Project Proposal

Adaptive Multimodal Behavior Generation for Embodied Agents in Conversations

Brooke (Xuchen) Gong
xg54@duke.edu
November 25, 2021

Abstract

Nowadays, embodied agents are expected to play the roles of a coach, interviewer, or assistant in the areas of healthcare and education. However, their lack of emotional intelligence is the key challenge for making their applications effective and more widely accepted. Therefore, this proposal aims to develop an embodied agent (either a humanoid robot or a virtual human) that can engage in a task-oriented conversation with EQ. To endow a robot with such intelligence, I focus on giving a robot the ability to 1) adapt its behaviors to the user's responses timely and 2) have vivid and coherent verbal and non-verbal expressions. To achieve these goals, I propose a multimodal behavior generator in the robot system that instantly considers the user's current status to adjust its multimodal behaviors.

Keywords — Social robot, Human-robot interaction, Multimodal perception, Dialogue processing, Robot behavior generation

1 Introduction

Embodied conversational agents (ECAs) are computer-generated characters that simulate key properties of human face-to-face conversation, such as verbal and nonverbal behaviors [2]. Humanoid robots and virtual humans, for example, are two typical types of ECAs.

A humanoid robot can be physically present, so unlike a virtual agent, it can carry out physical tasks such as being a support worker or a working assistant. On the other hand, a virtual agent can be applied in VR/AR, and nowadays it is popularly used to construct brand ambassadors [1] because of its lifelike appearance and expressiveness. However, despite their differences, they both have wide applications in clinical diagnosis, healthcare, and education, such as mental health diagnosis, autism therapy, and interview coaching.

One advantage of applying the above types of robots in these areas is that interacting with a robot can reduce stress and fear because of its non-judgmental nature [1, 20]. Moreover, by virtue of not being humans, robots can be around the clock when life humans cannot be, 24/7.

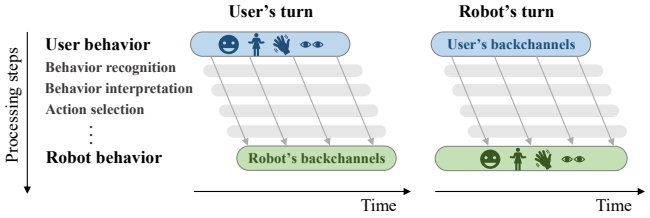

Figure 1. Proposed dynamic interaction between the robot and the user. The robot can provide backchannels as a listener and give instant adaptive behaviors as a speaker.

Therefore, the main challenge of the robots is left with whether they can engage in dynamic interactions with emotional intelligence. For example, a robot coach should provide friendly interactions and pay extra attention to the user's display of confusion, boredom, or detachment. Similarly, a mental health robot companion should keep track of the user's emotions and provide empathetic responses at appropriate moments.

2 Objectives

A socially intelligent embodied agent is expected to converse with humans interactively and vividly. Therefore, the goal of this project is to add EQ to a robot system so that a humanoid robot or a virtual human can have task-oriented conversations with a participant, ideally, as a human does. The specific task of the conversation can be generically defined for different application scenarios (e.g., assisting, diagnosing, and coaching), but the so-called human-like robot should always meet the expectations of 1) adapting its actions to the user's feedback timely and 2) expressing itself with coherent non-verbal behaviors along with its speech.

1) To endow a robot with the ability to adapt its behaviors to the present situation, I aim to develop methods to process the user's behaviors continuously and constantly incorporate this information in the robot's decision of what action to take in the next time unit. As shown in Fig. 1, when a user speaks, the robot should always be attentive and provide backchannels such as pronouncing "Huh" and

nodding at appropriate times. When it is the robot's turn to speak, it should continue interpreting the user's visual and vocal cues. Suppose the user displays confusion after watching a part of the demonstration made by the robot, for example. In that case, the robot coach should query the user's confusion and offer clarifications before continuing the demonstration.

2) To endow a robot with coherent non-verbal behaviors, I aim to develop techniques to generate faces and body movements that agree with what the robot says. For example, the emphasis made by hands and nods should be consistent, the lip shapes (if applicable) should appear pronouncing the correct spoken language, the facial expressions should agree with the emotions embodied in the voice intonations, etc.

3 Related Work

3.1 Task-oriented Embodied Agents

For the embodied agents designed particularly for a task, healthcare is the most active area for application, with the most research investigating autism spectrum disorder and mood disorders [2]. In some applications, an embodied agent plays the role of a coach, providing training for mindfulness [31], job interview [21, 23], or social interaction [22] and giving feedback on the users' performance. The second type of role the embodied agents play is a counselor. For example, in the work of Pontier et al. [24], a virtual agent is responsible for guiding users through an online version of the Beck Depression Inventory questionnaire. In addition, many social robots are particularly designed to company the elderly [25] or take care of people with dementia [26].

3.2 Continuous Dialogue Processing

The robot system needs to enable continuous dialogue processing to adapt its behaviors to its interlocutor's feedback timely.

When it is the user's turn, a robot is expected to provide empathetic responses such as backchannels and head nods. However, most of the current applications of backchannels and head nods in human-robot interaction use rule-based methods. For example, one simple method locates backchannel places at 200*ms* after a low pitch region [8]. In another work [6], a virtual human interviewer named SimSensei gives a set of pre-defined neutral backchannels encoded in rules and determines the time of providing continuation prompts based on the duration of participant speech. One exception of these rule-based methods is an early work [9], where a sequential probabilistic model is trained on the humans' prosody, words, and eye gaze during face-to-face interactions, which can output a probability of a backchannel for every frame.

When it is the robot's turn, the robot should timely adjust its actions according to the user's current affective status, engagement, etc. However, research on making timely adaptive behaviors when a robot itself is speaking is scarce. One close example is the robot Bartender, an assistant that supports multiparty interaction in a bartending scenario [4]. Similar to [5], it makes inferences from the behavioral signals of the user to quantify user state across the conversation, but that information can only be used to update the style and content of the follow-up questions. Such deferred update is acceptable in an interview or a chat, where the robot's speech or questions are relatively short. However, when the robot is expected to perform storytelling or didactic coaching, timelier behavior adaptations become necessary.

3.3 Text-to-Behavior Generation

After the robot knows what to say in response, it needs to generate coherent verbal and non-verbal behaviors. The current work that achieves this task employs two paradigms, the pipeline paradigm or the synchronous paradigm.

The pipeline paradigm first uses a text-to-speech (TTS) model to synthesize audio from text data and then uses the synthesized audio to generate non-verbal behaviors. Thanks to the increasing accessibility to public gesture generation datasets [12, 13, 14, 15], many parametric approaches, such as deterministic models [16, 17] and probabilistic models [18], have been proposed. There is also work utilizing both audio and text to generate gestures, typically having a multimodal attention block followed by an adversarial pose generator [19].

Another paradigm, the synchronous one, utilizes text data alone as input and simultaneously synthesizes audio and non-verbal behaviors. For example, an early rule-based system, Behavior Expression Animation Toolkit (BEAT) [10], uses natural language knowledge to identify the linguistic and contextual information conveyed in the text. Based on the lexical, syntactic, semantic, and rhetorical analyses, they hard code the mapping from text to the movements of the face, hands, arms, and voice intonation.

Alternatively, in terms of the statistical approach under the synchronous paradigm, two examples are found. Wang et al. [11] propose to jointly synthesize speech and gestures from text data in a single model by adding a block that synthesizes gestures parallel to the block that synthesizes audio. Since a major challenge of inferring gestures from text data is to align the two modalities, the authors align the two modalities by letting them share the learned attention, which therefore requires the model to be trained on audio-gestures paired data.

As another attempt to resolve the alignment problem, Yu et al. [7] propose to replace the end-to-end attention mechanism in the popular TTS model [27] with a duration prediction model. Such a duration prediction model can then be used for facial expression generation without relying on parallel speech and face data.

4 Proposed Project

Aimed to add EQ to an embodied agent so that it can have task-oriented conversations with humans intelligently, I propose a user-robot interaction system, as illustrated in Fig. 2. Similar to a typical social robot, this system has

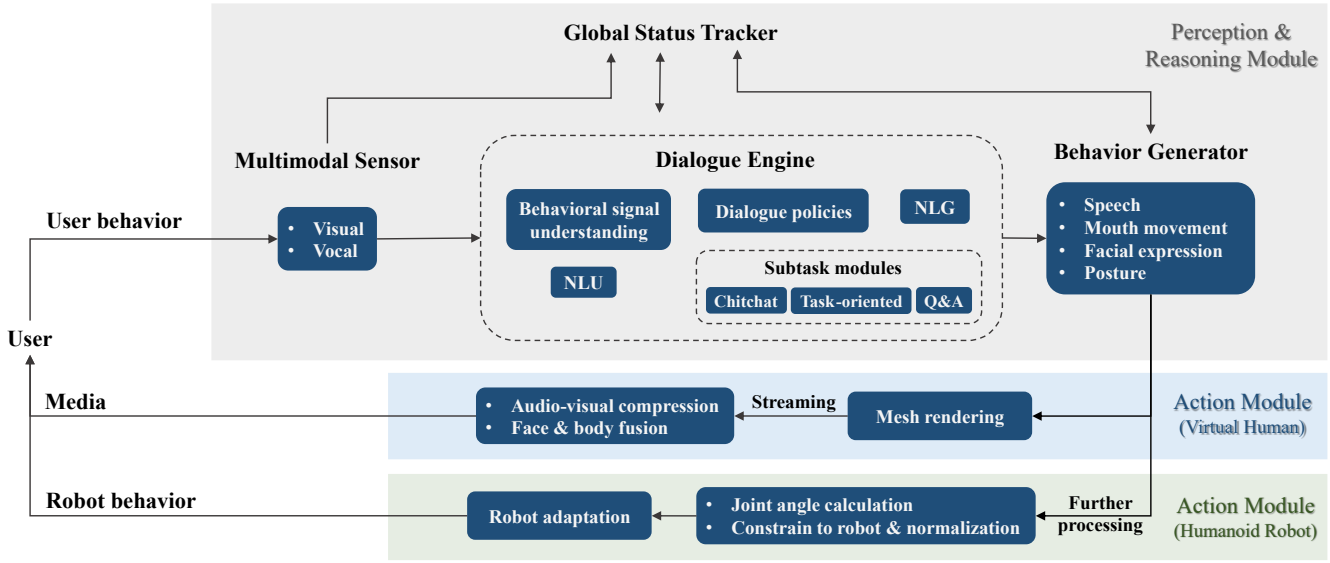

Figure 2. Proposed user-robot interaction system composed of a Perception & Reasoning module and an Action Module. The Multimodal Sensor continuously processes the visual and vocal cues from the user, sending preliminary conclusions about the user status to the Global Status Tracker. The Global Status Tracker stores and further interprets the user status and transmits signals to the Behavior Generator. The Behavior Generator generates verbal and non-verbal behaviors from the robot's language and the context information from the Global Status Tracker.

components for perception, reasoning, and action [2], with a slight difference in its union of the perception and reasoning components. The union of the sensing and reasoning step results from the "continuous adaptation" requirement because that will expect these two steps to execute both continuously and simultaneously. Moreover, in this system, the specific action module to employ depends on the type of embodied agent the task uses.

4.1 Multimodal Sensor

Once the interaction starts, the Multimodal Sensor begins to receive visual and vocal cues from the user continuously. It can have multiple functions implemented within, such as the automatic detection of facial landmark points, gaze fixations, head orientation, and body key points in real-time.

After the data acquisition, the sensor processes the cues via speech recognition, emotion recognition, etc. Since the robot system is generic and adaptable for different tasks, the importance of different cues can be defined according to the specific application scenario. For example, detecting the user's engagement is the priority for a robot coach, while the user's affective status is more informative for a mental health robot companion.

Next, after the important status information of the user is inferred, the sensor sends these statuses to the Global Status Tracker. The statuses can be a set of binary vectors that records the conclusions about the user's statuses, a collection of aligned/fused multimodal features of the user's behavioral signals, or both.

4.2 Global Status Tracker

The Global Status Tracker (GST) is a collection of variables containing the dialogue history information. In particular, it contains the user's engagement, affective status, cognitive status (could be inferred from the user's last response to the robot), etc. This tracker receives the intermediate conclusions from Multimodal Sensor and is then continuously accessed by the Dialogue Engine to help determine what action to take. This information will also be considered during the multimodal behavior synthesis since after the robot makes an initial decision about the response, the response is expected to be adaptive to the user's behaviors at present.

4.3 Dialogue Engine

The Dialogue Manager (DM) is the brain of the robot system, and it heavily relies on the knowledge base of the task of the system. DM contains a Natural Language Understanding (NLD) unit that comprehends the words said by the user and a Behavioral Signal Understanding (BSU) unit that interprets the user's non-verbal behaviors. These high-level understandings, such as the user's intent, opinion on topic, and personality, are also promptly sent to and stored in the GST, as shown in Fig. 2.

Furthermore, DM also contains dialogue policies, several sub-task modules, and a Natural Language Generator (NLG) unit. The dialogue policies are designed specifically for a task, and their role is to comprehend the dialogue progress and determine what action to take to promote the progress. By considering the information sent from the GST,

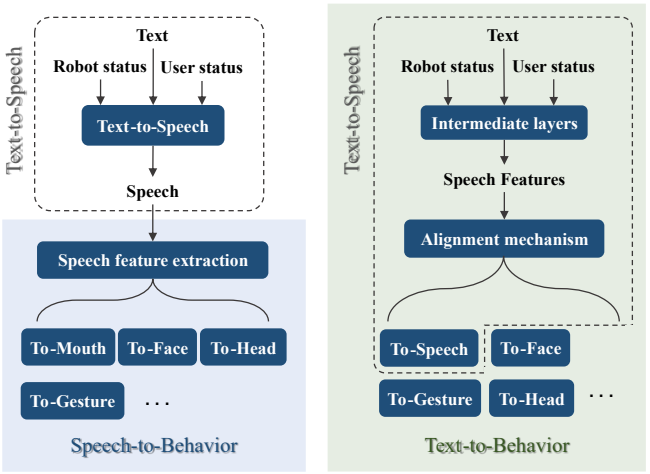

Figure 3. Text-to-Behavior, a synchronous paradigm method (right), is compared to the pipeline ones (left). In this method, non-verbal behaviors such as gestures and facial expressions are generated simultaneously with the speech from the input text and the information stored in the Global Status Tracker.

the policies both select what sub-task modules to call and utilize the selected subtask modules to carry out the task.

The sub-task modules can be either rule-based or data-driven. Take the subtask of giving empathetic responses as an example: a robot can either express empathy by imitation [29] or use a model trained on a dataset designed to automatically predict a listener's empathetic responses during face-to-face interaction, such as the OMG dataset [30]. Next, the NLG units associated with the selected subtask modules unit will convert the communication goals into natural language.

4.4 Synchronous Text-to-Behavior Generator

As concluded by Wang et al. [11], a typical neural speech-to-gesture model needs to extract audio features, such as the duration, pitch, and intensity of phonemes first [28]. Nevertheless, these features should have already been modeled in the text-to-speech model [27], so letting the speech-to-gesture model re-model features that another network has modeled is inefficient.

Therefore, in terms of the robot's gesture generation, I propose to jointly synthesize speech, facial expressions, gestures, and body movements synchronously in an integrated model, namely text-to-behavior. As shown in Fig. 3, the text data and the robot's and user's status information stored in the GST are taken as inputs to generate the acoustic features. With the text providing semantic information and the GST providing supplementary emotional and style information, the generated features should convey more contextual messages.

Next, different communicative modalities are inferred from the speech features in parallel, which are aligned through an alignment mechanism. In terms of this alignment, the traditional end-to-end attention mechanism [27] and the duration prediction model [7] will be experimented with. I will also explore methods to allow for unmatched multimodal data during training, as the duration prediction model does.

In all, this synchronous method has two advantages. First, because it generates speech and non-verbal behaviors synchronously, the updates in the GST can affect the generation of multiple modalities at present together. Therefore, it frees the model from repeatedly accessing and processing the input data. Second, sharing a learned alignment mechanism, the different modalities are expected to be more natural and coherent than those aligned by fixed time slots.

Further questions to explore:

- How to adapt a robot's behaviors to users' idiosyncratic personalities?
- How to flexibly encode different robot personalities in its multimodal behavior generator?
- For a robot's backchannels generated by statistical approaches, how to expect their effects on the user?
- How to let a robot coach understand team coordination so that a one-to-one coaching scenario can be extended to a one-to-multiple one?

6 References

[1] https://digitalhumans.com/what-are-digital-humans/

[2] Provoost, Simon, et al. "Embodied conversational agents in clinical psychology: a scoping review." Journal of medical Internet research 19.5 (2017): e151.

[3] Skantze, Gabriel. "Turn-taking in conversational systems and human-robot interaction: a review." Computer Speech & Language (2020): 101178.

[4] Foster, Mary Ellen, Simon Keizer, and Oliver Lemon. "Towards action selection under uncertainty for a socially aware robot bartender." Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction. 2014.

[5] Rizzo, Albert, et al. "Autonomous virtual human agents for healthcare information support and clinical interviewing." Artificial intelligence in behavioral and mental health care. Academic Press, 2016. 53-79.

[6] DeVault, David, et al. "SimSensei Kiosk: A virtual human interviewer for healthcare decision support." Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. 2014.

[7] Yu, Chengzhu, et al. "Durian: Duration informed attention network for multimodal synthesis." arXiv preprint arXiv:1909.01700 (2019).

[8] Ward, N., 1996. Using prosodic clues to decide when to produce backchannel utterances. In: Proceedings of the fourth International Conference on Spoken Language Processing, pp. 1728–1731.Philadelphia, USA.

[9] Morency, L.P., de Kok, I., Gratch, J., 2008. Predicting listener backchannels: A probabilistic multimodal approach. In: Proceedings of the Intelligent Virtual Agents, IVA. Springer, Tokyo, Japan, pp. 176–190.

[10] Cassell, Justine, Hannes Högni Vilhjálmsson, and Timothy Bickmore. "Beat: the behavior expression animation toolkit." Life-Like Characters. Springer, Berlin, Heidelberg, 2004. 163-185.

[11] Wang, Siyang, et al. "Integrated Speech and Gesture Synthesis." Proceedings of the 2021 International Conference on Multimodal Interaction. 2021.

[12] Paul Bremner, Oya Celiktutan, and Hatice Gunes. Personality perception of robot avatar tele-operators. In Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on, pages 141–148. IEEE, 2016.

[13] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta. 2017. Creating a gesture-speech dataset for speech-based automatic gesture generation. In International Conference on Human-Computer Interaction. Springer, 198–202.

[14] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-objective adversarial gesture generation. In Motion, Interaction and Games. 1–10.

[15] Ahuja, Chaitanya, et al. "Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach." European Conference on Computer Vision. Springer, Cham, 2020.

[16] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In 18th International Conference on Intelligent Virtual Agents. 79–86.

[17] Ondras, Jan, et al. "Audio-driven robot upper-body motion synthesis." IEEE transactions on cybernetics (2020).

[18] Wu, Bowen, Carlos Ishi, and Hiroshi Ishiguro. "Probabilistic human-like gesture synthesis from speech using GRU-based WGAN." GENEA: Generation and Evaluation of Non-verbal Behaviour for Embodied Agents Workshop 2021. 2021.

[19] Ahuja, Chaitanya, et al. "No Gestures Left Behind: Learning Relationships between Spoken Language and Freeform Gestures." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. 2020.

[20] Hart, John, Jonathan Gratch, and Stacy Marsella. "How virtual reality training can win friends and influence people." Fundamental Issues in Defense Training and Simulation. CRC Press, 2017. 235-249.

[21] Smith, Matthew J., et al. "Virtual reality job interview training in adults with autism spectrum disorder." Journal of autism and developmental disorders 44.10 (2014): 2450-2463.

[22] Hopkins, Ingrid Maria, et al. "Avatar assistant: improving social skills in students with an ASD through a computer-based intervention." Journal of autism and developmental disorders 41.11 (2011): 1543-1555.

[23] Smith, Matthew J., et al. "Virtual reality job interview training for individuals with psychiatric disabilities." The Journal of nervous and mental disease 202.9 (2014): 659.

[24] Pontier, Matthijs, and Ghazanfar F. Siddiqui. "A virtual therapist that responds empathically to your answers." International Workshop on Intelligent Virtual Agents. Springer, Berlin, Heidelberg, 2008.

[25] Broekens, Joost, Marcel Heerink, and Henk Rosendal. "Assistive social robots in elderly care: a review." Gerontechnology 8.2 (2009): 94-103.

[26] Mordoch, Elaine, et al. "Use of social commitment robots in the care of elderly people with dementia: A literature review." Maturitas 74.1 (2013): 14-20.

[27] Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[28] Kucherenko, Taras, et al. "A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020." 26th International Conference on Intelligent User Interfaces. 2021.

[29] Paiva, Ana, et al. "Empathy in virtual agents and robots: A survey." ACM Transactions on Interactive Intelligent Systems (TiiS) 7.3 (2017): 1-40.

[30] Barros, Pablo, et al. "The omg-empathy dataset: Evaluating the impact of affective behavior in storytelling." 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2019.

[31] Bodala, Indu P., Nikhil Churamani, and Hatice Gunes. "Creating a Robot Coach for Mindfulness and Wellbeing: A Longitudinal Study." arXiv preprint arXiv:2006.05289 (2020).