An Adaptive Virtual Agent Platform for Automated Social Skills Training

Takeshi Saga, Jieyeon Woo, Alexis Gerard, Hiroki Tanaka, Catherine Achard, Satoshi Nakamura, Catherine Pelachaud

An Adaptive Virtual Agent Platform for Automated Social Skills Training

TAKESHI SAGA, Nara Institute of Science and Technology, Japan
JIEYEON WOO, ISIR - Sorbonne University, France
ALEXIS GERARD, ISIR, France
HIROKI TANAKA, Nara Institute of Science and Technology, Japan
CATHERINE ACHARD, ISIR - Sorbonne University, France
SATOSHI NAKAMURA, Nara Institute of Science and Technology, Japan
CATHERINE PELACHAUD, CNRS - ISIR - Sorbonne University, France

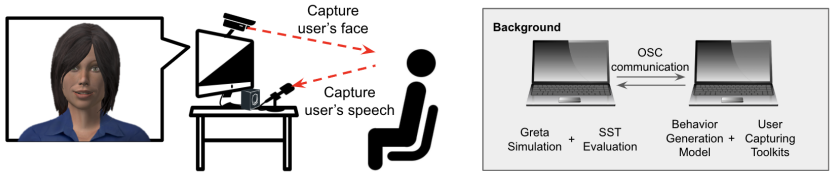

Fig. 1. System setup of the proposed system. The system runs the user-capturing toolkits, behavior generation model, Greta, and SST evaluation.

Interlocutors adapt their verbal and nonverbal behaviors as signs of engagement during face-to-face interaction. We aim to build engaging Socially Interactive Agents, SIAs, that can adapt their behaviors during interaction. With an adaptive behavior generation model, we drive SIAs' upper face and head movements in real-time. We evaluate this platform through a scenario for Social Skills Training, SST.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; • **Computer systems organization** → **Real-time system architecture**.

Additional Key Words and Phrases: virtual agent, adaptation, social skills training

ACM Reference Format:
Takeshi Saga, Jieyeon Woo, Alexis Gerard, Hiroki Tanaka, Catherine Achard, Satoshi Nakamura, and Catherine Pelachaud. 2023. An Adaptive Virtual Agent Platform for Automated Social Skills Training. In . ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

People use verbal and nonverbal behaviors to communicate with others. They adapt their behaviors to the other interlocutors to indicate their engagement in the conversation. This adaptation ability, which strengthens the bond between interlocutors, is important in human-agent interaction. In this demonstration, we propose a virtual agent system with adaptive real-time behavior generation

ICMI '23, October 9–13,2023, Paris, France

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

for upper face and head movements, which can provide face-to-face multimodal interaction [1]. As an example of an application system, we show automatic social skills training using the system. Social skills training (SST) is a validated rehabilitation program used in psychiatric hospitals or job training center [1]. It is geared toward persons suffering stress related to social anxiety. According to Bellack's definition, there are four basic social skills to be acquired to manage everyday activities: paying attention to an interlocutor's speech (LISTEN), conveying positive feelings (TELL), requesting something from an interlocutor (ASK), and refusing a request (DECLINE). SST involves offering these persons to act out scenarios related to these situations. We propose a virtual agent system for SST, which generates real-time adaptive behavior and renders SST evaluation and feedback.

2 SYSTEM SETUP

Figure 1 shows our system setup. A human participant sits in front of a screen displaying a close-up of a virtual agent (mainly its head and face). For behavior generation, the system captures the user's voice with a microphone and the user's face (head movement, gaze, and facial expressions) with a 1080p RGB webcam. We also use a speakerphone, a device that serves as both speaker and microphone, for the user's speech recognition and the agent's speech synthesis. To record the interaction, we use a microphone-embedded webcam with 1080p resolution. Two computers, with 2.4GHz Intel Core i9 mounted with NVIDIA Quadro RTX 4000 and 64GB RAM, are used to run the system. The first computer runs the Greta platform [3], an open-source virtual agent platform with an embodied conversational agent which can communicate verbally and nonverbally in real-time. The speech of the agent is rendered using the CereProc speech synthesizer. The second computer runs the ASAP model [6] generating adaptive agent's behavior in real-time, along with the user capturing toolkits of OpenFace (facial feature extraction) and openSMILE (prosodic feature extraction). We also implement the SST evaluation system on the first computer, which uses OpenFace, OpenPose (body points extraction), Praat (prosodic feature extraction) for feature extractions (for more details, see [4]). The computers communicate with each other via the OSC (Open Sound Control) communication protocol.

The behavior generation performs in real-time for a single synchronized system loop (acoustic and visual features extraction, and behavior generation) within the $0.04sec$ (time of a frame). Approximately $7GB$ RAM is required to run the system.

3 REAL-TIME ADAPTIVE BEHAVIOR GENERATION

To increase the user's engagement, real-time adaptive agent behavior is generated with the ASAP model [6] by constantly adapting to the user's behaviors. It models the interpersonal relationship of multimodal signals with the self-attention pruning technique to give the agent the reciprocal adaptation capability. It receives previous visual (eye movements, head rotations, six upper face Action Units (AUs) [2] of *AU1*, *AU2*, *AU4*, *AU5*, *AU6*, and *AU7*, and that of the smile *AU12*) and audio (fundamental frequency, loudness, voicing probability, and 13 Mel-frequency Cepstral Coefficient (MFCC)) features as input. We used features extracted with OpenFace and openSMILE from both the human user and the agent. The ASAP model generates synchronized agent's adaptive behaviors (outputting facial AUs and head/gaze movements) for every frame (at each time-step) at $25fps$.

4 INTERACTION MANAGEMENT

The system controls dialogue turns with a rule-based management mechanism, which enables smooth turn-taking. The system recognizes the ending of the user's speech by detecting the silence

[1]Demo video link: https://youtu.be/O_i1PrnV-_o

through a speech recognition module. The user's utterance is not taken into account for the dialogue management during the agent's speech.

The dialogue is managed by the Flipper [5] engine. The user's utterance text obtained by Google ASR [2] is passed to Flipper via ActiveMQ. The system selects predefined response utterances depending on keywords extracted from the user's utterance. The agent's adaptive behavior is generated with the ASAP model [6] while Greta's internal behavior realizer module computes the agent's lip movement.

5 APPLICATION: SOCIAL SKILLS TRAINING SYSTEM

A basic human-human SST starts with ice breaking and briefly introduces the SST goal corresponding to each training target skill to maximize the training effect. Then, the participant acts in role-play situations with the trainer. After that, the trainer gives feedback on the role-play toward further improvement. They may repeat the role-play feedback loop several times for better training effects. We replace this process with a fully-automated system.

Each session ends with evaluation feedback. It is based on Saga's system [4], comprised of evaluation and feedback modules with small modifications. The evaluation module estimates component scores ranging from one to five on the following three aspects: eye contact, facial expression, and vocal variation. For the estimation, we used random forest models based on multimodal features (average voice intensity, F0 frequency, smile, head poses, nodding, facial action units, and gestures. See [4] for more details). The feedback module selects a set of pre-defined sentences to reflect users' nonverbal behaviors during the interaction.

6 CONCLUSION

We propose a virtual agent system that can interact with human users in real-time. The agent adapts its behavior to that of the user. We demonstrated the flexible integration ability of our system by showcasing its use with the SST scenario.

ACKNOWLEDGMENTS

Funding was provided by the JST Core Research for Evolutionary Science and Technology, the Agence Nationale de la Recherche (ANR-JST CREST, TAPAS project, Grant No. JPMJCR19A5 for JST-CREST, ANR-19-JSTS-0001 for ANR).

REFERENCES

[1] Alan S. Bellack, Kim T. Mueser, Susan Gingerich, and Julie Agresta. 2004. *Social Skills Training for Schizophrenia: A Step-by-Step Guide* (2 ed.). Guilford Press, 370 Seventh Avenue, Suite 1200, New York, NY 10001-1020.

[2] Paul Ekman and Wallace V Friesen. 1976. Measuring facial movement. *Environmental psychology and nonverbal behavior* 1, 1 (1976), 56–75.

[3] Radoslaw Niewiadomski, Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. 2009. Greta: an interactive expressive ECA system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. ACM, New York, NY, United States, 1399–1400.

[4] Takeshi Saga, Hiroki Tanaka, Yasuhiro Matsuda, Tsubasa Morimoto, Mitsuhiro Uratani, Kosuke Okazaki, Yuichiro Fujimoto, and Satoshi Nakamura. 2023. Automatic evaluation-feedback system for automated social skills training. *Scientific Reports* 13, 1 (April 2023), 6856.

[5] Jelte van Waterschoot, Merijn Bruijnes, Jan Flokstra, Dennis Reidsma, Daniel Davison, Mariët Theune, and Dirk Heylen. 2018. Flipper 2.0: A pragmatic dialogue engine for embodied conversational agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. ACM, New York, NY, United States, 43–50.

[6] Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. 2023. ASAP: Endowing Adaptation Capability to Agent in Human-Agent Interaction. In *28th International Conference on Intelligent User Interfaces*. ACM, New York, NY, United States, 464—-475.

[2]https://cloud.google.com/speech-to-text