

Evaluation of Multi-party Virtual Agents

RESHMASHREE B. KANTHARAJU, ISIR, Sorbonne Université

CATHERINE PELACHAUD, CNRS - ISIR, Sorbonne Université

The Council of Coaches project has developed a platform that aims at providing tailored and personalized virtual coaching for ageing people to support them in improving their health and well-being. This paper presents the results of the user evaluations of the technical prototype that we conducted.

Additional Key Words and Phrases: Group Cohesion, Virtual Agent Platform, Evaluation

1 INTRODUCTION

The Council of Coaches project aims to develop a tool to provide virtual coaching for ageing people to improve their physical, cognitive, mental and social health. The council consists of a number of Embodied Conversational Coaches (ECCs), each specialised in their own specific domain. They interact with each other and with the user to inform and motivate them, and discuss issues related to their health and well-being. To this aim, it is important to develop agents that will be able to handle the differences in individual goals and to overcome these differences to decide and achieve the group goal [7].

In this paper, we present two evaluation studies conducted on the platform developed in the course of the project. The first study evaluates the final technical prototype (see Sec. 2.1) developed which integrates a dialogue manager with two virtual agent platforms, i.e., Greta and ASAP realizer. The second study focuses on the evaluation of the model (see Sec. 3.1) developed to generate cohesive non-verbal behaviours for a group of virtual agents.

2 EVALUATION STUDY 1: TECHNICAL PROTOTYPE

In this section we present the evaluation of the prototype developed. A multiparty interaction scenario where both the user and virtual coaches can interact with each other and among themselves is designed. A Unity3D scene acts as the user interface for the platform and the users can interact with the agent through on-screen buttons with pre-selected responses. The latest version of the platform will be made accessible to the community

2.1 System

The initial prototype used for this evaluation implements a small dialogue manager that is able to steer the scripted dialogue and control the possible user input and user interface [4]. The dialogue manager is responsible for selecting the next move in the dialogue, controlling the user interface and listening to the feedback provided by the ASAP realizer. The Greta and ASAP platforms are used for multimodal behaviour generation and for visualising Embodied Conversational Agents (ECA) in the Unity3D engine. A system with the platform pre-installed is used for the experiment.

Joint workshop on Games-Human Interaction (GHItaly21) and Multi-party Interaction in eXtended Reality (MIXR21), July 12, 2021, Bolzano, Italy
Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
<https://www.agents-united.org>

Authors' addresses: Reshmashree B. Kantharaju, ISIR, Sorbonne Université, Paris, France, bangalore_kantharaju@isir.upmc.fr; Catherine Pelachaud, CNRS - ISIR, Sorbonne Université, Paris, France, catherine.pelachaud@upmc.fr.

2.2 Stimulus

The council of coaches consisted of four different coaches as shown in Figure ???. Each coach is modeled to have their own appearance, name, role, and expertise. The dialogues developed for the functional prototype related to the topic of weight management is used [2].

- (1) Francois (Diet Coach): proposed a healthy recipe based on the user's dietary preferences that were collected during the interaction.
- (2) Olivia (Physical activity Coach): recommended the user go for a walk around the block once a day around meal time.
- (3) Emma (Social Coach): suggested the user to have a friend or a family member accompany them during the walk.
- (4) Carlos (Peer): provided supportive dialogue emphasizing the expertise of the coaches and the efficacy of their coaching.

2.3 Questionnaire

We made use of the Godspeed questionnaire to measure the Animacy, Anthropomorphism, Likeability and the Perceived intelligence of our coaching agents. We did not utilise the Perceived safety questionnaire, as we did not expect the discussed subject to have a strong emotional impact with regards to anxiety, agitation, or surprise. Further, we modified the System Usability Scale questionnaire to suit our study. We removed the questions related to use of the product in terms of a new technology since we used a simple computer interface. However, we still retained some questions related to the ease of use. Finally, we asked two general open-ended questions about the participant's opinion of the system and the agents to capture the overall impression, and to find out if they would recommend the system to others.

2.4 Procedure

The experiment involved one user interaction that lasted approximately five minutes. The experiment was conducted in English. The scenario consisted of four coaches, where two coaches are ASAP agents and two Greta agents. Every participant interacted with the system individually. The participants were first asked to read the information letter and sign the informed consent form, as well as ask any questions they had. A brief description of the experiment and the tasks to be performed by the participant was explained. During the interaction the virtual coaches introduced themselves and provided an interactive coaching session on healthy weight management. After the interaction, the participants were asked to answer the questionnaire.

2.5 Participants

The prototype was setup at Sorbonne University and the University of Twente. We collected responses from 7 and 13 participants respectively. In total we had 20 participants, with 40% being female ($n=8$) and 60% being male ($n=12$). A total of 30% of the participants were below 55 years old while 70% were aged 55 years or older, thus falling within the primary target population for Council of Coaches. A total of 75% of the participants had never interacted with a virtual agent prior to this study.

2.6 Results and Discussion

The results indicate that participants rated the agents high on likability ($m = 3.67$) and perceived intelligence ($m = 3.63$), but did not have high score on anthropomorphism ($m = 2.76$) and animacy ($m = 2.93$). Since 70% of population



Fig. 1. A screenshot of the interaction with the coaches.

were above 55 years, their expectations of agents were probably a bit technically unrealistic. Furthermore, we need to consider the fact that 75% of them had never interacted with a virtual agent before and their perception was probably influenced by media (films).

The opinion of participants about the system and agents was mixed. While several participants mentioned enjoying using the system, and felt comfortable, few participants remarked that the gaze behaviour of agents was not genuine. This could be related to the remarks about the interaction not being human enough yet. Thus, we aim to develop a model to improve the gaze behaviour of agents during the interaction.

The agents were particularly perceived to be very friendly ($m = 4.05$) and kind ($m = 3.85$). 70% of the participants thought that the system was well integrated. 55% of the participants indicated that they would use the system frequently and 50% felt confident using the system. However, 35% of the participants reported that they would need technical assistance. Even though the interface was simple (clicking buttons), the overall system might have looked complicated for the older participants. If it was just launching an app or a browser, the user might have not felt that they required assistance. Overall, 80% of the participants said that they would recommend the system to their friends and family.

3 EVALUATION STUDY 2: COHESIVE GROUP MODEL

Group cohesion is prominent when the main goal of the group is decision making or problem solving. Cohesion describes the tendency of group members' shared bond/atraction that drives the members to stay together and to want to work together [3]. It is a group phenomenon that emerges over time in teams [8] and a key variable for effective team performance [1]. Hence, we believe cohesion is an important phenomenon for our multiparty conversation model. In this evaluation study we aim to understand the perception of cohesive behaviours of the virtual coaches by the participants. In particular, we are interested in evaluating the perceived level of cohesiveness of the group, the trust in the agents and their persuasiveness.

3.1 System

The main aim of a group behaviour model is to generate the non-verbal behaviours for all the agents present in the discussion based on their roles i.e., listener or speaker as the conversation proceeds. In this work, the goal of this component is to enable the agents to display cohesive group behaviour. In order to model group cohesion in multi-party interactions, we first annotated non-verbal behaviours i.e., gaze behaviour, facial expressions, head movement and

laughter and perceived cohesion level on 2-min video segments of the Patient Consultation Corpus (PCC) [9]. The annotated segments were grouped as either low or high cohesion based on a threshold. A one-way Anova revealed that mutual gaze, laughter and head nods are prominent cues frequently observed in high cohesion segments. Moreover, gaze and laughter (smile) information was used to automatically recognize cohesive video segments, reaching an accuracy higher than 75%. We can infer that these two cues play a very important role in perception of cohesion by external observers. For this study, we make use of head nods in addition to these two behaviours to display a cohesive group of agents. We make use of an LSTM network using Keras with optimised hyperparameters to predict speaker and listener behaviours. The input to the model is the one hot encoded gaze direction for each participant along with the binary encoding of smile and head nods. The output is translated into a BML file that needs to be executed by each agent. The model is trained on the cohesive video segments only. The model generates the gaze target every 30 frames and a BML file is selected. Also, this network triggers when an agent has to display a smile or head nod.

3.2 Stimulus

The scenario consists of the same four coaches (virtual agents)(see Sec. 2.2) interacting with each other with a different appearance. Three different dialogue samples were developed on weight loss, stress management and sleep [5]. For this evaluation study we choose only one topic, i.e., stress management based on Bales' model.

3.3 Questionnaire

We make use of two pre-study questionnaires and three post-study questionnaires. The pre-study questionnaires measure the Negative Attitude towards Robots Scale (NARS) adapted to virtual agents (4-items) and persuadability of the user (5-items). The post-questionnaire measures the cohesiveness (4-items), the credibility (3-items) and the persuasiveness (3-items) of the group.

3.4 Design

The goal of the study was to understand the impact of cohesive group of agents. We have developed a between-group study with two groups. The first group interacted with agents that display behaviours generated by a random behaviour generator. The second group interact with agents that display cohesive group behaviours generated by our model. The behaviours we focused on are gaze, smile and head nods. The agent appearance and dialogue content remain the same for both the groups. Based on the results from our previous study on persuasiveness, we assigned the role of providing advice to an older authoritative agent while the supportive coach was assigned to a younger peer coach [6]. We also made use of vicarious persuasion techniques where one agent presents an argument to persuade another agent while indirectly persuading the user. The dialogue on stress management lasted for about 3 minutes.

The study was initially planned to take place in a laboratory setting. However due to the recent health regulations, respecting the social distancing, we have modified the experimental setup. Since the technical setup is not on-the-fly and requires several software platforms to be installed along with their licensing, asking the participants (age group above 50) would have been a challenging task. Therefore, we modified our setup to present pre-recorded videos to the participant to imitate interactions in real-time. We used a survey platform to generate the flow of the conversation, provide options for the user to provide their response and based on the response selected the next video to be played.

3.5 Procedure

The participant read a general instruction form and provided their consent to take part in the study. The participants filled in the pre-study questionnaire. An introductory video of a virtual agent was presented to familiarise the user to the virtual agent, their behavioural capabilities and the type of interaction. We then started the session where the user is asked to imagine a situation and then interact with the group of agents. The user is presented with a recording of the interaction and prompted for a response when required. Once the user selects an option, the video interaction continues. Once the interaction is complete, the user is notified and the post-study questionnaire is displayed and we collected basic demographic information.

3.6 Participants

The participants were recruited online using a survey hosting platform named Prolific. We had set three specifications to recruit participants, i.e., aged above 50, proficient in English and has been diagnosed with chronic disease. In total we had 32 participants taking part in our evaluation study where 10 participants were in the age group of 51-60 and 22 were in the age group above 60. 36% of the participants were male while 64% were female.

3.7 Results and Discussion

A one-way Anova was used for the analysis of the responses. The perceived level of cohesion was slightly higher for the condition using our model in comparison to random behaviour model for all the participants ($n=30$). However, the difference was not statistically significant ($p > 0.05$). We calculated the persuadability score of each participant and retained those with a score higher than three. In total we had 16 participants equally distributed between the two conditions who reported to be persuadable. Results indicate that the perceived level of cohesion was higher for the videos generated by our model ($m=4.03$) than the random behaviour model ($m=3.53$) and the difference was slightly significant ($p=0.1$). There was no statistically significant difference between the two conditions for the perceived level of trust. We computed the mean score of trust for only persuadable participants in both conditions. Even though the rating was higher for the condition using our model, the difference was not statistically significant. We further grouped the participants based on NARS questionnaire, and we did not find any significant results. Finally, the perceived level of persuasiveness was rated equally for both the conditions with no difference.

In this evaluation study we tried to measure the perceived level of cohesion and how this in turn aspects the trust in the agents and their persuasiveness. In order to do this we designed an online evaluation study with two conditions. We used our model to generate cohesive behaviours for one condition and for the other we used a random behaviour model. We found there was no significant difference in the perceived level of cohesion for both the condition for all the participants. However, when we filtered out participants based on their persuadability score we found a slightly significant difference where participants found the condition using our model to be highly cohesive group of virtual agents. The study had to be done online with pre-recorded videos which hindered the quality of videos. Even though we tried our best to record high-quality videos, we are not sure whether the participants were able to watch them in the same setting. Since the differences in a listener executing a smile or nod is very subtle the participants might have missed it. Also, the environmental conditions could affect the results which we were not able to control. Regarding the perceived persuasiveness, we found there was no significant difference. This could be attributed to the fact that we used the same dialogue content and agents for both the conditions and only the non-verbal behaviours were different.

Some participants found the automatic text-to-speech generated audio to be very artificial which could have affected their rating. Overall, the participants found the study to be quite interesting and an enjoyable experience.

3.8 Conclusion

We described two evaluation studies conducted in the context of the Council of Coaches project. The first study focuses on the evaluation of the technical prototype. The results indicate that participants rated the agents high on likability and perceived intelligence, however, the gaze model for group interaction of the agents need to be developed. Further, the users enjoyed the interaction and said they would recommend the system to others. Using the feed back received from this study, we developed a group behaviour model that handles cohesive non-verbal behaviour generation for the group of agents with focus on gaze, smile and head nods. Results indicated that the perceived level of cohesion was slightly higher than the random behaviour model. Overall, the participants found the study to be quite interesting and an enjoyable experience. In future, we aim to expand the cohesive behaviour generation model to include other non-verbal behaviours.

ACKNOWLEDGMENTS

This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement number 769553. This result only reflects the authors’ views and the EU is not responsible for any use that may be made of the information it contains. We are grateful to our project collaborators at University of Dundee, University of Twente and Universitat Politècnica de València who were mainly responsible for development of the technical prototype.

REFERENCES

- [1] Daniel J Beal, Robin R Cohen, Michael J Burke, and Christy L McLendon. 2003. Cohesion and performance in groups: a meta-analytic clarification of construct relations. *Journal of applied psychology* 88, 6 (2003), 989.
- [2] Tessa Beinema, Harm op den Akker, Lex van Velsen, and Hermie Hermens. 2021. Tailoring coaching strategies to users’ motivation in a multi-agent health coaching application. *Computers in Human Behavior* 121 (2021), 106787.
- [3] Milly Casey-Campbell and Martin L Martens. 2009. Sticking it all together: A critical assessment of the group cohesion–performance literature. *International Journal of Management Reviews* 11, 2 (2009), 223–246.
- [4] Gerwin Huizing, Brice Donval, Mukesh Barange, Reshmashree Kantharaju, and Fajrian Yunus. 2019. of deliverable Final prototype description and evaluations of the virtual coaches. (2019).
- [5] Gerwin Huizing, Randy Klaassen, and Dirk Heylen. 2021. Designing Effective Dialogue Content for a Virtual Coaching Team Using the Interaction Process Analysis and Interpersonal Circumplex Models. In *Persuasive Technology*. Springer International Publishing.
- [6] Reshmashree B Kantharaju, Dominic De Franco, Alison Pease, and Catherine Pelachaud. 2018. Is two better than one? Effects of multiple agents on user persuasion. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 255–262.
- [7] Reshmashree B. Kantharaju, Alison Pease, Dennis Reidsma, Catherine Pelachaud, Mark Snaith, Merijn Bruijnes, Randy Klaassen, Tessa Beinema, Gerwin Huizing, Donatella Simonetti, Dirk Heylen, and Harm op den Akker. 2019. Integrating Argumentation with Social Conversation between Multiple Virtual Coaches. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) (IVA ’19). Association for Computing Machinery, New York, NY, USA, 203–205. <https://doi.org/10.1145/3308532.3329450>
- [8] Jessica M Santoro, Aurora J Dixon, Chu-Hsiang Chang, and Steve WJ Kozlowski. 2015. Measuring and monitoring the dynamics of team cohesion: Methods, emerging tools, and advanced technologies. In *Team cohesion: Advances in psychological theory, methods and practice*. Emerald Group Publishing Limited.
- [9] Mark Snaith, Nicholas Conway, Tessa Beinema, Dominic De Franco, Alison Pease, Reshmashree Kantharaju, Mathilde Janier, Gerwin Huizing, Catherine Pelachaud, and Harm op den Akker. 2021. A multimodal corpus of simulated consultations between a patient and multiple healthcare professionals. *Language resources and evaluation* (2021), 1–16.