IAVA: Interactive and Adaptive Virtual Agent

Jieyeon Woo, Michele Grimaldi, Catherine I Pelachaud, Catherine Achard

IAVA: Interactive and 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

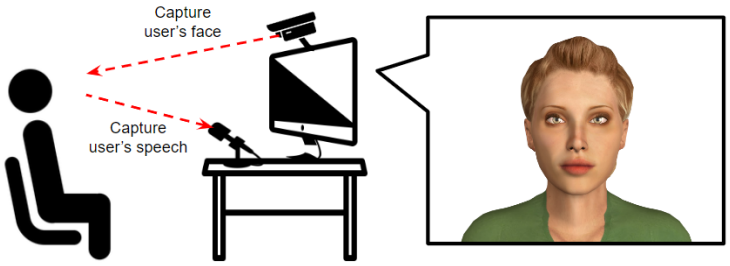

Figure 1: The proposed IAVA system enables users to interact with an interactive virtual agent with adaptation capability. It is equipped with a webcam to capture the user's face and a microphone to capture the user's speech. The virtual agent is displayed in front of the user.

ABSTRACT

During an interaction, partners adapt their behaviors to each other. Adaptation can have several functions such as being a sign of engagement and enhancing human users' interaction experience. It is important that virtual agents acting as interaction partners should continuously adapt their behaviors to those of their interlocutors in real time. This paper focuses on creating an interactive virtual agent that is capable of rendering real-time adaptive behaviors in response to its human interlocutor. It ensures the two aspects: generating real-time adaptive behavior and managing natural dialogue. We propose a system of an adaptive virtual agent and choose the e-health application of Cognitive Behavioral Therapy (CBT), which is a mental health treatment that restructures automatic thoughts into balanced thoughts, as a proof-of-concept to showcase the benefit of endowing behavior adaptation to the agent. The virtual agent adapts to the user via the display of nonverbal behaviors, which are generated via a deep learning model, throughout the whole interaction while acting as a therapist helping human users to detect their negative automatic thoughts.

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

ACM ISBN 978-1-4503-9994-4/23/09…$15.00
https://doi.org/10.1145/3570945.3607326

CCS CONCEPTS

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

KEYWORDS

Virtual agent, adaptation, real-time system

ACM Reference Format:
Jieyeon Woo, Michele Grimaldi, Catherine Pelachaud, and Catherine Achard. 2023. IAVA: Interactive and Adaptive Virtual Agent. In *ACM International Conference on Intelligent Virtual Agents (IVA '23), September 19–22, 2023, Würzburg, Germany*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3570945.3607326

1 INTRODUCTION

In a conversation, people exchange their thoughts and feelings verbally via words and nonverbally through gestures and prosody. During the exchange, people not only convey their message by sending social signals but also adapt to their interacting partner [8]. The adaptation is a key element of conversation notably for interpersonal relationships [10] and can be seen in multiple levels of aligning linguistically through the verbal channel, adapting our behaviors nonverbally, or even changing our conversational strategies (the perceived impression of ourselves such as giving a warm or competent image of oneself) [8]. This adaptation acts as a sign of engagement and strengthens the relation between the partners [12, 18].

Virtual agents, which are computer-generated virtual characters, are designed to interact with human users. One of their ultimate

goals is to make their users fully engaged in the interaction. As a way of attaining this aim, we hypothesize that agents can increase their users' engagement by adapting their behaviors depending on those of their users. It is vital for them to generate adaptive behavior continuously in real time and also to assure a fluid dialogue such as managing the turn-taking, the agent interpreting whether the user is giving their speaking turn (addressing single responses made up of several utterances linked with pauses) and whether the user is reacting with backchannels (not aiming to take the speaking turn).

The creation of such virtual agents that renders real-time adaptive behaviors in response to their human user is a hard task. The aforementioned two aspects of rendering real-time adaptive behavior and managing natural dialogue need to be addressed.

The deployment of virtual agents has been seen in numerous applications. Their use can be easily observed for education [17, 22], assistance (as a companion or guide) [6, 31], and healthcare [7, 26, 30]. Especially in [26], virtual agents have proven to be promising tools for providing medical care. In their study, they used virtual agents to provide virtual psychiatric interviews. Their study shows that virtual agents can gain high user trust and acceptance.

To showcase the benefit of endowing behavior adaptation to the agent, we choose the e-health application of Cognitive Behavioral Therapy (CBT) as a proof-of-concept.

Cognitive Behavioral Therapy (CBT) [4] is a mental health treatment that restructures automatic thoughts. The treatment helps people to recognize and change their automatic thoughts into balanced thoughts. Sometimes thoughts pop up suddenly and are unconsciously triggered by a certain action or event. As they occur unexpectedly, we are not aware of them. These thoughts that come out of our conscious awareness, which are mostly illogical and poisonous, are called automatic thoughts. They affect our mood by evoking negative feelings or misleading positive feelings. The aim of CBT is to restructure these thoughts, notably negative automatic thoughts, to reduce the effects of negative thoughts and to brighten people's moods. This is done by identifying biased and mistakenly perceived ideas and helping people to rectify their thoughts through some fact-finding questions.

In this paper, we propose a novel architecture, Interactive and Adaptive Virtual Agent (IAVA), that allows computing in real time the behavior of an agent from the behavior of its human interlocutor to simulate the dynamic behavior adaptation between the interlocutors. IAVA endows the agent with adaptation capability. The agent adapts to the user via the display of interactive nonverbal behaviors, which are generated via a deep learning model (ASAP model [36]) throughout the whole interaction. For our IAVA system, we focus on developing an interactive virtual agent that is capable of generating real-time adaptive behaviors in response to its human interlocutor. It ensures the two aspects: generating real-time adaptive behavior and managing natural dialogue. The agent adapts to its interlocutor linguistically (choosing its next conversational move via the dialogue manager) and nonverbally (displaying reciprocally adaptive facial gestures via a deep learning model). In this study, as we choose CBT as our proof-of-concept, the agent acts as a therapist helping human users to detect their negative automatic thoughts.

Our system can be employed for any other applications concerning human-agent interaction. The adaptation is essential for all types of interaction to improve the communication and engagement of interlocutors. As such, our system can help in teaching/coaching (serving as a tutor), assisting (taking the role of a companion or guide), and providing medical care (delivering other types of clinical treatment). The main contribution of our architecture is its enablement of adaptive behavior for real-time interaction and its possibility of usage in various applications.

The rest of the paper is organized as the following: Section 2 summarises the literature; Section 3 introduces the main system functionalities; Section 4 describes the requirements of the proposed IAVA system; Section 5 explains the details of the system architecture; Section 6 presents the novelty brought by our system; Section 7 discusses the future work; and Section 8 concludes the paper.

2 RELATED WORK

Previous works have been done on developing embodied conversational agents (virtual agents and robots) that adapt their behaviors according to those of their human users. The agents show adaptation at different levels (verbal, nonverbal, and/or conversational strategy) and forms (such as backchannels or mimicry). They are used for various applications: teaching/coaching, assisting, and providing medical care.

Conversational agents have the common goal of improving communication and the user's experience (engagement, rapport, and liking). Several works have focused on enhancing the interaction itself and the user experience. Huang et al. [20] created a virtual agent that produces visual backchannels for the role of a listener learned from conditional random fields (CRFs) through gaze, prosody, and lexical features. They demonstrated that such an agent reinforces the rapport that it builds with the human interlocutor and is perceived as more natural. Bailenson and Yee [2] proposed a virtual agent based on the mimicry present in human-agent interaction. They rendered the agent's mimicry behavior by imitating the user's head movements (delay of 4*s*). Their results show that a mimicking agent is perceived as more positive and persuasive than that without mimicry. Ritschel et al. [29] looked into the influence of the robot's personality through linguistic style. A reinforcement learning model was used to model the robot's personality which adapts to the user's engagement level (estimated from gaze and posture). They demonstrated that a robot adapting its personality can improve the user's engagement. Weber et al. [35] investigated the adaptation of the user's sense of humor by proposing a joke-making robot. The real-time adaptation was based on the user's smile and laughter using reinforcement learning without explicit user feedback. They were able to significantly perform better in terms of amusement level by making jokes adapted to the user's humor compared to those that were presented in a random fashion.

Agents have also shown their usefulness in teaching and coaching. Anderson et al. [1] propose a virtual agent framework for social coaching in job interviews that adapts to the user's multimodal signals (face and hand gestures). It generates the virtual recruiter's nonverbal behaviors with predefined animation commands. Their system facilitates self-reflection and provides more flexible and personalized coaching. Pereira Santos et al. [25] developed an embodied agent for obstetric simulation training. The agent plays the

role of a digital patient and its facial expressions are adapted in real time. The agent behavior is commanded by the user via on-screen controls at each frame.

The usage of agents can also be seen for assistance serving as a guide or companion. Biancardi et al. [6] built a virtual agent that is capable of adapting its behaviors when interacting with a human interlocutor, serving as a virtual museum guide, with the goal to optimize the user's engagement. They observed that an adaptive agent is more positively perceived than a non-adaptive agent. Their system adapts at different levels of: behavioral and conversational which display the agent behavior selected from a set of possible pre-scripted ones, and signal levels displaying predicted behaviors across a certain time window [13]. Sidner et al. [31] developed a real-time architecture applicable to companion agents (virtual agents and robots) for the elderly with the goal to provide companionship and reduce isolation. The agent interacts with a human user through dialog and adapts its gesture via face detection and motion perception of the user's behavior.

The employment of agents has also been seen for medical applications. Raffard et al. [27] looked into the effect of virtual agents displaying mimicry (delay varying between $0.5s$ and $4s$) with participants with schizophrenia and healthy ones. They observed that a mimicking agent improves the rapport and interaction synchrony for both participant groups. They showed the meaningfulness of an agent mimicking in real time in enhancing human-virtual agent interaction which may lead to improvement of patients' engagement in medical treatment. Several conversational agents have also been proposed for our chosen use case of CBT treatment. Ring et al. [28] proposed an affectively-aware virtual therapist for depression counseling which is based on theories of emotions in psychotherapy. The CBT dialogue is managed with user speech input and speech-based affect detection. They also display the agent's nonverbal behavior by generating it automatically using the Behavior Expression Animation Toolkit (BEAT) [11]. They demonstrate the potential efficacy of affectively aware agents in guiding users through CBT sessions. More recently, Shidara et al. [30] implemented a virtual agent that helps users to identify and evaluate automatic thoughts. They look at mood improvement and study its relation with the user's facial expressions. They remark that using fact-finding questions for the CBT dialogue of the virtual agent (to evaluate automatic thoughts) significantly ameliorated their users' moods.

Our goal is to develop a virtual agent for real-time agent behavior adaptation by automating the generation of adaptive behavior and displaying it at the frame-level.

3 IAVA FUCTIONALITIES

IAVA assures interactive communication with its two main functionalities which are as follows.

3.1 Real-time adaptive behavior generation

Real-time adaptive behavior is generated with the aim to favor user engagement in the interaction. IAVA produces agent's behavior that is adapted to that of the user. To render such adaptive agent behavior, we integrate the Augmented Self-Attention Pruning (ASAP) model [36]. The model endows the agent with reciprocal adaptation capability and generates the next agent's behavior using the previous visual and audio features of both the human user and the agent. The agent's behavior is predicted for every frame. The integration of the ASAP model and the rendering of the agent's behavior at the frame-level is further detailed in Section 5.

3.2 Natural dialogue management

The management of natural dialogue ensures the fluid flow of the interaction. For our proof-of-concept, the CBT scenario presented in [30] is employed as the dialogue content. As in [30], the user's response is verified by an automatic thought classifier model to check if it corresponds to an automatic thought. The next conversational move is selected depending on whether the user's answer is an automatic thought or not. The smooth turn-taking between the user and the agent is also guaranteed. The agent is able to assure the dialogue flow by managing the turn-taking. The technical details are later elaborated in the upcoming Section 5.

In order to perform the aforementioned functionalities of real-time adaptive behavior generation and natural dialogue management, the system needs to do the following:

- detect the user's head movements and facial expressions;
- capture the user's speech to get the content of the user's utterance and the intonation;
- know when the user is speaking and when the agent can respond back to the user (i.e. manage turn-taking).

4 REQUIREMENTS

The proposed IAVA system requires a virtual agent platform and several toolkits. The required platform and toolkits are necessary for each main component (virtual agent visualization, adaptive agent behavior generation, and dialog/turn-taking management) for our architecture. For each one of them, we use state-of-the-art technologies that are adequate for our goal of making a real-time virtual agent. The chosen requirements are as follows.

Greta platform. The Greta platform [23] is an open-source virtual agent platform. It models a real-time autonomous three-dimensional embodied conversational agent capable of communicating verbally and nonverbally. It can simultaneously talk and display nonverbal behaviors such as facial expressions, gestures, gaze, and head movements. It defines the agent's communicative intentions and behavior based on the architecture of the SAIBA framework [34]. For the animation, it follows the MPEG4 [24] animation standards. IAVA is built upon the Greta platform to render concurrently verbal and nonverbal behaviors.

OpenFace. OpenFace [3] is an open-source toolkit that extracts facial features such as head movements, gaze, face Action Units (AUs) [15], and facial landmarks.

openSMILE. openSMILE [16] is an open-source toolkit for audio feature extraction. The prosodic features such as the fundamental frequency, loudness, voicing probability, and Mel-frequency Cepstral Coefficient (MFCC) [21] can be obtained.

Automatic Speech Recognition. Automatic Speech Recognition (ASR), also known as Speech-to-Text (STT), is the technology of transcribing the audio of spoken words into written text. Full

phrases are identified and converted into text as the user is speaking. Its employment can be seen in conversational AI assistants such as Google, Alexa, and Siri. The proposed system uses the Google ASR [1].

Flipper2.0. Flipper2.0 (or Flipper) [33] is a dialogue engine that enables flexible management of conversation dialogues. The CBT dialogue is managed by Flipper to direct the conversational flow.

Ogre3D. Ogre3D [2] is an open-source scene-oriented 3D rendering engine. It is integrated and used to visualize the final animation of the virtual agent in the Greta platform.

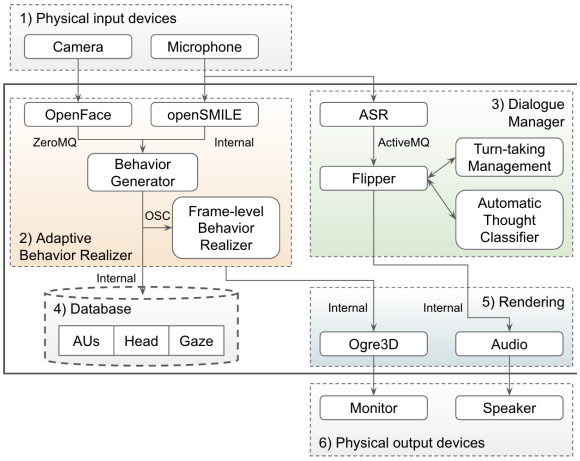

Figure 2: IAVA system consists of 6 parts: 1) physical input devices; 2) Adaptive Behavior Realizer; 3) Dialogue Manager; 4) database; 5) rendering; and 6) physical output devices.

5 SYSTEM ARCHITECTURE

IAVA is composed of six parts, as illustrated in Figure 2, which are: 1) physical input devices; 2) Adaptive Behavior Realizer; 3) Dialogue Manager; 4) database; 5) rendering; and 6) physical output devices.

5.1 System Inputs and Outputs

Our system makes use of various physical devices as input and output, and communicates multiple signals via different communication protocols.

5.1.1 Physical devices. The system uses a 1080p RGB webcam to capture the user's face, a pin microphone to capture the user's speech, a speakerphone to render the agent's speech utterance, and a monitor to display the virtual agent (in a close-up of their face, head, and shoulders) as shown in Figure 1.

5.1.2 Signals. The input and output signals communicated within the system are as follows.

Visual features: The visual features of the user are extracted in real time at $30fps$ by processing the webcam-rendered images of the user using OpenFace. To be more specific, the visual features of eye movements (around the x and y axes), head rotations (around the x, y, and z axes), 6 upper face AUs (which are *AU1*, *AU2*, *AU4*, *AU5*, *AU6*, and *AU7*) along with that of the smile (*AU12*) are passed to the model to generate the agent behavior.

Audio features: The audio features of both human user and agent are obtained separately in real time at *100 Hz* from the user's speech captured by the microphone via openSMILE. To detail, the fundamental frequency, loudness, voicing probability, and 13 MFCCs are fed to the model for the prediction.

Utterance text: The text of the user's utterance is acquired by ASR from the microphone captured user's speech. The text utterance is given as input to the Flipper engine to manage the dialogue.

Agent animation: The agent animation realized for each frame is visualized with Ogre3D and displayed on the monitor.

Agent speech: The selected agent's speech is generated via the Greta platform's Audio module, to transform the text selected by the Dialogue Manager to audio, and the audio is rendered with the speakerphone.

5.1.3 Communication protocols. The signals are passed between different toolkits and modules via communication protocols which are:

ZeroMQ:. ZeroMQ [3] [19] is an asynchronous network messaging library that is used for distributed and concurrent systems. Messages such as binary data, serialized data, and simple strings can be sent without a dedicated message broker. In our system, it is used to transmit real-time OpenFace signals directly to the model.

ActiveMQ:. ActiveMQ [4] [32] is an open-source message broker which can foster multi-client or multi-server communication. IAVA employs ActiveMQ messages to send the user's utterance recognized by the ASR to Flipper.

OSC:. OSC (Open Sound Control) [5] [37] is a lightweight and flexible protocol for real-time message communication. The advantages of OSC are its possibility to receive signals from other computers and platforms, and its availability in multiple programming languages. Our system makes use of OSC to communicate between the computational model externally running in Python and the Ogre3D of the Greta platform operating in Java.

5.2 Adaptation Behavior Realizer

To generate real-time adaptive behavior, we implement the Adaptation Behavior Realizer (ABR) module. The ABR module consists of two main components which are the Behavior Generator module and the Frame-level Behavior Realizer module as seen in Figure 3.

5.2.1 Behavior Generator module. The Behavior Generator module integrates a pre-trained computational model, ASAP model [36], which generates the agent behavior that is reciprocally adaptative.

[1] https://cloud.google.com/speech-to-text
[2] https://www.ogre3d.org/
[3] https://zeromq.org
[4] https://activemq.apache.org
[5] https://opensoundcontrol.org

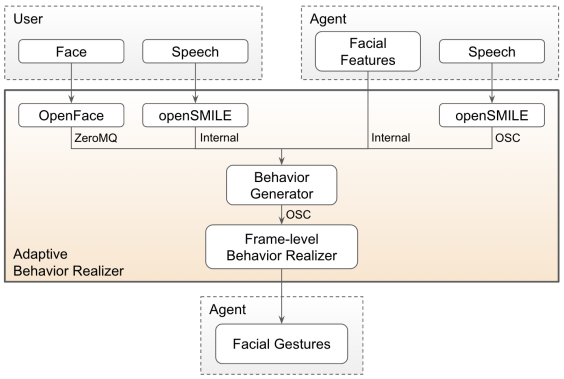

Figure 3: The Adaptation Behavior Realizer generates the agent's adaptive behavior and visualizes it at the frame-level. The agent's behavior is predicted with the Behavior Generator module via the ASAP model [36] which considers the face and speech signals from both human user and agent of the past time-steps. The generation is then rendered for each frame at $25fps$ via the Frame-level Behavior Realizer module.

The model takes the past 100 time-steps of both the human user's and the agent's behavior (visual and audio features) to predict the agent's visual behavior at the next time-step. The ASAP model learns interpersonal relationship from real human-human interactions [9]. It models the reciprocal adaptation capability and endows it to the agent from multimodal signals exchanged within a dyadic interaction with its data augmentation and self-attention pruning techniques. It generates the agent's adaptive behavior (outputting facial AUs and head/gaze movements) while assuring movement continuity via autoregressive adaptive online prediction for every frame (at each time-step) at $25fps$.

To obtain the agent's behavior at $25Hz$, the Behavior Generator module first extracts the features individually with different sampling rates as the following:

- User's audio features via openSMILE at $100Hz$ and communicated internally;
- User's visual features via OpenFace at $30Hz$ and communicated with ZeroMQ;
- Agent's audio features via openSMILE at $100Hz$ and communicated with OSC;
- Agent's visual features via the computational model at $25Hz$ and communicated internally.

We sync the different sampling rates to $25Hz$ (i.e. $25fps$) which is the computational model's sampling rate. The last 100 time-steps' signals are stocked and updated of all four feature categories with internal objects for the agent's behavior prediction of the next time-step. Each prediction, composed of the agent's facial expression (*AU1*, *AU2*, *AU4*, *AU5*, *AU6*, *AU7*, and *AU12*), head rotations, and gaze, is sent via OSC to the Frame-level Behavior Realizer module to display the agent's behavior at the frame-level. After each prediction, the four feature categories of user and agent are saved into the database in a CSV format at the sampling rate of $25Hz$.

5.2.2 Frame-level Behavior Realizer. The Frame-level Behavior Realizer module receives the agent's behavior generated by the Behavior Generator module via OSC. The Greta platform's original Behavior Realizer module [23] generates the agent's behavior by passing the user's raw input data through the Intent Planner module and the Behavior Planner module. It realizes the behavior in sequences that corresponds to the command sent by the Intent Planner. Our Frame-level Behavior Realizer module, which can be seen in Figure 3, differs from the original Behavior Realizer in the sense that it enables the generation of behaviors at the frame-level (at each time-step) which allows the virtual agent to continuously show smooth behavior throughout the whole interaction. Moreover, it produces the agent's behavior directly from the raw user input data. It is also possible to select the types of agent behavior that will be displayed via an interactive window. The types of agent behavior that can be activated are the following:

- Each upper face Action Unit (*AU1*, *AU2*, *AU4*, *AU5*, *AU6*, and *AU7*);
- Smile (*AU12*);
- Blink (*AU45*) which is automatically generated internally;
- Gaze (around the x and y axes);
- Head movement along each axis (x, y, and z);
- Mouth movement.

The IAVA system checks which agent behavior types are activated, at the beginning of the interaction, and displays them. For the ones that are deactivated, the agent will show the default behavior (value of 0 for the intensity of the AUs and 0 degree for the head rotations and gaze angles). The selected combination of the agent's behavior is passed directly to the Ogre3D for rendering.

5.3 Dialogue Manager

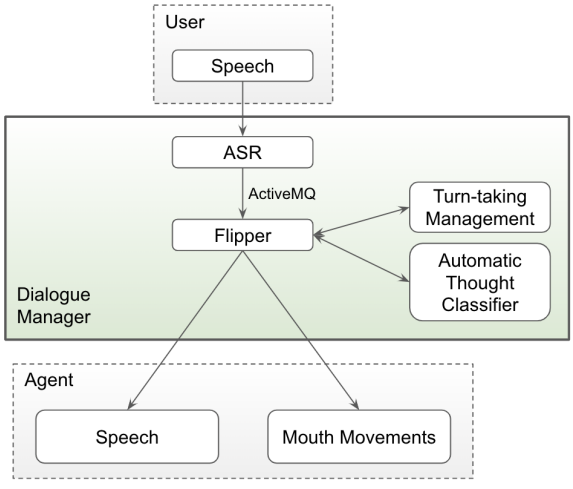

Figure 4: The Dialogue Manager manages the conversation dialogue. It selects the next conversational move while assuring the natural flow of the interaction by constantly communicating with the Turn-taking Management module. For the CBT application, the Automatic Thought Classifier module was integrated into the Dialogue Manager

The natural flow of the dialogue is managed by the Dialogue Manager. The dialogue is controlled by the Flipper engine which continuously communicates with the Turn-taking Management module, as illustrated in Figure 4, to choose the next conversational move. For the application of CBT, the Automatic Thought Classifier module is integrated into the Dialogue Manager. The process is as follows. Flipper first receives via ActiveMQ the utterance text of the user's response from the ASR. For each new utterance, it checks whether the utterance corresponds to an automatic thought via the Automatic Thought Classifier module and directs the conversational flow with the Turn-taking Management module. The modules are further explained below. The communicative intentions selected by Flipper are then instantiated into mouth movements which are combined and synchronized with the agent's speech via the Greta platform's standard treatment of passing by the Greta platform's original modules of Behavior Planner, Behavior Realizer, and Speech Synthesizer. The produced agent's mouth movements and speech are each sent to the Orgre3D and Audio module for rendering, as shown in Figure 2. This process is repeated for each user's utterance throughout the interaction.

5.3.1 Turn-taking Management module. Turn-taking is managed with the Turn-taking Management module to assure a smooth and natural flow of the conversation. The module keeps track of the speaking state of the agent and that of the human user. It handles conversational turn-taking by looking at both speaking states. By observing these two states, the agent interprets whether the user has finished answering and is giving their speaking turn (to address single responses made up of several utterances linked with pauses) and whether the user is reacting with backchannels (i.e. not aiming at taking the speaking turn), and thus decide when to take the speaking turn. After the agent decides to take the turn, it proceeds with its next conversational move.

5.3.2 Automatic Thought Classifier module. For CBT interaction, to proceed with the CBT scenario proposed in [30], a semantic analysis of the user's utterance needs to be done to identify whether the user has answered with an automatic thought or not. The structural content of the dialogue is processed by the Automatic Thought Classifier module. The module integrates the classifier model presented in [30] using the classifier algorithm of Support Vector Machine (SVM; linear kernel) with the French word embeddings from Bidirectional Encoder Representations from Transformers (BERT) [14], which is a pre-trained language model for word representations. As in [30], the raw text is tokenized and a part-of-speech tag is associated with each token. All input sentences are covered with [CLS] and [SEP] tokens, which are placed at the beginning and the ending respectively, and are fed to BERT with a hidden vector of 768 dimensions. These tags are used as the inputs of the classifier model. The model identifies automatic thoughts by performing binary classification on the user's utterance. Depending on whether the user's response is an automatic thought or not, the next agent's utterance is decided.

5.4 Animation Rendering

The final animation of the generated agent's behavior, which consists of the agent's facial gestures obtained at the frame-level by the Adaptation Behavior Realizer and the agent's mouth movements and speech produced by the Dialogue Manager is rendered by the Animation Rendering module. The agent's facial gestures and mouth movements, visualized together via Ogre3D, and the agent's utterance, produced by Greta platform's Audio module, are each passed to their corresponding physical output devices (monitor and speaker respectively).

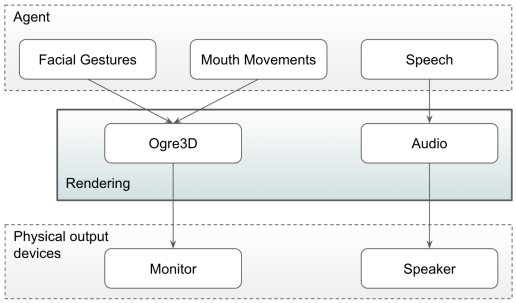

Figure 5: The Animation Rendering module displays the generated agent's behaviors, which are the agent's facial gestures obtained by the Adaptation Behavior Realizer and the agent's mouth movements sent by the Dialogue Manager, and renders the agent's speech produced by the Dialogue Manager.

5.5 System Performance and Specifications

The IAVA system works in real time, executing a single system loop every $0.04s$. The single system loop consists of:

- perception of $0.03s$ with OpenFace at $30fps$ and openSMILE at $100Hz$;
- adaptive behavior generation of $< 0.01s$ via the ASAP model;
- communication and visualization of $< 0.01s$.

All signals within the system are synced without any delay for it to function in $25Hz$, and thus generate and display the agent's behavior every $0.04s$.

For the functioning of the system, a space requirement of approximately $7GB$ is needed which consists of: $2GB$ for platform visualization, $2GB$ for OpenFace and openSMILE, and $3GB$ for execution and data saving.

In addition to the memory space requirement, hardware specifications must be met which are two computers with $2.4GHz$ Intel Core i9 mounted with NVIDIA Quadro RTX 4000 and $64GB$ RAM.

6 NOVELTY

The IAVA is designed with the previously discussed components. In the following, we highlight some applications that may result from the newly implemented modules.

6.1 Real-time Frame-level Realization

The Adaptive Behavior Realizer module allows the realization of the agent's behavior at the frame-level. The module allows the generation of the agent's behavior at each time-step (i.e. each frame). It can be easily employed by other computational models for real-time interaction. To plug in another model, a simple replacement of the pre-trained model, replacing the computational model with

another one, is sufficient. This allows different systems to also render the agent's behavior at the frame-level.

6.2 Dialogue Flow Management

The management of the dialogue flow assures smooth turn-taking between the virtual agent and the human user. This can be used for any other interaction with another scenario using the module of Turn-taking Management. If the new scenario depends on the decision of an external computational model, the module of Automatic Thought Classifier can be transformed into the new application. Thus, the dialogue flow can be managed for various applications with different dialogue scenarios and different conversational decision models.

6.3 Integration of openSMILE to Greta platform

Since the openSMILE toolkit has been integrated into the Greta platform, the prosodic features are now available within the platform itself without needing external software. The audio features of both user and agent can be extracted separately via openSMILE. For other systems that require prosodic features, the integrated openSMILE can be used to extract and make use of such features in real time.

6.4 Integration of OSC protocol to the Greta platform

With the integration of OSC protocol, the Greta platform can communicate messages in real time in a lightweight and flexible way with high accuracy. OSC can be used for communication between multiple modules for various applications.

7 FUTURE WORK

With the implementation of the IAVA system, there are a few directions for future work.

A first direction is to verify the effectiveness of providing adaptive agent behavior in real time through a user study. Ongoing work is being done for the same proof-of-concept of CBT by conducting a user study, where the user interacts in real time with the agent with adaptive behavior generated by the ASAP model [36]. The effect of displaying adaptive behavior will be assessed by comparing the display of three different experimental conditions of: reciprocally adaptive behavior, mismatched (nonadaptive) behavior, or still position. We will evaluate how showing agent adaptive behavior plays on the user's perception of the agent (agent's behavior naturalness and human-likeliness, synchrony, engagement, and rapport) and the efficacy of the CBT (improvement in the user's mood, anxiety, psychological distress, and cognitive change). We also plan to demonstrate the usefulness of our IAVA system by applying it to another application which is the Social Skills Training (SST) [5]. SST is a behavioral therapy for improving social skills in people. We envision seeing a positive effect on SST by presenting adaptive agent behavior during the training. We plan to evaluate SST performance and check whether showing such behaviors can indeed improve people's social skills.

Another direction is to perform a cultural comparison of the impact of adaptive virtual agents. Different cultures have different behavior styles. In this sense, the adapting behavior may differ across cultures and thus may have a different impact. We intend to study if the adaptive virtual agent presents such differences depending on the culture. We will check if the adaptation mechanisms can be generalized across different cultures by evaluating if our system based on a specific culture can be applied to another culture. Also, we will see if the system tuned for the comparing culture demonstrates the same adaptation mechanisms.

As a third direction for future development, we will further develop the generation of the agent's adaptive behavior. The current behavior generation method via the ASAP model [36] produces agent behavior that adapts to that of the human user. For instance, when the user is engaged in the interaction the agent will display behaviors to show its engagement by providing expressive facial expressions and head movements, which are generated by the computational model. However, this may be a limitation of our system. If the human user is inexpressive, the agent adapts to it and will tend to show expressionless behavior. Thus, the virtual agent should keep on maintaining, eliciting, and/or regaining the engagement of its users using different behaviors. We intend to assure the continuity of engagement by adding a conversation strategy to the agent to detect the occurrence of such a situation and by casting the intention on the agent's behavior during its generation.

8 CONCLUSION

In this paper, we propose a novel system of an Interactive and Adaptive Virtual Agent (IAVA) which captures the user's facial expressions and spoken utterances to show adaptive agent behavior and naturally manage the dialogue. The real-time realization at the frame rate of $25fps$ is secured for the display of the agent's facial gestures. Furthermore, the smooth flow of the interaction is ensured via the management of turn-taking. The implementation of IAVA system has added several new modules to the Greta platform which offers novel usage of the platform for various applications.

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] Keith Anderson, Elisabeth André, Tobias Baur, Sara Bernardini, Mathieu Chollet, Evi Chryssafidou, Ionut Damian, Cathy Ennis, Arjan Egges, Patrick Gebhard, et al. 2013. The TARDIS framework: intelligent virtual agents for social coaching in job interviews. In *Advances in Computer Entertainment: 10th International Conference, ACE 2013, Boekelo, The Netherlands, November 12-15, 2013. Proceedings 10.* Springer, 476–491.

[2] Jeremy N Bailenson and Nick Yee. 2005. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science* 16, 10 (2005), 814–819.

[3] 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.

[4] Judith S Beck and Aaron T Beck. 2011. Cognitive behavior therapy. *New York: Basics and beyond. Guilford Publication* (2011), 19–20.

[5] Alan S Bellack, Kim T Mueser, Susan Gingerich, and Julie Agresta. 2013. *Social skills training for schizophrenia: A step-by-step guide.* Guilford Publications.

[6] Beatrice Biancardi, Soumia Dermouche, and Catherine Pelachaud. 2021. Adaptation Mechanisms in Human–Agent Interaction: Effects on User's Impressions and Engagement. *Frontiers in Computer Science* 3 (2021), 696682.

[7] Timothy Bickmore. 2022. Health-related applications of socially interactive agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on*

Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application. 403–436.

[8] Judee K Burgoon, Lesa A Stern, and Leesa Dillman. 1995. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press.

[9] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth Andre, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. 350–359. https://doi.org/10.1145/3136755.3136780

[10] Joseph N Cappella. 1991. Mutual adaptation and relativity of measurement. *Studying interpersonal interaction* 1 (1991), 103–117.

[11] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 477–486.

[12] 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.

[13] Soumia Dermouche and Catherine Pelachaud. 2019. Generative model of agent's behaviors in human-agent interaction. In *2019 International Conference on Multimodal Interaction*. 375–384.

[14] 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).

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

[16] 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.

[17] Foteini Grivokostopoulou, Konstantinos Kovas, and Isidoros Perikos. 2020. The effectiveness of embodied pedagogical agents and their impact on students learning in virtual worlds. *Applied Sciences* 10, 5 (2020), 1739.

[18] 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.

[19] Pieter Hintjens. 2013. *ZeroMQ: messaging for many applications*. " O'Reilly Media, Inc.".

[20] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2010. Learning backchannel prediction model from parasocial consensus sampling: a subjective evaluation. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*. Springer, 159–172.

[21] Beth Logan. 2000. Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*. Citeseer.

[22] Caitlin Mills, Nigel Bosch, Kristina Krasich, and Sidney K D'Mello. 2019. Reducing mind-wandering during vicarious learning from an intelligent tutoring system. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20*. Springer, 296–307.

[23] 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.

[24] Igor S Pandzic and Robert Forchheimer. 2003. *MPEG-4 facial animation: the standard, implementation and applications*. John Wiley & Sons.

[25] Carlos Pereira Santos, Joey Relouw, Kevin Hutchinson-Lhuissier, Alexander van Buggenum, Agathe Boudry, Annemarie Fransen, Myrthe van der Ven, and Igor Mayer. 2023. Embodied Agents for Obstetric Simulation Training. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 515–527.

[26] Pierre Philip, Lucile Dupuy, Marc Auriacombe, Fushia Serre, Etienne de Sevin, Alain Sauteraud, and Jean-Arthur Micoulaud-Franchi. 2020. Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients. *NPJ digital medicine* 3, 1 (2020), 2.

[27] Stéphane Raffard, Robin N Salesse, Catherine Bortolon, Benoit G Bardy, José Henriques, Ludovic Marin, Didier Stricker, and Delphine Capdevielle. 2018. Using mimicry of body movements by a virtual agent to increase synchronization behavior and rapport in individuals with schizophrenia. *Scientific reports* 8, 1 (2018), 17356.

[28] Lazlo Ring, Timothy Bickmore, and Paola Pedrelli. 2016. An affectively aware virtual therapist for depression counseling. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) workshop on Computing and Mental Health*. 01951–12.

[29] Hannes Ritschel, Tobias Baur, and Elisabeth André. 2017. Adapting a robot's linguistic style based on socially-aware reinforcement learning. In *2017 26th ieee international symposium on robot and human interactive communication (ro-man)*. IEEE, 378–384.

[30] 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.

[31] Candace L Sidner, Timothy Bickmore, Bahador Nooraie, Charles Rich, Lazlo Ring, Mahni Shayganfar, and Laura Vardoulakis. 2018. Creating new technologies for companionable agents to support isolated older adults. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 3 (2018), 1–27.

[32] Bruce Snyder, Dejan Bosnanac, and Rob Davies. 2011. *ActiveMQ in action*. Vol. 47. Manning Greenwich Conn.

[33] 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.

[34] Hannes Vilhjálmsson, Nathan Cantelmo, Justine Cassell, Nicolas E. Chafai, Michael Kipp, Stefan Kopp, Maurizio Mancini, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, et al. 2007. The behavior markup language: Recent developments and challenges. In *Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, September 17-19, 2007 Proceedings 7*. Springer, 99–111.

[35] Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. 2018. How to shape the humor of a robot-social behavior adaptation based on reinforcement learning. In *Proceedings of the 20th ACM international conference on multimodal interaction*. 154–162.

[36] 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*.

[37] Matthew Wright. 2005. Open Sound Control: an enabling technology for musical networking. *Organised Sound* 10, 3 (2005), 193–200.