality. Researchers have successfully incorporated social skills (e.g., active listening, mimicry, gestures) into virtual human systems [1], [2], [3]. Indeed, compared to their predecessors, virtual humans with such social skills increase feelings of connection and rapport, namely the experience of harmony, fluidity, synchrony, and flow felt during a conversation [2]. Enhanced with these skills, virtual humans could be particularly useful as agents in a variety of applications. For example, a user experience can be better standardized with virtual humans than with human beings. Recently it was also shown that virtual humans can increase willingness to disclose by providing a "safe" environment where participants do not feel judged by another human interlocutor [4]. This is particularly helpful in the case of healthcare scenarios where disclosure of sensitive information is prominent. An automatic system behind the virtual human brings other advantages as well. Utilizing advances in natural language understanding, facial expression analysis and gesture recognition scientists have the tools to analyze and quantify verbal and nonverbal behavior exchanges during real-time interactions. This capability brings forward new opportunities in applications where measuring and assessing human behavior is important (for example in healthcare scenarios [5]).

With this demonstration, we present the SimSensei system, a fully automatic system designed to conduct semi-structured interviews related to psychological distress [6]. In the live demo we emphasize the perception part of the system, a multimodal framework which captures and analyzes user state behavior for both behavioral understanding and interactional purposes.

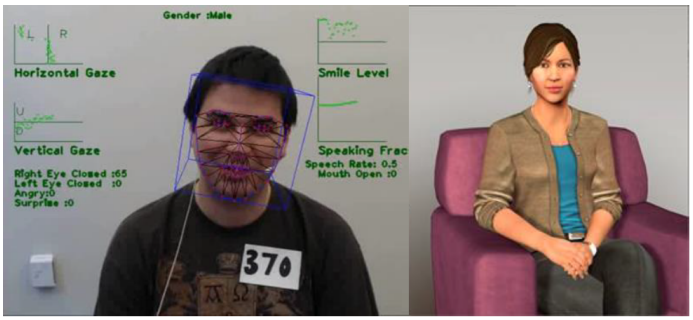

Fig. 1. SimSensei interface with visualization of live tracking. This is a side by side view of a participant interacting with Ellie, the SimSensei agent (seen on the right side). On the left side, we see an overlay of the live tracking by MultiSense, the multimodal perception system used in SimSensei, on the participant camera view. The participant's tracked state informs the automatic system and enhances the interaction.

II. SimSensei System Overview

SimSensei, seen in Fig.1 is a fully automatic system designed to create an engaging face-to-face interaction with a user. It was developed as a clinical decision support tool that would complement existing self-assessment questionnaires by giving healthcare providers quantifiable measures of the user behaviors that are correlated with psychological distress. As such, the main design goals were to create an interactional environment that allows the user to feel comfortable talking during a semi-structured interview, so that elicited behaviors can be automatically measured and analyzed under the specific context.

The SimSensei system has been designed and developed over a two year iterative process described in detail in DeVault et al. [6]. It is implemented using a modular approach (similar architecture seen in Hartholt et.al [7]) and its main components include a dialogue processing component (supported by individual modules for speech recognition, language understanding and dialogue management), a nonverbal behavior

generation component for the virtual human, a rendering engine and a multimodal perception system that analyzes audiovisual streams in real time.

III. Multimodal Perception System Overview

The perception system was tailored specifically for this application (following details in DeVault et al. [6]). Specifically, it should serve a double purpose: i) communicate the necessary nonverbal behavior signals to the other components of the system so that the agent is sensitive to the user's nonverbal behavior, and ii) recognize automatically and quantify the nonverbal behaviors. As an example, tracking the smile intensity of the participant serves both of these purposes: smile is a signal that has been tied to investigations of depression [8] and also plays an important role in a dyadic interaction [9]. As a basis for the perception system, SimSensei uses the MultiSense framework, which is a flexible system for multimodal real-time sensing. MultiSense allows for systematic capture of different modalities, and provides a flexible platform for real-time tracking and multimodal fusion. MultiSense dynamically leverages the measures from current technologies into informative signals such as smile intensity, 3D head position, intensity or lack of facial expressions like anger, disgust and joy, speaking fraction, gaze direction etc. As mentioned above, these informative signals serve two purposes. First, they contribute to the indicator analysis; constructs like affect, gesture and emotional variability and elements of engagement can be measured by the automatic system and inform the distress assessment. Second, they are broadcasted to the other components of SimSensei system using the PML standard [10] to inform the virtual human of the state of the participant and assist with turn taking, listening feedback, and building rapport by providing appropriate non-verbal feedback.

IV. System Evaluation

First, with respect to the user experience, it was possible to assess the rapport the users felt with the system through self-report questionnaires in each phase. Overall the results were promising; participants reported willingness to disclose, willingness to recommend and general satisfaction with both the wizard-driven and the fully AI version of the system [4]. Also, there is no significant difference in felt rapport as reported by the users between face-to-face interactions with another human and the final AI system [6] (different stages of the system and the data collected with the SimSensei framework are described in [11]). Interestingly, participants felt more rapport when interacting with the wizard-driven system than they did in face-to-face interviews. This is further explored with the fully AI system showcasing that even with the same virtual human interface, the belief that participants interact with a fully automatic system versus a human operator behind the scenes increases their willingness to disclose [4]. These results suggest that automated virtual humans that can engage participants to disclose information can help overcome a significant barrier to obtaining truthful patient information.

Moreover, with respect to providing an environment where informative behaviors can be elicited and analyzed in relation with psychological distress, relevant work on different modalities such as the verbal, audio and video (face and body) [12], [13], [14], [15] showed that a virtual human interaction is still rich in behaviors and information that allow for analysis and assessment of distress via fully automatic methods. Specifically, user state behaviors captured via the MultiSense perception system have enabled investigation of nonverbal behaviors of subjects with psychological disorders (such as anxiety, depression or post-traumatic stress disorder) replicating findings from clinical literature via automatic analysis, for example, the display of less facial expressivity was shown as a gender independent indicator of depression [15] and voice quality differences between distressed and non-distressed participants were quantified [14] in SimSensei interactions. In the work mentioned above, indicator behaviors captured by the system were used to automatically assess if a participant is suffering from a psychological condition with promising performance.

Acknowledgment

This work is supported by DARPA under contract W911NF-04-D-0005 and the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] T. Bickmore, A. Gruber, and R. Picard, "Establishing the computerpatient working alliance in automated health behavior change interventions," *Patient Education and Counseling*, vol. 59, no. 1, pp. 21 – 30, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0738399104003076

[2] J. Gratch, S.-H. Kang, and N. Wang, "Using social agents to explore theories of rapport and emotional resonance," *Social Emotions in Nature and Artifact*, p. 181, 2013.

[3] K. Anderson, E. Andre, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, A. Egges, P. Gebhard, H. Jones, M. Ochs, C. Pelachaud, K. Porayska-Pomsta, P. Rizzo, and N. Sabouret, *The TARDIS framework: intelligent virtual agents for social coaching in job interviews*, ser. Lecture Notes in Computer Science. Springer, 2013.

[4] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency, "It's only a computer: Virtual humans increase willingness to disclose," *Computers in Human Behavior*, vol. 37, no. 0, pp. 94 – 100, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563214002647

[5] J. Gratch, L.-P. Morency, S. Scherer, G. Stratou, J. Boberg, S. Koenig, T. Adamson, A. A. Rizzo *et al.*, "User-state sensing for virtual health agents and telehealth applications." in *MMVR*, 2013, pp. 151–157.

[6] D. DeVault and et al., "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, ser. AAMAS '14. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1061–1068. [Online]. Available: http://dl.acm.org/citation.cfm?id=2617388.2617415

[7] A. Hartholt, D. R. Traum, S. C. Marsella, A. Shapiro, G. Stratou, A. Leuski, L.-P. Morency, and J. Gratch, "All together now - introducing the virtual human toolkit," in *IVA*, 2013, pp. 368–381.

[8] L. I. Reed, M. A. Sayette, and J. F. Cohn, "Impact of depression on response to comedy: a dynamic facial coding analysis." *Journal of abnormal psychology*, vol. 116, no. 4, p. 804, 2007.

[9] J. K. Burgoon, L. K. Guerrero, and K. Floyd, *Nonverbal communication*. Allyn & Bacon Boston, MA, 2010.

[10] S. Scherer, S. C. Marsella, G. Stratou, Y. Xu, F. Morbini, A. Egan, A. Rizzo, and L.-P. Morency, "Perception markup language: Towards a standardized representation of perceived nonverbal behaviors," in *The 12th International Conference on Intelligent Virtual Agents (IVA)*, Santa Cruz, CA, Sep. 2012.

[11] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, *The Distress Analysis Interview Corpus of human and computer interviews*. LREC, 2014.

[12] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, A. S. Rizzo, and L.-P. Morency, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, 2014.

[13] D. DeVault, K. Georgila, R. Artstein, F. Morbini, D. Traum, S. Scherer, A. S. Rizzo, and L.-P. Morency, "Verbal indicators of psychological distress in interactive dialogue with a virtual human," in *Proceedings of the SIGDIAL 2013 Conference*. Metz, France: Association for Computational Linguistics, August 2013, pp. 193–202. [Online]. Available: http://www.aclweb.org/anthology/W13-4032

[14] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd." in *Interspeech*, 2013, pp. 847–851.

[15] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and ptsd: the effect of gender," *Journal on Multimodal User Interfaces*, pp. 1–13, 2014.