# Conducting Cognitive Behavioral Therapy with an Adaptive Virtual Agent

Jieyeon Woo, Michele Grimaldi, Catherine Achard, Catherine Pelachaud

# Conducting Cognitive Behavioral Therapy with an Adaptive Virtual Agent

Jieyeon Woo
woo@isir.upmc.fr
ISIR - Sorbonne University
Paris, France

Michele Grimaldi
michele.grimaldi@isir.upmc.fr
ISIR - Sorbonne University
Paris, France

Catherine Pelachaud
pelachaud@isir.upmc.fr
CNRS - ISIR - Sorbonne University
Paris, France

Catherine Achard
achard@isir.upmc.fr
ISIR - Sorbonne University
Paris, France

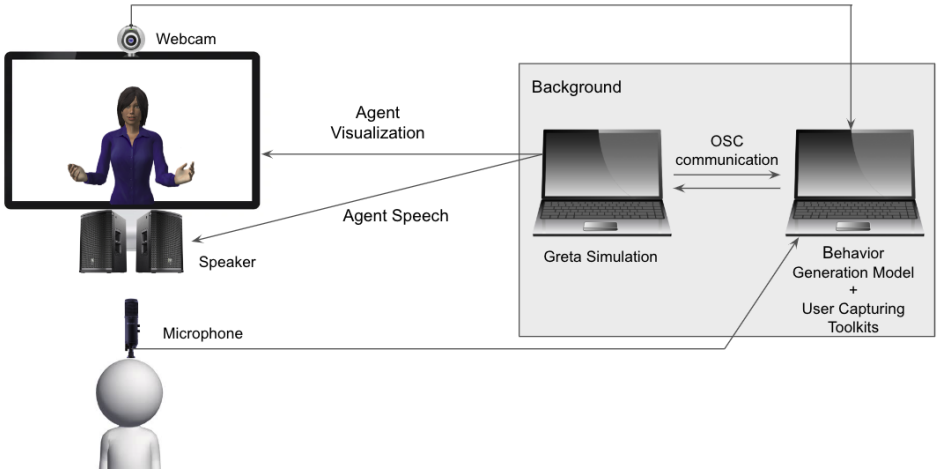


**Figure 1: The proposed system delivers CBT to the user via a virtual agent that adapts its behaviors to the user in real time. It captures the user's face with a webcam and the user's speech with a microphone. The virtual CBT agent is displayed in front of the user on a monitor and its speech is rendered via a speech synthesizer.**

## ABSTRACT

When conversing, people adapt their behaviors to one another to show their engagement. Virtual agents, acting as interaction partners, should also adapt to their interlocutors in real time. In this paper, we introduce a virtual agent delivering Cognitive Behavioral Therapy (CBT) and adapting its behaviors in real time. The system focuses on the real-time generation of adaptive behavior and management of natural CBT dialogue.

## CCS CONCEPTS

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


*IVA '23, September 19–22, 2023, Würzburg, Germany*

ACM ISBN 978-1-4503-9994-4/23/09.
https://doi.org/10.1145/3570945.3607334

## KEYWORDS

Virtual agent, adaptation, cognitive behavior therapy

**ACM Reference Format:**
Jieyeon Woo, Michele Grimaldi, Catherine Pelachaud, and Catherine Achard. 2023. Conducting Cognitive Behavioral Therapy with an Adaptive Virtual Agent. In *ACM International Conference on Intelligent Virtual Agents (IVA '23), September 19–22, 2023, Würzburg, Germany*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3570945.3607334

## 1 INTRODUCTION

People express their thoughts and feelings to others by passing their message through behaviors that are verbal (spoken and written words) and nonverbal (gestures and voice prosody). Behaviors are not only sent but also adapted depending on those of the interlocutor. This allows the interacting partners to indicate their engagement [3, 7] and to build a stronger bond between them [10].

Virtual agents should make their human users engaged in the conversation when interacting with them. We hypothesize that one way for the agent to achieve such a goal is for it to adapt to the human behavior. They should be capable of generating continuous

adaptive behavior in real time and converse fluidly by managing their dialogue.

In this demonstration, we propose a virtual agent adapting in real time with a natural flow of dialogue [1]. Our virtual agent system can provide face-to-face multimodal interaction. For this demonstration, the agent delivers Cognitive Behavioral Therapy (CBT) [2]. CBT is a medical treatment of cognitive restructuring which helps people to identify and correct their automatic thoughts. Thoughts sometimes come up instantaneously and outside of conscious awareness in response to a trigger (action or event). These thoughts, called automatic thoughts, happen unconsciously, making us unaware of them but they still affect our mood. They are often irrational and harmful, and elicit negative emotion or misleading positive emotion. Recognizing these negative automatic thoughts and rectifying them with balanced and rational thoughts to improve people's moods are the key aspects of CBT. The agent's adaptive behaviors are generated with a computational model (ASAP model [14]) and the dialogue flow is managed throughout the entire CBT session.

The contributions of this paper are the following:

- We built a virtual agent which adapts its behavior in real time and assures natural dialogue flow.
- CBT is delivered through the display of the agent's adaptive behavior.

## 2 SYSTEM SETUP

### 2.1 Dialogue Scenario

For the utterances spoken by the virtual agent, the CBT scenario presented in [11] is used. The scenario consists of 14 fact-finding questions (virtual agent's questions) to help the user to rectify their automatic thought. It also includes questioning for the identification of the user's automatic thoughts. This is done to help the user to identify their automatic thought which is not always evident.

### 2.2 Experimental Setup

For the experimental setup, the system displays a virtual agent in front of the user to deliver CBT, as depicted in Figure 1. The user speaks into a pin microphone and the user's face (head movement, gaze, and facial expressions) is captured by a 1080p RGB webcam. The virtual agent is shown on a monitor and its spoken utterance is rendered using a speech synthesizer. 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 [9], an open-source virtual agent platform with an embodied conversational agent which can communicate verbally and nonverbally in real time. The second computer runs the behavior generation model (ASAP model [14]) generating adaptive agent's behavior in real time, along with the user capturing toolkits of OpenFace [1] (facial feature extraction) and openSMILE [6] (prosodic feature extraction). The computers communicate with each other via the OSC (Open Sound Control) [2] [15] communication protocol.

[1]Demo video link: https://youtu.be/9aZeSUxhf60
[2]https://opensoundcontrol.org

### 2.3 System Performance and Specifications

The proposed system performs in real time with an execution time of $0.04s$ for a single system loop ($0.03s$ for perception with OpenFace at $30fps$ and openSMILE at $100Hz$, $< 0.01s$ for behavior generation, and $< 0.01s$ for communication and visualization) with the signals all synced without any delay.

For the system to function, a memory space of approximately $7GB$ is required for the setup and use ($2GB$ for platform visualization, $2GB$ for OpenFace and openSMILE, and $3GB$ for execution and data saving). Hardware specifications are as follows. Two computers with $2.4GHz$ Intel Core i9 mounted with NVIDIA Quadro RTX 4000 and $64GB$ RAM.

## 3 REAL-TIME ADAPTIVE BEHAVIOR GENERATION

The virtual agent generates real-time adaptive behavior with the aim to increase the user's engagement. To render such agent behavior, constantly adapting to that of the user, the Augmented Self-Attention Pruning (ASAP) model [14] is employed. The model endows the agent with reciprocal adaptation capability by modeling the interpersonal relationship of multimodal signals with the self-attention pruning technique. It receives previous visual (eye movements, head rotations, 6 upper face Action Units (AUs) [5] 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) [8]) features, from OpenFace and openSMILE respectively, of both human user and agent. The ASAP model generates the agent's adaptive behavior (outputting facial AUs, head/gaze movements) that are in sync with that of its human interlocutor. The prediction is made and realized for every frame (at each time-step) at $25fps$ via Ogre3D [3].

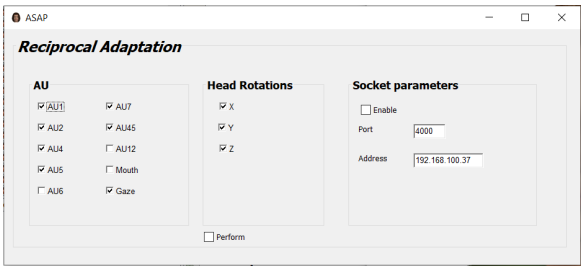


**Figure 2: The user can activate the agent behavior types that he/she wishes to display via the interactive window.**

The system also provides an interactive window that allows the selection of the types of agent behavior to be displayed, as shown in Figure 2. Before the CBT session, the activated behavior types are checked by the system and displayed. Deactivated behavior types display the default neutral behavior.

## 4 NATURAL CBT DIALOGUE MANAGEMENT

The fluid dialogue flow is directed by the system. It ensures smooth turn-taking between the user and the agent. The system constantly

[3]https://www.ogre3d.org/

checks the speaking state of both the agent and the human user. When the user stops speaking after their turn, the agent interprets this as it can take the speaking turn. This is done by detecting the silence of the user via the speaking states during his/her speaking turn.

The dialogue is managed by the Flipper [13] engine. The user's utterance text obtained by Google ASR [4] is passed to Flipper via ActiveMQ [5] [12]. The dialogue content, corresponding to the aforementioned CBT scenario [11], is chosen depending on the user's utterance. An automatic thought classifier model, Support Vector Machine (SVM) with linear kernel presented in [11] trained with the French word embeddings from Bidirectional Encoder Representations from Transformers (BERT) [4], verifies whether the utterance is an automatic thought or not. The next conversational move depends on this verification of automatic thought. The selected move (agent's speech utterance) is instantiated into the corresponding agent's lip movements and speech. These two are then combined and synced with the agent's adaptive behavior generated by the ASAP model [14].

## 5 CONCLUSION

We propose a virtual CBT agent capable of adapting its behavior in real time to the interacting user. It provides CBT by displaying real-time adaptive agent behavior generated via a computational model. Along with the continuous adaptation of the agent's behavior, the natural flow of the CBT dialogue is also guaranteed by the system. We are working on testing the system performance on CBT and studying its effectiveness in different cultures.

## ACKNOWLEDGMENTS

This work is performed as a part of ANR-JST-CREST TAPAS (ANR-19-JSTS-0001) and IA ANR-DFG-JST Panorama (ANR-20-IADJ-0008) projects.

## REFERENCES

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.
[2] Judith S Beck and Aaron T Beck. 2011. Cognitive behavior therapy. *New York: Basics and beyond. Guilford Publication* (2011), 19–20.
[3] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. 2012. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing* 3, 3 (2012), 349–365.
[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[5] Paul Ekman and Wallace V Friesen. 1976. Measuring facial movement. *Environmental psychology and nonverbal behavior* 1, 1 (1976), 56–75.
[6] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.
[7] Aman Gupta, Finn L Strivens, Benjamin Tag, Kai Kunze, and Jamie A Ward. 2019. Blink as you sync: Uncovering eye and nod synchrony in conversation using wearable sensing. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 66–71.
[8] Beth Logan. 2000. Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*. Citeseer.
[9] 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*. 1399–1400.
[10] Richard C Schmidt and Michael J Richardson. 2008. Dynamics of interpersonal coordination. In *Coordination: Neural, behavioral and social dynamics*. Springer, 281–308.
[11] Kazuhiro Shidara, Hiroki Tanaka, Hiroyoshi Adachi, Daisuke Kanayama, Yukako Sakagami, Takashi Kudo, and Satoshi Nakamura. 2022. Automatic thoughts and facial expressions in cognitive restructuring with virtual agents. *Frontiers in Computer Science* 4 (2022), 8.
[12] Bruce Snyder, Dejan Bosnanac, and Rob Davies. 2011. *ActiveMQ in action*. Vol. 47. Manning Greenwich Conn.
[13] 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*. 43–50.
[14] 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*.
[15] Matthew Wright. 2005. Open Sound Control: an enabling technology for musical networking. *Organised Sound* 10, 3 (2005), 193–200.

[4] https://cloud.google.com/speech-to-text
[5] https://activemq.apache.org

## A REQUIREMENTS

For the proposed system to work, several requirements need to be met. The requirements are as follows.

### A.1 Physical devices

The system requires the following physical devices:

- webcam: to capture the user's face;
- microphone: to capture the user's speech;
- loudspeaker: to produce the audio of the agent's speech utterance;
- monitor: to display the virtual agent (in a close-up of their face, head, and shoulders);
- 2 computers: to run the system in real time.

### A.2 Platform and Toolkits

The platform and toolkits that are necessary are the following:

- Greta platform [9]: open-source virtual agent platform with an embodied conversational agent (including a speech synthesizer) which can communicate verbally and nonverbally in real time;
- OpenFace [1]: open-source toolkit that extracts user's facial features such as head movements, gaze, face Action Units (AUs) [5], and facial landmarks;
- openSMILE [6]: open-source toolkit for extracting prosodic features such as the fundamental frequency, loudness, voicing probability, and Mel-frequency Cepstral Coefficient (MFCC) [8];
- Google ASR: automatic speech recognition which transcribes the audio of the user's utterance into written text;
- Flipper2.0 [13]: dialogue engine for conversation dialogue management;
- Ogre3D: open-source scene-oriented 3D rendering engine for animation visualization.