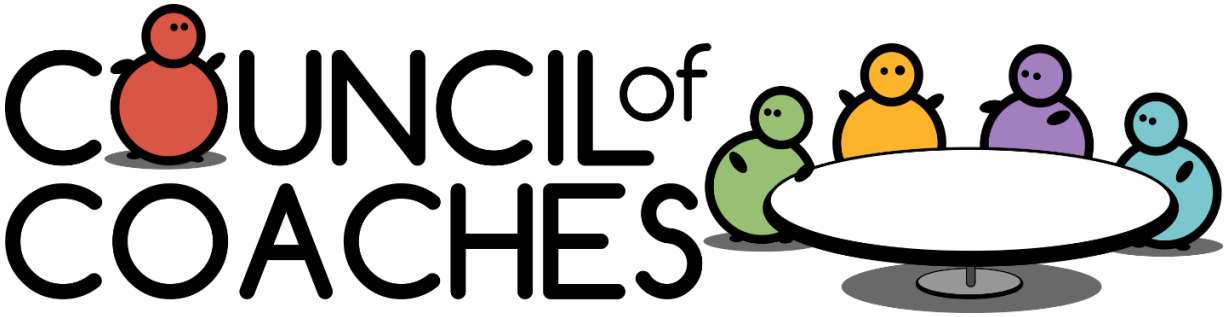

D6.6: Final virtual coach design and model

Dissemination level: Public
Document type: Report
Version: 1.0.0
Date: August 31st, 2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement #769553. This result only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains.

Document Details

Project Number	769553
Project title	Council of Coaches
Title of deliverable	Final virtual coach design and model
Due date of deliverable	August 31st, 2020
Work package	WP6
Author(s)	Daniel Davison (CMC), Gerwin Huizing (CMC), Randy Klaassen (CMC), Brice Donval (SU), Mukesh Barange (SU), Reshmashree Kantharaju (SU), Fajrian Yunus (SU), Catherine Pelachaud (SU)
Reviewer(s)	Harm op den Akker (RRD), Konstantina Kostopoulou (iSPRINT), Mark Snaith (UDun)
Approved by	Coordinator
Dissemination level	PU: Public
Document type	Report
Total number of pages	62

Partners

- University of Twente – Centre for Monitoring and Coaching (CMC)
- Roessingh Research and Development (RRD)
- Danish Board of Technology Foundation (DBT)
- Sorbonne University (SU)
- University of Dundee (UDun)
- Universitat Politècnica de València, Grupa SABIEN (UPV)
- Innovation Sprint (iSPRINT)

Abstract

The goal of this Work Package (WP6) is to design, implement, and evaluate the Human Computer Interaction aspects of the Council of Coaches. In this deliverable D6.6, we present the work done so far on the development of the final technical prototype of the Council of Coaches system. We describe the improvements made to the individual components and the group model. We also present the evaluation studies that analyse the effect of the user interface, multi devices, verbal conflict, peer agent and cohesive non-verbal cues during the multiparty interaction.

Table of Contents

List of figures

List of tables

Symbols, abbreviations and acronyms

AMI	Augmented Multiparty Interaction
ASAP	Articulated Social Agents Platform
AU	Action Unit
BML	Behaviour Markup Language
CMC	Centre for Monitoring and Coaching
COUCH	Council of Coaches
D	Deliverable
DBT	Danish Board of Technology Foundation
EC	European Commission
ECA	Embodied Conversational Agent
ECC	Embodied Conversational Coach
GUI	Graphical User Interface
HCI	Human-Computer Interaction
iSPRINT	Innovation Sprint
M	Month / Mean
MS	Milestone
RRD	Roessingh Research and Development
SAIBA	Situation, Agent, Intention, Behaviour, Animation
SD	Standard Deviation
SU	Sorbonne University
UDun	University of Dundee
UPV	Universitat Politècnica de València
UT	University of Twente
WP	Work Package

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. In the duration of the project, two prototypes of the Council of Coaches system were developed, to research the interaction between a virtual council of coaches, a functional demonstrator and a technical one. This deliverable is dedicated to the technical prototype were the council consists of a number of Embodied Conversational Coaches (ECCs), each specialised in their own specific domain. The coaches interact with each other and with the user, to inform and motivate them, and to discuss issues related to their health and well-being.

In this deliverable D6.6, in Section 3, we describe adjustments on the final prototype developed for the Council of Coaches project. In particular, we describe the improvements made on the Greta and ASAP platforms which are used for multimodal behaviour generation and for visualising Embodied Conversational Agents (ECA) into the Unity3D engine.

In the remainder of this deliverable, we present seven evaluation studies with the final Council of Coaches Technical Prototype in the context of the Council of Coaches system (see Section 4). These studies are listed here below:

- **Section 4.1:** UI Usability #1: 2D User Interfaces Evaluation
- **Section 4.2:** UI Usability #2: Speech-based user interface evaluation
- **Section 4.3:** Multi-device interaction evaluation
- **Section 4.4:** Verbal conflict presentation style impact on group discussion evaluation
- **Section 4.5:** Peer mediator coach presence and behaviour impact on group discussion evaluation
- **Section 4.6:** Gesture generation evaluation
- **Section 4.7:** Cohesive group evaluation

In Table 1 below, we provide an overview of all the studies that are described in this document, including the number and type of study subjects and evaluation methods used.

Table 1: Overview of studies described in D6.6, including the study Name, method, setting, N, number of participants younger and older than 50 years, and number of participants that have a chronic health condition.

Study	Method	Setting	*N*	Participants <50	Participants >50	Participants with health conditions (DM-II, CP)
UI Usability #1: 2D User Interfaces Evaluation	User confrontation, Qualitative, semi-structured interviews and questionnaire	Face to face	5	-	5	-
UI Usability #2: Speech-based user interface evaluation (planned)	Quantitative and qualitative, observations, questionnaires and semi-structured interviews	Face to face and online	-	-	-	-
Multi-device interaction evaluation	Qualitative, questionnaires and semi-	Face to face and online	15	9	6	-

	structured interviews					
Verbal conflict presentation style impact on group discussion evaluation	Online interaction followed by questionnaires (no interaction with researchers)	Online study	242	-	242	-
Peer mediator coach presence and behaviour impact on group discussion evaluation	Online interaction followed by questionnaires (no interaction with researchers)	Online study	243	-	243	-
Gesture generation evaluation	Online followed by questionnaires (no interaction with researchers)	Online study	28	-	28	-
Cohesive group evaluation	Online interaction followed by questionnaires (no interaction with researchers)	Online study	32	-	32	32
Totals			**565**	**9**	**556**	**32**

In this wave of evaluations, focusing on the technical Council of Coaches platform, we managed to involve a large number of participants that are in the target age group for our virtual coaching context (N=556).

2 Objectives

The main objective of this deliverable (D6.6) is to describe the final adjustment of the technical prototype of the virtual coach dialogue platform where virtual coaches are interacting with each other and the user and the evaluations of different aspects of this model.

3 Adjustments on the final prototype

3.1 ASAP and Flipper

Flipper (van Waterschoot et al., 2018) is used to generate and plan conversational intents for all coaches in the interaction. Behaviour of individual coaches is controlled by the ASAP (van Welbergen et al., 2014) and Greta behaviour realizers. To support fluent multi-agent interactions, Flipper has been further extended and integrated with the Coaching Engine (CE), the Dialogue and Argumentation Framework (DAF) and the ASAP and GRETA virtual agent platforms. This results in the following overall dialogue flow.

At the start of each interaction, based on its coaching strategy models, the CE selects a dialogue topic which is appropriate for the specific user at that time. Flipper then instructs ASAP and GRETA to load the various coaches that play a role in the dialogue into the Unity scene. DAF loads the dialogue rules and generates initial move-sets for each coach, after which Flipper selects a move and instructs the respective ASAP or GRETA agent to perform the behaviours associated with that move. Flipper monitors the progression of the move. When a move is completed, Flipper requests a next move-set from the DAF. This process repeats until an end-state in the dialogue is reached. Flipper then informs the CE that the dialogue topic has been completed and requests a next topic, repeating all the above steps.

While an individual coach is performing a move, the other coaches in the scene receive updates from Flipper about the progress of the dialogue. This includes information about the previous speakers, the current speaker, and the addressees. Additionally, each coach receives an estimate of their "eagerness to speak" based on the number of available moves which they have at that point in the dialogue. GRETA uses this information to generate socially appropriate gestures and gaze behaviours, as described in the next sections.

3.2 GRETA

3.2.1 Gesture generation model

In this section, we present a technique to predict the timing of communicative gestures based on the speech prosody. We consider two classes of gestures: beat and other gesture types (i.e. iconic, metaphoric, concrete deixis, abstract deixis, nomination deixis, and emblems). We compute the gestures class based on the speech prosody. We learn their relationship by using a recurrent neural network with an attention mechanism. The model takes a sequence as the input and yields another sequence as the output. The input is the speech prosody broken into time-steps and the output is the sequence of gesture classes. Our input features are the fundamental frequency (F0, the F0 direction score, and intensity). These three features have been found to be highly correlated with gesture production (Kendon, 1980). Pitch accent perception is also affected by intensity (Niebuhr & Pfitzinger, 2010), which suggests that intensity might also be linked to the production of gestures. Lastly, it is also observed that when there is more gesturing activity, the speaker also speaks with a higher and more variable pitch and intensity (Voigt, Podesva, & Jurafsky, 2014). These findings suggest that F0, F0 direction score, and intensity may have a relation with gestures generation. By limiting the number of features to only three features, we also mitigate the problem of the curse of dimensionality.

We use the Gest-IS English corpus (Saint-Amand, 2018). The corpus consists of 9 dialogues of a dyad, a man and a woman, discussing various topics in English in a face-to-face setting. The total duration is around 50 minutes. In those dialogues, the speakers are talking about physical description of some places, physical description of some people, scenes of two-person interactions, and instructions to assemble a wooden toy. The corpus has been annotated along different layers (Saint-Amand, 2018) (see Figure 1): gesture phases (preparation, pre-stroke hold, stroke, post-stroke hold, partial retraction, retraction, and recoil), gesture types (iconic, metaphoric, concrete deixis, abstract deixis, nomination deixis, beat, and emblems), chunk boundaries, classification annotations on whether the gesture is communicative (i.e. it contributes to the dialogue discourse) or non-communicative (i.e. it does not contribute to the dialogue discourse, such as rubbing the eyes or scratching nose), the transcription, and the transcription timestamp for each word. The gesture annotations only consider gestures which are performed by at least one hand. The transcription timestamps include the starting timestamps and

the ending timestamps of each word. We also extract eyebrow movements using OpenFace (Baltrusaitis, Zadeh, Lim, & Morency, 2018). We have extended the recurrent neural network with attention mechanism to perform the prediction (see Deliverable 6.4). The recurrent neural network with attention mechanism is an extension of the encoder-decoder model. The standard encoder-decoder model compresses the entire information from the input sequence into a fixed-length vector, namely the last encoder node. The attention mechanism adds an attention map between the encoder and the decoder. The map itself is a neuron matrix. If wij is the weight in the attention map at position $\langle i, j \rangle$, then wij represents the weight of the input at timestep i on the output at timestep j. This neuron matrix enables focusing the "attention" toward some specific input timesteps. If the input at timestep i is pertinent on the output of timestep j, then the wij would be high. Those weights are learned during the training, similar to all other weights in the network. Because this is a multi-class classification problem where the output of each timestep belongs to one of the gesture classes, we use a one-hot encoding.

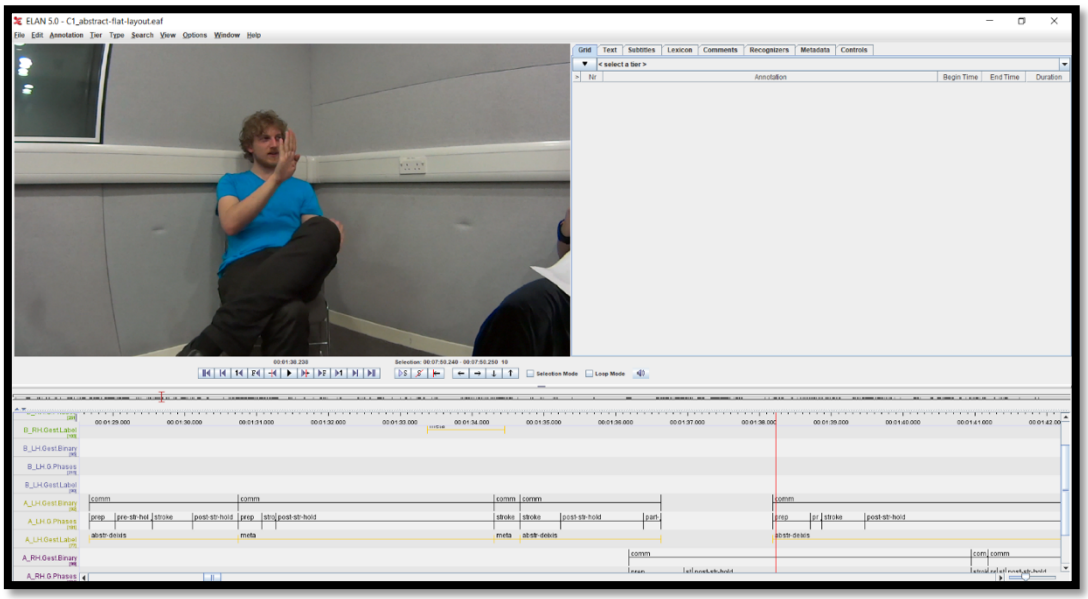

Figure 1: Gest-IS video with gesture animation.

3.3 Group Model

Group conversation is a frequently used form of communication for exchanging ideas and making decisions. Group cohesion is one of the most common phenomena that emerges over time. Cohesion in general describes the members' attraction towards the group and the desire to be a part of the group (Casey-Campbell & Martens, 2009). Identifying the most prominent social cues that increase or decrease the perception of cohesion in a group is therefore the first step towards building such a model. In particular, we focus on gaze direction, head movement, laughter and social attention. We then developed a model to generate cohesive behaviours for our agents. In this section we first present the prominent cohesive cues and then present our model.

3.3.1 Non-verbal cues

In this section, we present the data corpus used for our research and an overview of the annotations performed. A portion of the Augmented Multiparty Interaction (AMI) corpus (Carletta, et al., 2005) was annotated for task and social cohesion by (Hung & Gatica-Perez, 2010). They extracted 120 two-minute segments randomly from the entire corpus and obtained cohesion scores using a questionnaire. A group of 21 annotators were divided into 10 groups of 3, and each group annotated 12 segments. Each segment was annotated by three different annotators. A 27-item questionnaire was developed on a 7-point Likert scale. Out of the 27-items, 6-items related to task dimension and 18-items related to social dimension and 3-items marked miscellaneous. We obtained the raw annotation scores (ranging from 1: low to 7: high) from the first author of the paper and computed our cohesion labels for the 120 segments.

In order to assign a label to a given segment, we first clean up the raw scores, calculate the inter-rater agreement to ensure the ratings are reliable and finally compute the final cohesion score. To clean up the raw data, we remove the scores for the three questions grouped under miscellaneous and replace any missing scores with the median value for that segment. Next, we calculate the inter-rater agreement using the Inter-class Correlation Coefficient (ICC). We compute ICC between the three annotators belonging to a group using a one-way, average, consistency measure. The ICC value between the three annotators for all the 10 groups were above 60\% representing fair agreement for all the 120 segments.

For each segment, we calculate the average score of the 24 questions to obtain three annotator-dependent values. Then, we calculate the average of those three annotator-dependent values to obtain the final score. The obtained values range from 2.36 (lowest) to 6.30 (highest). The mean is 4.63 with standard deviation of 0.89. We categorised the 120 segments as either low cohesion or high cohesion using the mean value as the threshold. Our dataset consists of 64 segments labelled as high cohesion and 56 segments as low cohesion.

3.3.1.1 Gaze Direction

The annotation schema we follow is based on the MUMIN annotation scheme (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007) and has been adapted to multi-party interactions. We define four different gaze targets for a given participant i.e., the other three participants in the group and “others”. We use OpenFace (Baltrusaitis, Zadeh, Lim, & Morency, 2018) to automatically extract the gaze angle of the eyes and the head rotation. The tool outputs values in both horizontal (left-right) and vertical (up-down) direction for gaze. These values range between -1 to +1 based on the direction. We remove the values whose detection confidence score is lower than 95%. To reduce the manual effort, we partially automated the gaze annotation task. We classify the gaze angle data points using a k-nearest neighbour clustering method to group the data-points into four clusters. These clusters represent looking left, right, straight and down. We extract continuous data-points belonging to the same cluster and mark the time boundaries to find the start and end timestamps. We then automatically correct these annotations based on the head rotation (we consider yaw). Finally, we manually correct the annotation values and adjust the time boundaries to match our annotation schema.

3.3.1.2 Facial Action Units

We utilise OpenFace (Baltrusaitis, Zadeh, Lim, & Morency, 2018) to automatically extract the facial AUs individually for each participant of the group. The toolkit outputs the confidence score of detection along with the intensity (on a scale of 0 to 5) and the presence of 17 AUs in total. For this study, we make use of the AU intensity values as we are interested in the level of variation of facial expressions displayed by the participants during the interaction. To get the most accurate values we remove the frames where the confidence score is below 95%. To remove the noisy data, all the intensity values below 0.7 (threshold found empirically by observing the data) is replaced by zero. We calculate the duration of continuous activation of a given AU. We discard the values that were activated for less than 200ms. For each annotation, we calculate the non-zero intensity value. The final set consists of annotations of 17 AUs with the start and end time points and the average intensity of activation for that period.

3.3.1.3 Head Movement

As we are interested in understanding the head movement that conveys agreement or comprehension, for this study we focus on head nods: signals agreement, comprehension, or a positive response. We manually annotated the segments for head nods. For high cohesion segments the number of instances annotated are 1042. Similarly, for low cohesion, concordance instances are 750.

3.3.1.4 Laughter

The transcription file available with the corpus provides time marker for instances of laughter. We use these files to extract individual instances of laughter. There was some discrepancy with regard to the start and end time of laughter annotations which were corrected manually. The number of laughter instances extracted in total are 784, from which 205 instances were extracted from the 56 low cohesion segments and 579 instances from the 64 high cohesion segments.

3.3.2 Prediction model

To predict whether or not each segment belongs to low or high cohesion, we first use a Support Vector Machine (SVM). We perform a grid-search approach to find the optimum value for cost and gamma and report the results with this setting. We use a 10-fold stratified cross-validation approach. The representation of features for group interaction is not as direct as it is for dyads. One of the most common methods employed, is calculating the aggregate of the individual features to represent a single group value or calculating pairwise values and then computing their average. In this study, we use two different types of representation of features extracted from individuals to represent the group feature namely, Concatenate: features from each participant is represented individually and concatenated to form a feature vector, and Average: features from the four participants are aggregated to have one value.

The best performing feature for the prediction task was facial AUs with an accuracy of 78.33%. We performed several tests using the intensity and duration values of the 17 AUs. The performance was consistently higher when we used a subset of 10 AUs (significantly discriminatory) instead of the complete set of 17 AUs. Further, we found that AU activation duration performs better than intensity but the best performance is achieved when both the intensity and duration features are used. We also extracted AUs displayed by a given participant during speech and during non-speech. Results showed that AUs extracted during non-speech segments of the participant performs better than during the speech. From these results, we can interpret that the information conveyed by the listener is important for estimating cohesion as this is an indication of feedback.

The next best performing feature is laughter. We trained the model on both average instances and average duration of laughter. We achieved a prediction accuracy of 70.24% based on the average duration. Overall, head nods achieved moderate performance of 65%. We also trained the model on the gaze direction extracted as a group-level feature i.e., total duration of mutual gaze in a group. The total duration of mutual gaze between any two participants in the group achieved an accuracy of 75.83%. This shows that the group-level features are highly important for predicting cohesion. From the results obtained we can conclude that non-verbal social cues indeed convey information regarding the level of cohesion and they can be used for automatic prediction tasks. We also observe that group-level features are equally important as individual-level features. The results from this work contributes towards developing the computational model to simulate a cohesive group of virtual agents.

3.3.3 Behaviour Generation Model

This section describes the specifications of the module that generates cohesive turn taking and the associated non-verbal behaviours for speaker and listener agents for the Council of Coaches platform. The main aim of the group behaviour model is to choose the appropriate agent to perform the next dialogue move and 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.

The two main tasks of the group behaviour model are Speaker selector component which decides the next possible speaker based on the dialogues available and the conversation history and the second component is the non-verbal behaviour generator that will generate the appropriate behaviours and select the relevant BML files to be performed by the agents.

The speaker section selects the next speaker and updates Flipper. This component gets the input from Flipper about all the available agents with dialogue move and based on the conversation history of speakers and selects the next speaker. We implemented a LSTM model using Keras to predict the next speaker. The input to the model is the sequence of previous speakers and the output is the next speaker when the dialogue move is available. The component sends the template ID of the agent to Flipper.

The behaviour model generates the appropriate non-verbal behaviour patterns for the agents present in the coaching session. The main goal of this component is to enable the agents to display cohesive group behaviour. From the study described previously we found that gaze and laughter (smile) performed the best with an accuracy higher than 75%. Therefore, we can conclude that these two cues play a very important role. For this study, we make use of head nods in addition to these two behaviours to display a cohesive group of agents. The model follows a distributed architecture where each agent

decides the behaviour individually based on the history of conversation. We make use of an LSTM network using Keras with optimised hyperparameters. 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 the interpreted BML file ID 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. Figure 2 shows the overall architecture. In the next section we present an evaluation study of this model.

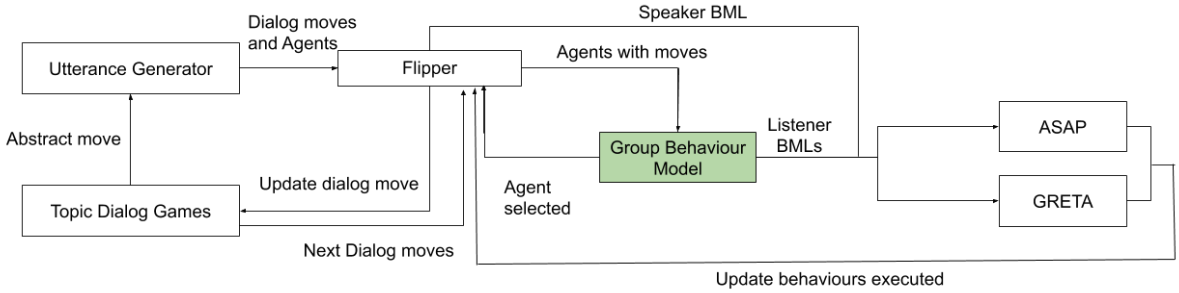

Figure 2: Architecture with the group model.

4 Evaluation studies

In this section, we describe seven evaluation studies with the technical demonstrator and functional demonstrator to evaluate the different aspects of the prototype.

The first and second study focuses on the evaluation for the user interface usability. The third study aims to evaluate the quality of the interaction with multiple devices. The fourth study measures the impact of verbal conflict presentation style on group discussion. The impact of peer coach presence and behaviour on group discussion is addressed in the fifth study. The evaluations of the gesture generator model and of the cohesive group model correspond to the sixth and seventh studies.

In the sections below, we start each new study on a new page for readability of the document.

4.1 UI Usability #1: 2D User Interfaces Evaluation

To investigate the usability of the user interface we conducted two studies. The first described here, focuses on traditional 2D interfaces through which a user interacts with the coaches. The second explores a speech-based interface where the user and coaches interact through spoken dialogue (see Section 0).

In this study we evaluate the 2D user interface of the Council of Coaches system. We compare three different user interfaces in a within-subject study design. Five adults of 50 years and older participated in this study. The study is conducted in English. Participants were asked to imagine a scenario where they wanted to be more physical active. Their task was to discuss a new physical activity goal with a council of coaches during the interaction. Through a questionnaire and semi structured interview we gathered insights into how users worked with the three interfaces. Results showed an overall preference for the "WhatsApp-style"-user interface.

Table 2: Summary table for the study: "UI Usability #1: 2D User Interface Evaluation".

Study	Method	Setting	*N*	Participants <50	Participants >50	Participants with health conditions (DM-II, CP)
UI Usability #1: 2D User Interfaces Evaluation	User confrontation, Qualitative, semi-structured interviews and questionnaire	Face to face	5	0	5	0

4.1.1 Objectives

The objective of this study is to evaluate the developed user interfaces of the Council of Coaches system. We want to analyse the usability and user experience of the different user interfaces.

4.1.2 Participants

Five participants (see Table 3) participated in this face-to-face study. Inclusion criteria to participate this study were 1) participants should be 50 years or older, 2) be able to understand the English language and 3) already part of the household or exciting contact of the household of the researcher.

Table 3: Overview of participants in the UI Usability Study #1.

Participant	Gender	Age Category	Experiences	
			Computers	Virtual Coaches
#1	M	50-55	High	Low
#2	F	50-55	Average	None
#3	M	61-65	Moderate	None
#4	F	56-60	Average	None
#5	F	56-60	Low	Low

4.1.3 System

For this study three new user interfaces were developed for the latest version of the technical prototype. One new Unity scene (Figure 5) and two designs of a chat environment (Figure 4 and Figure 7). For this study the excising goal setting dialogue game was adapted and extended. The coaches started with introducing themselves. After that, general topics such as the weather and how the user is doing at the moment are discussed. After that the user and the 2 coaches will try to agree on a physical activity goal for the user.

The user interfaces are based on the ideas presented in D6.2. The Unity scene (Figure 5) is inspired on the user interfaces of Skype and Microsoft Teams. In this scene, each coach has his or her own window and the upper body of the coach is visible. A webcam is used to represent the user in the scene.

The user interface presented in Figure 4 is inspired by a traditional chatroom environment. The name and messages of the coaches are presented in a text area and the users can select possible moves via the blue buttons and send the message to the chat.

The user interface presented in Figure 7 is inspired by the WhatsApp application. Coaches and the user are chatting with each other as they are member of a WhatsApp group. Users can select possible moves via a menu and send the message to the group.

For this study 3 different combinations of user interface and devices were prepared.

- Condition 1: The technical user interface on a laptop (in this case a Lenovo Y540-15IRH 15" laptop) (see Figure 3). Coaches were represented as 3D agents with verbal and nonverbal behaviour. User input via buttons on the laptop screen
- Condition 2: The Skype inspired Unity scene on a laptop in combination with the traditional chatroom on a tablet (Figure 6). Coaches were represented as 3D agents with verbal and nonverbal behaviour. User input via chatroom on a tablet.
- Condition 3: The WhatsApp inspired user interface on smartphone (in this case a Samsung S20) (Figure 7). Coaches were text-based agents on the smartphone screen. User input via the same user interface on a smartphone.

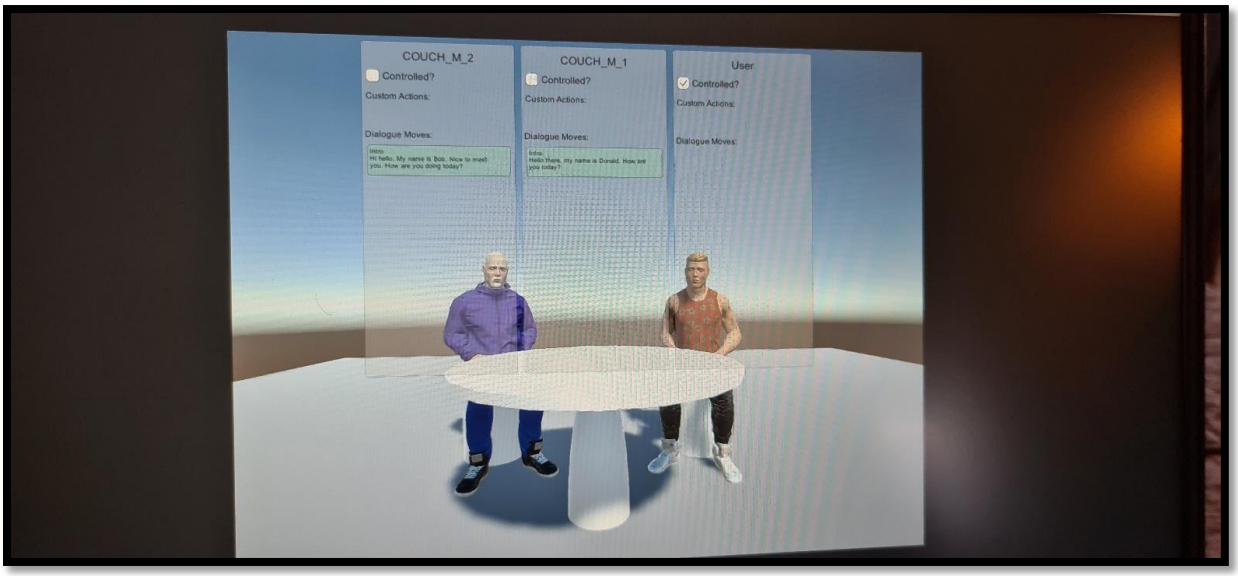

Figure 3: The user interface of the technical demonstrator. In green the buttons to select possible move by the users.

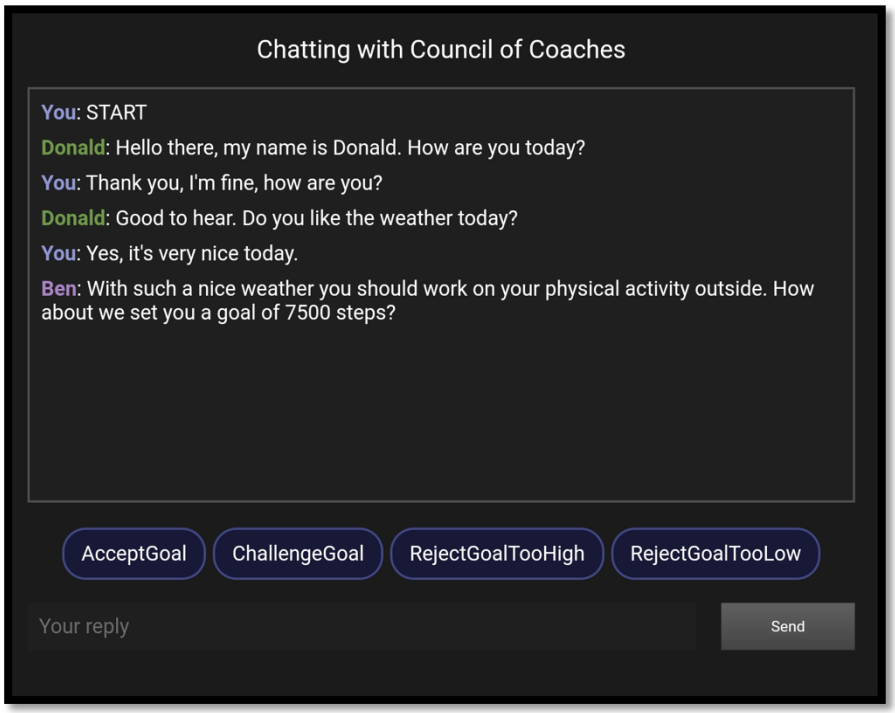

Figure 4: Traditional chatroom user interface on a table. In the text area the history of the chat. The blue buttons are the possible moves of the user that can be selected and send to the chat.

Figure 5: Unity scene inspired by online video conferencing tools like Skype and Microsoft Teams. Possible moves are presented in the chatroom on a tablet.

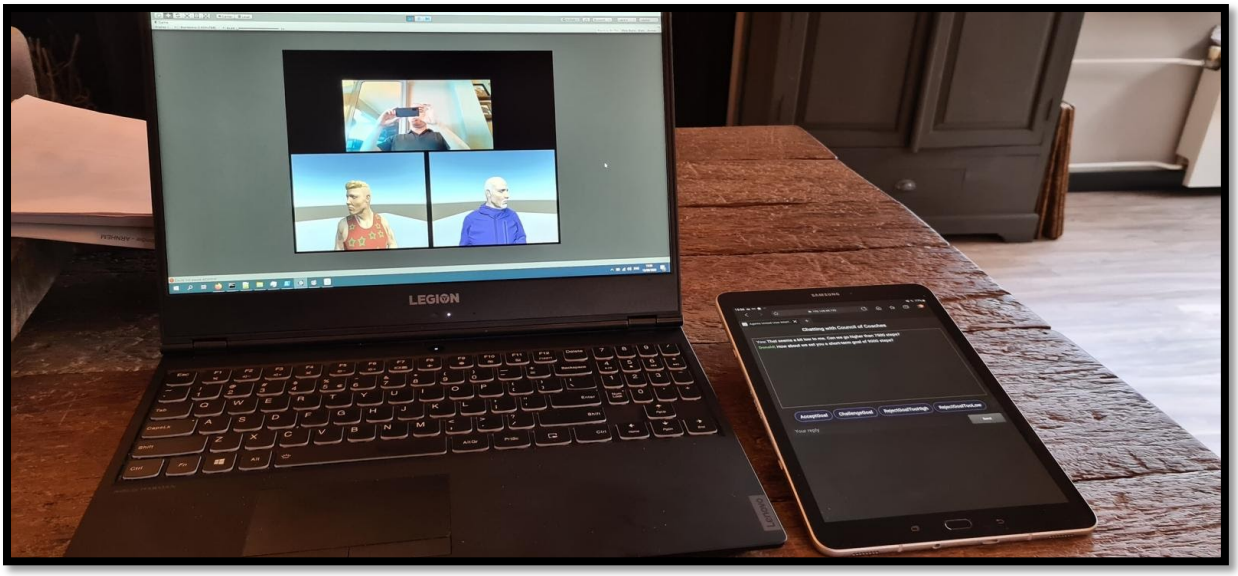

Figure 6: Combination 2 of the user evaluation. The new Unity scene and chatroom on a tablet.

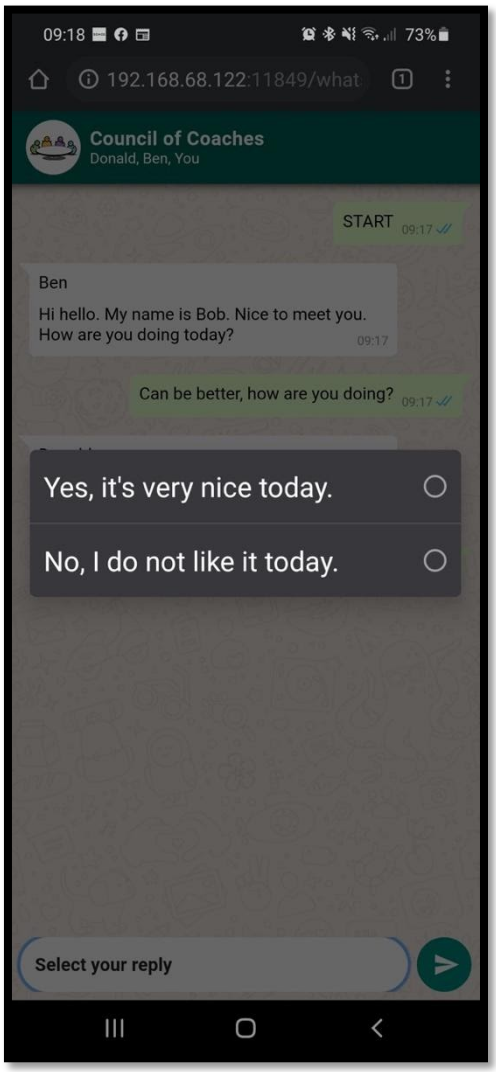

Figure 7: WhatsApp inspired user interface on a smartphone.

4.1.4 Questionnaire

The System Usability Scale (Brooke, 1986) was used to measure the usability of the different (combinations) of user interfaces. A small semi-structured interview was designed to investigate the user experience. Topics that are discussed during the semi-structured interview are the general impression of the interaction, the interaction with the user interface and the interaction with the coaches.

Results of the System Usability Scale are analysed as proposed by Brooke (Brooke, 1986). Results of the semi-structured interview are analysed by a thematic analysis.

4.1.5 Design

A within subject study with 5 older adults was designed. Participants were asked to image them in the situation to improve their physical activity. To do so they needed to set a physical activity goal (number of steps per day) in a conversation with the coaches in the system.

Participants interacted with the user interfaces in the following order, condition 1, condition 2 and condition 3. In each condition the same dialogue is used. Based on the input of the user the dialogue can slightly differ.

4.1.6 Procedure

This study is approved by the ethical committee of the faculty of EEMCS of the University of Twente under the reference number RP 2020-122.
Participants first received an information brochure that explained the goal of the study, the data that will collected, the kind of tasks they need to perform and the possible risks that are involved by participating. After reading this, they could ask questions or discuss any point that was not clear from the information brochure. When the participant accepted to participate in the study, both participant and researcher signed an informed consent.

The researcher started by introducing the scenario. For condition 1 and condition 2 the context in the scenario was that the participant was at home. For condition 3 the context was on the go. At the start of each condition the user interface and possible ways to interact with the user interfaces was explained verbally. After each condition the participant was asked to fill in the System Usability Scale questionnaire. Additionally, aspects about the user experience were asked in the short semi structured interview. The semi-structured interview was audio recorded for analysis. Recordings are deleted after the analysis.

4.1.7 Results

The results of the System Usability Scale and the semi structured interviews are presented in this section.

System Usability Scale

Table 4: System Usability Scale (SUS) scores for the UI Usability #1 study.

Participant	Score		
	Condition 1	Condition 2	Condition 3
#1	60	65	82,5
#2	90	95	95
#3	90	57,5	100
#4	72,5	87,5	67,5
#5	87,5	72,5	87,5
Mean (SD)	**80 (13,3)**	**75,5 (15,6)**	**86,5 (12,6)**

Semi-structured interview

Condition 1:

Three out of five participants (#1, #2, #3) stated that the system was reacting slowly. All participants thought that the user interface was clear. The buttons for user input and the verbal and nonverbal output of the coaches were clear. One participant (#1) mentioned that the realness of the coaches, and the setting (at a table) did not match with the slowness of the system and affected the flow of the interaction. The coaches look real, but react slowly. The setting, a table with the coaches, created a realistic setting for a coaching dialogue, but also creates distance between the coaches and the user. Two Participants (#1, #3) mentioned that the appearance of the virtual coaches did not match the people they would meet in daily life. One participant (#3) mentioned that the coaches were dressed too informal to talk about a serious topic like a goalsetting. Three participants (#2, #4, and #5) mentioned that the virtual coaches in the scene were fun to interact with. These participants could imagine that this could help staying motivated to use the system. Two participants (#3, #4) started talking back to the coaches at the start of the interaction.

Condition 2:

Two participants (#1, #2) mentioned that the use of the interface of the laptop and the tablet with the chatroom environment made the interaction more direct and more personal. Participant #1 mentioned that the concept was based on what he was used to in daily life. Showing only the upper body of the coaches made the interaction more personal and less distant. He mentioned that the webcam view did not add anything extra to the interaction. Participant #2 mentioned that the interaction felt more personal, but the coaches were not always looking at the user during the interactions. Two participants (#4, #5) mentioned that using a touchscreen was easier to use during the interaction. Participant (#3) mentioned that using two devices for a coaching dialogue was too much and not user friendly. He thought that it should work on one device.

Condition 3:

Two participants (#4, #5) mentioned that the screen of the smartphone was too small to interact with the virtual coaches. Participant #1 and participant #3 stated that the user interface was the most pleasant one to use and the most user-friendly user interface in the study. Participant #1 mentioned that the user interface is cleaner, and it fits the way he is communicating with people in daily life. There were less elements in the user interface which was less distracting. His experience was as if he was communicating with real persons. Participant #3 was explicitly mentioning that he was not missing the virtual coaches or the speech output. Participant #4 and participant #5 were missing the virtual

coaches. Without the virtual coaches the interaction was less fun. All participants thought the interaction was in this condition was the fastest.

Preferences

Participants were asked to arrange the user interfaces in order of preferences to use. Table 5 shows the overview of the preferences of the participants.

Table 5: Condition preference of participants (from high to low).

Participant	Order of condition preference (high to low)
#1	3, 2, 1
#2	3, 2, 1
#3	3, 2, 1
#4	2, 1, 3
#5	2, 1, 3

4.1.8 Discussion

On average the results of the System Usability Scale are acceptable. All user interfaces scored above 70. A score of >=70 is considered as acceptable (Bangor, 2009). The lowest score (57,5) is in condition 2 and is still considered as OK (Bangor, 2009).

Results of the semi structured interview showed that the participants thought the interaction in condition 1 was slow. All participants thought the interaction in condition 3 was the fastest. All user interfaces were equally fast in all conditions. A possible explanation could be the fact that there was no text to speech present in condition 3, and the user could read the text of the coaches at their own speed. The presence of the virtual coaches could contribute to make the interaction more direct, more personal and more fun to use. On the other hand, most participants prefer the condition where only text was shown, when asked to arrange them by order of preference of use. Two participants mentioned the appearance of the virtual coaches as too informal. The virtual coaches and other assets used in this study are based on the open source version of the Open Agent platform. Only free or open source assets were used, no paid third-party assets were used in this scene.

The results in this small user study are based on a very small set of participants. On one side these participants fit the target group of the Council of Coaches system, on the other hand the participants were relatives of the researcher. A social desirability bias in the results could be present.

4.2 UI Usability #2: Speech-based user interface evaluation

Typically, users interact with our coaches by selecting one of several predefined multiple-choice options. Although this is a robust user-friendly method to capture user input, it may not be ideal for elderly for several reasons, like loss of fine motor function, physical disabilities, or bad eyesight. Spoken dialogues may offer a better user experience, but also come with many complexities. In this study we implemented a speech recognition system as part of the functional demonstrator. Its impact on the user experience will be examined with target users and proxy users after the completion of the COUCH project. See (Bosdriesz, 2020) for an in-depth analysis and discussion of the opportunities and challenges for adding speech interfaces in COUCH.

Table 6: Summary table for the study: "UI Usability #2: Speech-based user interface evaluation".

Study	Method	Setting	*N*	Participants <50	Participants >50	Participants with health conditions (DM-II, CP)
UI Usability #2: Speech-based user interface evaluation (planned)	Quantitative and qualitative, observations, questionnaires and semi-structured interviews	Face to face and online	-	-	-	-

4.2.1 Objectives

We explored to what extent spoken interaction may offer a valuable addition to the interaction with the Council of Coaches system.

4.2.2 Approach

From related work and literature on spoken dialogue systems we gathered insights into the opportunities and pitfalls inherent to this technology, especially when targeted towards older adults. Taking these lessons into account, we constructed an extension to the COUCH platform to support spoken interaction.

Using this system, we plan to conduct two user studies to investigate how the interaction is influenced:

Study 1: During this study we test the robustness and usability of the application in a real-life setting with our target user group. Older adults who previously participated in our Functional Demonstrator user studies are invited to also use the new speech-based application in their own home using their own tablet/pc and microphone. They will interact with the system for multiple sessions. We will use diaries and/or semi-structured interviews to gather participants' experiences with the system. We aim for ~5 participants of age 50+.

Study 2: In this study we make a closer comparison between the original functional demonstrator application and the new speech-based application. This will be a single-session interaction with older adults and proxy-users. We will use a within-subjects method. The interaction will take place in a more controlled setting, ensuring minimal background noise, using a high-quality microphone and a laptop or desktop computer. We will use questionnaires and semi-structured interviews to compare the interaction with the system in terms of the usability and user experience between the two applications. We aim for ~25 participants.

4.2.3 System

As a basis for this study we used the Functional Demonstrator platform, which has the advantage of being lightweight, stable, and easy to distribute to end-users. Additionally, there is a large collection of relevant WOOL-based dialogue content available, allowing us to explore speech-based interactions over a longer period of time and throughout multiple sessions.

Our speech recognition system uses a state-of-the-art Kaldi-based engine (Povey, et al., 2011) with Dutch language models. To integrate the speech recognition in the HTML and JavaScript-based web interface of the Functional Demonstrator we used the dictate.js library.

To enable spoken interaction with the system we use the following approach. At each appropriate moment in the dialogue, recognised user speech is matched to the available user-reply options. Each reply has certain automatically-derived and manually-authored keywords associated with it. If a keyword is found in the user's speech, the corresponding reply is selected and the dialogue continues. If no keyword is found the user is asked to repeat themselves.

Additionally, to support a natural two-way spoken dialogue with the coaches we implemented text-to-speech synthesis using Dutch WaveNet voices offered by the Google Cloud API. A coach's lines in the dialogue are synthesised and played through the computer or tablet's speakers.

The user interface has been extended to show the state of the speech recognition and the text-to-speech: icons were added to show when a coach was speaking and when the system was listening (see Figure 8).

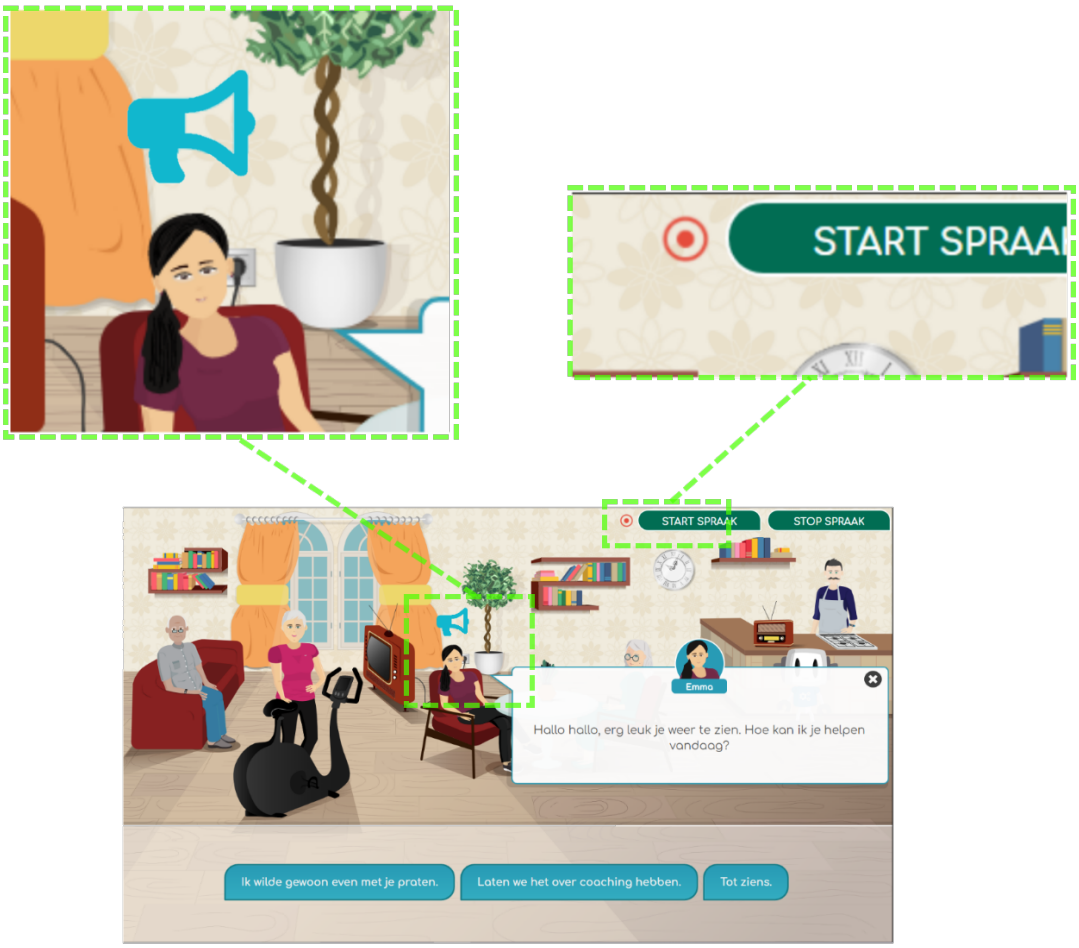

Figure 8: A speaker icon is displayed above the coach when they are speaking and a record icon is displayed when the system is listening for user speech.

4.2.4 Results and discussion

Insights from related work and literature teach us that speech recognition is a complex process, wherein many problems can arise. Variations in input make it difficult for an ASR to accurately recognize the spoken words. Environmental noise and speech characteristics and pronunciation of elderly worsen this problem. The models used by speech recognizers are trained on and for "average" people's voices. This means they may yield lower performance for groups of people, like elderly, that are not part of these *average* voices group. As people grow older, a natural ageing of the voice takes place. The characteristics of an aged voice have been found to be less easily recognized by standard ASR systems. Furthermore, the conversation mechanisms in spoken language, such as grounding, turn taking and conversational repair, are orders of magnitude more complex than for well-defined multiple-choice text-and-button-based interactions.

Creating a usable spoken dialogue system is clearly not a trivial task. Two user studies are currently in preparation, where we will investigate to what extent the user experience and system usability is influenced when confronting older adults with our system.

4.3 Multi-device interaction evaluation

As part of this evaluation we have focused on bringing the Council of Coaches to Virtual Reality: in this way users interact with their coaches face-to-face in an immersive 3D setting. In several exploratory studies we confronted users with various 3D environments, interaction methods and accessibility features. We used a combination of screen-based interactions, virtual reality, and interactive videos to test the various aspects of our prototypes. User experiences were measured through questionnaires and semi-structured interviews. Participants were 6 users of 50+ and 9 proxy users. Results showed that users preferred interacting with the coaches in an outdoor nature setting, and appreciated the inclusion of subtitles to increase the accessibility of the system. The process of designing and evaluating our virtual reality prototypes is discussed in more detail in (van der Werff, 2020) and (Petersen, 2020).

Table 7: Summary table for the study: "Multi-device interaction evaluation".

Study	Method	Setting	*N*	Participants <50	Participants >50	Participants with health conditions (DM-II, CP)
Multi-device interaction evaluation	Qualitative, questionnaires and semi-structured interviews	Face to face and online	15	9	6	-

4.3.1 Objectives

We investigate how factors in the 3D environment and VR-specific interaction modalities influence the user experience and engagement with the Council of Coaches system.

4.3.2 Participants

Participants in this study were all from the same households as the researchers, to avoid the risk of introducing new social contacts. None of the participants had prior experience with VR. We recruited a total of 6 users aged 50+ and 9 proxy users. Proxy users were instructed to imagine themselves in the position of an older adult and were given a short scenario to help them empathise with such a target user.

4.3.3 System design and implementation

Through the process of ideation, we explored the design space relevant to the VR user experience. From literature and related work, we gathered input for brainstorming sessions on relevant 3D environments and interaction modalities.

3D environment – The environment and setting in which the interaction with the coaches takes place may have an influence on how users engage with the coaches. In the functional demonstrator the coaches are situated in a living room setting, whereas in previous versions of the technical demonstrator the interaction took place in an otherwise empty space. We focused on designing and developing an appropriate 3D VR environment for the technical demonstrator. Guided by user persona's, we explored environmental aspects related to being indoors/outdoors, cosy/formal, private/public, cluttered/tidy, industrial/natural and real/sci-fi. Through exploratory interviews with two users (aged 50+) we narrowed down to two concepts:

- An indoor beach house environment, with a calm private appearance (an impression is given in Figure 9).

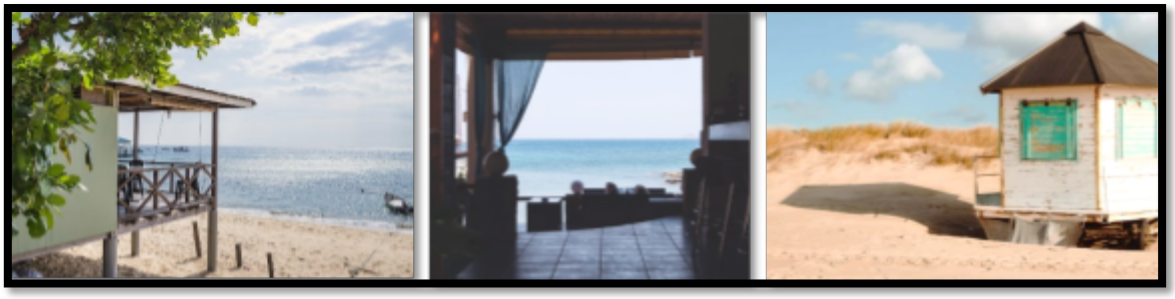

Figure 9: Mood design images for the indoor beach house environment.

- An outdoor forest environment, with a wide-open view of nature (an impression is given in Figure 10)

Figure 10: Mood design images for the outdoor forest environment.

Interaction and accessibility – The interaction modalities offered by VR are different than those used in regular 2D screen-based interactions. Additionally, VR can offer certain accessibility features to further enhance the user experience for older adults. The interaction design was based on insights from related work and literature, as well as design guidelines from leading VR manufacturers. Possible options and situations were sketched, as shown in Figure 11. A set of five interaction and accessibility features were conceptualised in an office environment:

- Location of the input options: on a virtual table or floating in the air.
- Timing of the visibility of the input options: always visible or only when prompted.
- Mode of selecting an input option: grabbing an option or pressing a virtual button.
- Appearance of the user's hands in VR: appear as controllers or appear as virtual hands.
- Availability of subtitles or interaction history as an accessibility feature: a coach's verbalisations are shown as subtitles in a floating text balloon, or are displayed in an overview of the interaction history.

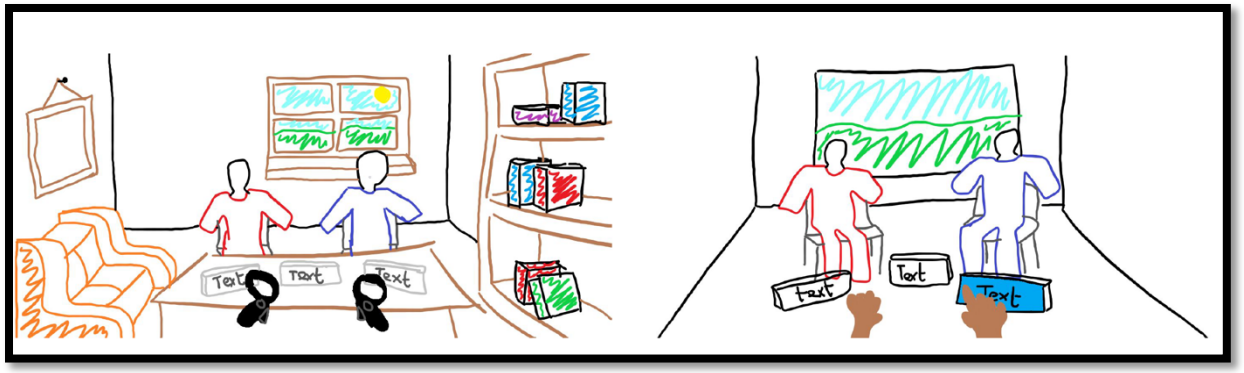

Figure 11: Sketches of various user interaction features.

Figure 12: Implemented prototypes of the various 3D environments and interaction modalities.

The various concepts following from the ideation process were translated into working prototypes for the Council of Coaches technical demonstrator. The Unity scene was extended to support Oculus VR headsets by importing the Oculus Integration Package. Assets were created in Autodesk Maya where necessary, or were imported from third party assets. Figure 12 shows an overview of the various developed VR scenes. Figure 13 shows the interaction history displayed in a picture frame to the side of the coaches. Figure 15 shows an example of the coach's verbalisations appearing in the scene as subtitles in a text balloon, and it shows the user's hands as they are pressing a floating button with an extended index finger. Conversely, Figure 14 shows the user's controllers as they select an input option on the table by grabbing it.

Figure 13: The interaction history is displayed on a picture frame on the wall to the side of the coaches.

Figure 14: Showing the user's controllers as they select an input option on the table by grabbing it.

Figure 15: Subtitles appearing as a text balloons, and a visualisation of the user's hand in VR while pressing one of the floating buttons with an extended index finger.

4.3.4 Methods

Various exploratory studies were conducted with the available prototypes. We used a combination of real-time in-person testing in VR and on screen, and pre-recorded interactive videos. The studies used a within-subject group design, allowing participants to compare environments and features across prototypes.

We used a shortened version of the User Engagement Scale (UES) (O'Brien, Cairns, & Hall, 2018) and semi-structured interviews to measure the engagement and user experience while evaluating the various prototypes.

4.3.5 Results and discussion

Regarding the environments, participants mentioned that they considered the looks, ambience and view as important aspects for increasing engagement. They showed a preference for having a dialogue with the coaches in the forest environment in VR, as it was seen as more realistic, giving a better experience and more immersion. The participants preferred the outdoor experience offered by this environment because of the unlimited view of the surrounding nature. Regarding the indoor beach house and office environment, some participants mentioned they liked the inclusion of furniture, as it added to a feeling of warmth and cosiness and was a recognisable setting for a coaching conversation. However, others mentioned that having a table between themselves and the coaches introduced a certain social distance. Interestingly, a similar remark was made during the 2D interface evaluation study, previously described in Section 4.1.1.

Regarding interaction, users indicated that input options should appear only when the user is being prompted, and should disappear when a selection is made. Although there was no clear preference for the location of input options, users seemed to prefer selecting an option as if they were pressing a real button with their finger; the visual feedback and greater sense of realism resulted in a better user experience. The representation of the user's virtual hands instead of controllers further increased the sense of realism. Regarding accessibility, the inclusion of text balloons in the scene was welcomed by all users, as it increased the understandability and ease of use. Additionally, being able to re-read the interaction history might help users remember what had been previously discussed during the session.

Although we involved several participants from our target user group at various stages in these studies, only two of them experienced the system in virtual reality. The others experienced the interaction with the coaches through interactive videos, showing a pre-recorded interaction that took place in VR. Although the videos displayed the interaction mechanisms, participants watching the videos were unable to experience the interaction modalities first hand. Our proxy users are likely more familiar with new technologies, and may have picked up on VR more easily than older adults would have.

4.4 Verbal conflict presentation style impact on group discussion evaluation

A part of the feedback from participants in the study described in Deliverable 6.5, section 7.2 was on the rather competitive and aggressive way the coaches spoke to each other when they discussed their opinions. Participants remarked this could be improved. This could make the coaches more realistic, likeable, and effective at coaching. In this experiment, we compared several different conflict presentation styles in group discussions, and looked at how they came across to participants.

Table 8: Summary table for the study: "Verbal conflict presentation style impact on group discussion evaluation".

Study	Method	Setting	*N*	Participants <50	Participants >50	Participants with health conditions (DM-II, CP)
Verbal conflict presentation style impact on group discussion evaluation	Online interaction followed by questionnaires	Online study	242	0	242	?

4.4.1 Objectives

We tried to investigate how different kinds of presentation styles of conflict impacted the perception participants formed of a virtual coaching team.

4.4.2 Participants

A total of 242 (81 men, 161 women, 0 other, 0 did not want to say) participants between 50 - 87 years of age (M = 57.50, SD = 6.66) were recruited through the online platform of Prolific. This was done to avoid contact in person, as well as to reach a larger group of participants. They were compensated for their participation with £5.

4.4.3 System

The coaches and environment that we designed for the system were made to be very simple 2D coaches sitting at a table. The interactions with the coaches were also kept simple. The coaches communicated using text balloons, and the participants could respond by selecting their response on a button. This was done for three reasons. The first reason was to reduce the impact of the appearance of individual coaches on how their message was perceived. The second reason was to make the system easy to interact with for participants, as they would interact with it online based on instructions and without the ability to get clarification from the researcher. The third and final reason was to make the system be able to run for any participant using a desktop with ease. The original prototype can be seen in Figure 16.

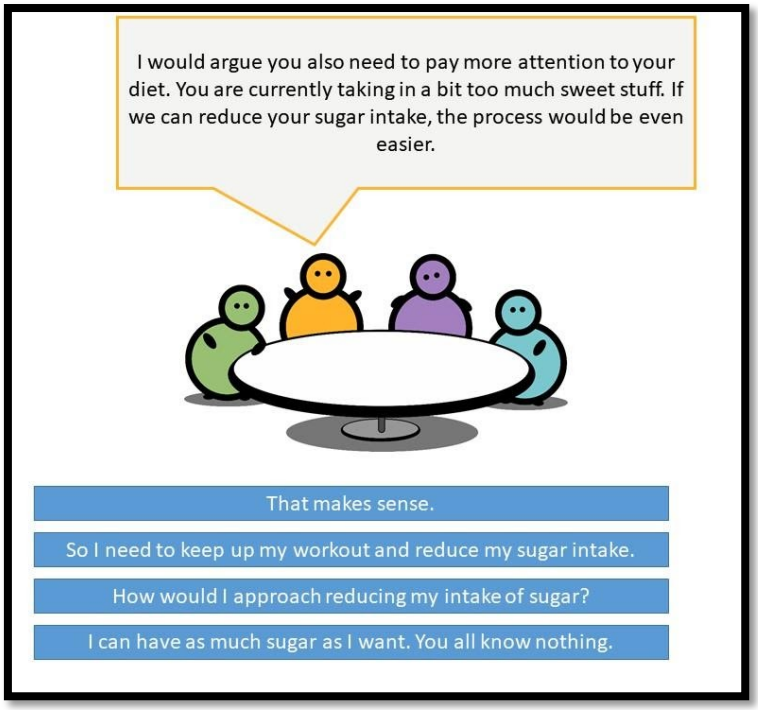

Figure 16: One of the interactions in the prototype system.

We based the two conflict presentation styles we wanted to compare on two theoretical models about interpersonal communication from the field of psychology. For each model, we modelled one of the four coaches after one of the four subsections of the possible behaviours in that model. The same was done by giving participants the opportunity to give their own input at a few points during the interactions, by giving them four options and modelling each option after one of the four subsections of the possible behaviours in that model.

The first was the Interaction Process Analysis (IPA) model (Bales, 1951), and the second was the Interpersonal Circumplex (ICP) model (Leary, 1957). Table 9 below should summarize the models and how they were implemented in a few key words to give an understanding of the types of behaviour we modelled. Both models can be further studied in the referenced sources.

Table 9: Summary of the two interpersonal interaction models.

Interaction Process Analysis (IPA)	**Interpersonal Circumplex (ICP)**
Social-emotional: positive Shows solidarity, shows tension release, or agrees	Dominant-hostile Narcissistic, competitive, sadistic, or aggressive
Task-related: giving of information Gives suggestion, gives opinion, or gives orientation	Submissive-hostile Rebellious, distrustful, self-effacing, masochistic
Task-related: asking for information Asks for suggestion, asks for opinion, or asks for orientation	Submissive-friendly Docile, dependent, cooperative, over-conventional
Social-emotional: negative Shows antagonism, shows tension, or disagrees	Dominant-friendly Hypernormal, responsible, autocratic, managerial

Six dialogues were written based on these models, with three dialogues per model. For these three dialogues per model, there were three general health goals to discuss with the team (losing weight, stress management, and sleeping better). These topics were the same for each model, and contained the same advice for the participant given by the coaching team. Each of the four coaches in each dialogue represented one of the four categories from the model for that dialogue (see Table 9) throughout all that they said. When the participant could give their own input, this also consisted of four options that were each one of the categories of the model for that dialogue. . Each dialogue was a group discussion in which the team was trying to help the participant achieve one of the aforementioned general health goals. In each discussion, there was conflict to some extent. In some cases this was inter-coach conflict, and in some cases the coaches clashed with the participant. There were several of such conflicts in each interaction, though it depended on the extent to which the participant answered in a negative and hostile manner how many of these conflicts they encountered. At least one conflict was always presented, which was always a conflict between the coaches about the advice they were giving as a team.

To make sure the dialogues we wrote based on the models were of good quality and fit the models, a pilot study was conducted before the main experiment. In this pilot study we had four participants of fifty (50) years or older evaluate our dialogues. They interacted with 2D coaches in Microsoft PowerPoint while screen sharing with us and thinking out loud. Clicking buttons on the slides automatically navigated them to the slide with the reply by the coaches. These interactions were recorded after they went through these interactions, we had them evaluate our dialogues by filling out questionnaires and taking part in a brief interview that was also recorded. Based on the answers given by participants, the dialogues were evaluated with regards to their fit to the models they were based on, and were further modified in preparation of the experiment based on their feedback. For an example of the interactions with the final system, see Figure 17.

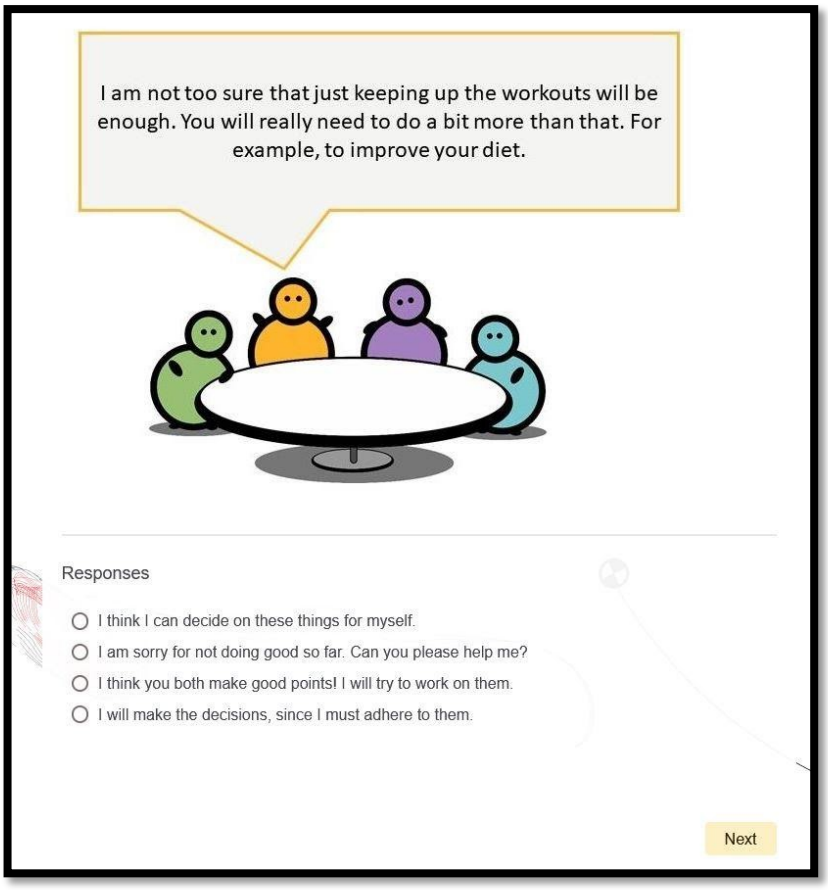

Figure 17: One of the interactions with the four coaches in the final system.

4.4.4 Questionnaires

We used several questionnaires measuring different constructs. For all the questionnaires we used, we rephrased the statements and questions to be about the coaching team as a whole, instead of about the individual coaches. For each questionnaire we presented the questions in a random order.

The first set of questionnaires we used was the Godspeed Questionnaire Series (Bartneck C. , 2020). This series consists of questionnaires measuring five constructs about how agents come across to participants. These are anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety (Bartneck, Croft, & Kulic, 2009).

We measured the construct of perceived quality of the coaching given by the coaching team. To do so, we used an adjusted version of the Coaching Behavior Scale for Sport (Côte, Yardley, Hay, Sedgwick, & Baker, 1999), or CBS-S for short. The original CBS-S we found contained several items that were not relevant to the interactions in our experiment. For that reason, we did not use items on the scales of physical training and fitness, technical skills, competition strategies, personal rapport, and negative personal rapport (i.e. items 1 to 15, and items 27 to 47). On the scale of mental preparation, we did not use the item regarding performance under pressure (i.e. item 16), as the dialogues with the coaching team were about general health goals, and did not address performing under pressure. On the scale of goal settings, we did not use the items regarding monitoring of progress, identifying target dates for attaining goals, and setting long-term goals (i.e. items 22, 24, and 25), as the dialogues with the coaching team were not about progress, planning, or setting long-term goals. They were more focused more on how to behave in the short term, and on giving more general tips. Furthermore, several items relevant to the interactions in our experiment were added under a new ``coaching quality'' scale. These were the following items:

1. My coaching team helps me to be motivated and inspired by others.
2. My coaching team helps me to discover which things help me to attain and maintain my (fill in health goal) better.
3. My coaching team had the right knowledge and abilities to give good coaching.
4. My coaching team gives advice of good quality.

Another construct measured was group cohesion. This was done using a selection of items from an earlier study on estimating group cohesion in small groups (Hung & Gatica-Perez, 2010). As these items were developed for analysing audio-visual nonverbal behaviour, and our material was mostly varying in what was “said” through the use of text, we had to pick items that fit our setup. Furthermore, we rephrased all of them to statements that our participants could agree or disagree with. We decided on six items from the task cohesion category, and six items from the social cohesion category. These were the following items:

Task cohesion

1. The team members share the same purpose/goal/intentions.
2. The team seems to share the responsibility for the task.
3. Overall, the team members appear to be collaborative.
4. Every team member seems to have sufficient time to make their contribution.
5. Overall, the team members give each other a lot of feedback.
6. The morale of the team is good.

Social cohesion

7. Overall, the team members appear to be in tune/in sync with each other.
8. Overall, the team members seem involved/engaged in the discussion.
9. Overall, there appears to be equal participation from the team members.
10. The team members appear to be receptive to each other.
11. Overall, the team members appear to be supportive towards each other.
12. Overall, it feels like the team operates spontaneously.

To measure the construct of persuasiveness of the team, we posed them with two items of our own. The first item they could indicate their agreement to. The second item asked them to explain why they

answered the previous item the way they did. Answering that item was optional. These were the following items:

1. I would try out the behaviour the coaching team recommended, if I had (fill in health goal) as a health goal.
2. Why, or why not?

The final construct we measured was the interaction experience. We developed six items for this construct. Each of the items the participants could indicate their agreement with. These were the following items:

1. I am satisfied with the coaching conversation I had.
2. I am satisfied with the amount of freedom I had in my replies in the coaching conversation.
3. I am satisfied with the way the coaches spoke to me in the coaching conversation.
4. I would recommend having a conversation with the coaching team to my friends, family, and colleagues that could benefit from a conversation about this topic.
5. I would have a conversation with the coaching team about other topics.
6. I would have another conversation with the coaching team about this topic.

4.4.5 Design

We conducted a user study consisting of three smaller within-subject studies. In each of these studies, participants were presented in a random order with one dialogue with the 2D coaching team based on the IPA model and one based on the IPC model. These dialogues would be about the same general health topic, and contain the same general tips. The difference was in the way the coaches presented it. Each dialogue contained at several conflicts, of which at least one, which was between the coaches, was always presented. The others were presented if the participant themselves were confrontational, and were conflicts between the coaches and the participant.

4.4.6 Procedure

The entire user study consisting of three smaller studies was made available online, and was completed without interaction with the researchers. At the start of the study participants were presented with an information brochure that explained the goal of the study, the data that will collected, the kind of tasks they needed to perform and the possible risks that were involved by participating. If anything was unclear, or they had any issues, they could use the contact information provided to contact the main responsible researcher. The participant could then accept to participate in the study by signing the informed consent form.

The study then introduced the scenario, and showed how to interact with the coaches using an example. For each of the three smaller studies, the participant was then presented in a random order with the two dialogues about the same topic (one based on IPA model, one on IPC model), followed by the questionnaires about that conversation. The questionnaires were the exact same for each dialogue.

4.4.7 Results

A series of paired sample T-tests was conducted on all the scales and questionnaires that were used. This was done for all 242 participants. Even though the study was divided into three smaller studies, in each of them the comparison was made between the IPA model and IPC model. Hence, we merged the data and analyses for all participants. That way we could compare the performance of both models overall. The only exception was the persuasion question, which could only be evaluated for one of the three smaller studies, as an error was made in the question in the other two studies. Thus, for this question only the 82 participants in that study were used in the analysis of the persuasion construct.

The outcomes of our analyses can be found in Table 10 below. For the Godspeed questionnaire series, a positive Cohen's d means a higher rating on the construct for the IPA model, as compared to the IPC model. For all the other constructs, a negative Cohen's d means a higher rating on the construct for the IPA model, as compared to the IPC model.

Table 10: Results of verbal conflict presentation style in group discussion user study.

	N	Mean (IPA; IPC)	Standard Deviation (IPA; IPC)	Correlation between scores	Difference between scores	Cohen's d
Godspeed: Anthropomorphism	242	M = 3.33 M = 3.05	SD = .96 SD = .99	r = .577, p < 0.001	.28 (95% CI [.17, .39]), p < 0.001	.29
Godspeed: Animacy	242	M = 3.50 M = 3.25	SD = .92 SD = .96	r = .601, p < 0.001	.25 (95% CI [.14, .35]), p < 0.001	.26
Godspeed: Likeability	242	M = 3.64 M = 2.89	SD = .87 SD = 1.10	r = .372, p < 0.001	.75 (95% CI [.61, .89]), p < 0.001	.76
Godspeed: Perceived Intelligence	242	M = 3.69 M = 3.20	SD = .84 SD = 1.01	r = .510, p < 0.001	.49 (95% CI [.38, .61]), p < 0.001	.53
Godspeed: Perceived safety	242	M = 3.79 M = 3.31	SD = .90 SD = 1.08	r = .556, p < 0.001	.48 (95% CI [.37, .60]), p < 0.001	.49
Adjusted CBS-S: Mental preparation	242	M = 2.95 M = 3.25	SD = .94 SD = 1.04	r = .594, p < 0.001	-.29 (95% CI [-.41, -.18]), p < 0.001	-.30
Adjusted CBS-S: Goal setting	242	M = 2.39 M = 2.84	SD = .92 SD = 1.04	r = .553, p < 0.001	-.45 (95% CI [-.57, -.33]), p < 0.001	-.46
Adjusted CBS-S: Quality of coaching	242	M = 2.61 M = 3.06	SD = .93 SD = 1.06	r = .504, p < 0.001	-.45 (95% CI [-.57, -.32]), p < 0.001	-.45
Group cohesion: Task cohesion	242	M = 3.16 M = 3.53	SD = 1.20 SD = 1.22	r = .446, p <0.001	-.38 (95% CI [-.54, -.21]), p < 0.001	-.31
Group cohesion: Social cohesion	242	M = 3.30 M = 3.70	SD = 1.19 SD = 1.26	r = .395, p < 0.001	-.40 (95% CI [-.57, -.23]), p < 0.001	-.33
Persuasion	82	M = 2.52 M = 3.07	SD = 1.41 SD = 1.68	r = .582, p < 0.001	-.55 (95% CI [-.86, -.23]), p = 0.001	-.35
Interaction experience	242	M = 3.56 M = 4.43	SD = 1.61 SD = 1.73	r = .571, p < 0.001	-.87 (95% CI [-1.06, -	-.52

					.67]), p < 0.001	

4.4.8 Discussion

As can be seen in Table 10, all the scales measuring our constructs showed significant results. We will briefly discuss what all of them mean, and discuss the potential implications. We will interpret effect sizes of d = 0.2 as small, d = 0.5 as medium, and d = 0.8 as large, following the guidelines by Cohen (Cohen, 1988). As was mentioned in the results section, for the Godspeed questionnaire series, a positive Cohen's d means a higher rating on the construct for the IPA model, as compared to the IPC model. For all the other constructs, a negative Cohen's d means a higher rating on the construct for the IPA model, as compared to the IPC model.

First off, in the Godspeed questionnaire series we found that the coaching team using the IPA model came off as more anthropomorphic (d = .29), animate (d = .26), likeable (d = .76), intelligent (d = .53), and people felt safer too (d = .49) when compared to the coaching team using the IPC model. These results suggest that with regards to how the coaching team came off, the IPA model was preferred in all respects to the IPC model. Especially the likeability, intelligence, and perceived safety constructs stand out for their medium-large sized effects. All in all, there is a clear preference for the coaching team using the IPA model.

Next, taking a look at the adjusted CBS-S, we found that the coaching team using the IPA model came off as better able to help with mental preparation (d = -.30), as well as with goal setting (d = -.46), and was evaluated as giving a better quality of coaching (d = -.45) when compared to the coaching team using the IPC model. These results suggest that with regards to how the coaching team came off, the IPA model was preferred in all respects to the IPC model. Especially the goal setting, and quality of coaching constructs stand out for their near medium sized effects. All in all, there is a clear preference for the coaching team using the IPA model.

For the group cohesion questionnaire, we found that the coaching team using the IPA model came off as more cohesive both in task cohesion (d = -.31), as well as social cohesion (d = -.33) when compared to the coaching team using the IPC model. These results suggest that with regards to how the coaching team came off, the IPA model was preferred in all respects to the IPC model. The effects were small-medium, but meaningfully show a preference for the coaching team using the IPA model regardless.

For the persuasion question we found that the coaching team using the IPA model was more persuasive (d = -.35) when compared to the coaching team using the IPC model. This result suggest that the coaching team using the IPA model was better able to persuade participants than the coaching team using the IPC model. The effect was small-medium, but meaningfully shows a preference for the coaching team using the IPA model regardless.

Finally, in our interaction experience questionnaire we found that the coaching team using the IPA model gave the participants a more positive interaction experience (d = -.52) when compared to the coaching team using the IPC model. This result suggests that the coaching team using the IPA model had an interaction with the participants that was perceived as more pleasant than the coaching team using the IPC model. The effect was of medium size, showing a decent amount of preference for the coaching team using the IPA model.

All the results considered, it is clear that participants preferred their dialogue with the coaching team using the IPA model. This may have to do with the fact that the coaching team using the IPC model had two hostile coaches, whereas the coaching team using the IPA model only had one negative coach. This may have given the coaching team using the IPC model a more negative tone. Furthermore, the dialogues with the coaching team using the IPC model usually opened with these more negative coaches, which may have primed people negatively. Finally, the coaching team using the IPA model may have had a clearer coach that provides information and asks for clarification, as compared to the coaching team using the IPC model due to the task-related: giving of information, and task-related: asking for information coaching roles in the IPA model.

4.5 Peer mediator coach presence and behaviour impact on group discussion evaluation

In previous work (Dohsaka, Asai, Higashinaka, Minami, & Maeda, 2009), the impact of having a peer agent being present during an interaction was shown to be beneficial. One can imagine potential benefits in a coaching context. Having a peer who empathises with participants could improve the experience the participants have during the group interaction, and might make them more open to share information. The peer coach could also support proposals by coaches and emphasise how they used that advice to enhance their quality of life, potentially increasing the credibility of the coaches and the advice they give. Finally, they could mediate a discussion between coaches. In this experiment, we plan to compare a group coaching interaction including a peer mediator coach mediating conflicts to the previous study that did not include a peer mediator coach that we described in Section 4.3. The dialogues used in this study are the same as in the previous study, but for each conflict a brief interaction with the peer mediator coach was added.

Table 11: Summary table for the study: "Peer mediator coach presence and behaviour impact on group discussion evaluation".

Study	Method	Setting	*N*	Participants <50	Participants >50	Participants with health conditions (DM-II, CP)
Peer mediator coach presence and behaviour impact on group discussion evaluation	Online interaction followed by questionnaires	Online study	243	0	243	?

4.5.1 Objectives

We tried to investigate how different kinds of presentation styles of conflict when mediated by a peer coach impacted the perception participants formed of a virtual coaching team. We specifically wanted to compare this to the previous study where this peer mediator coach was not present.

4.5.2 Participants

A total of 243 (100 men, 143 women, 0 other, 0 did not want to say) participants between 50 - 76 years of age (M = 57.74, SD = 6.15) were recruited through the online platform of Prolific. This was done to avoid contact in person, as well as to reach a larger group of participants. They were compensated for their participation with £3.75.

4.5.3 System

The system was based off of the one in Section 4.3. It only had the addition of the peer mediator coach (red character), and their content (see Figure 18). The presentation style of the system remained the same for the same reasons, and the content from the original system was also used here. The only difference was the addition of the peer mediator coach character at the table, and the conflict mediation content that was added between them and those involved in the conflicts in the dialogues. That way, any difference with the previous study should come from the addition of the peer mediator coach and their mediation of the conflicts.

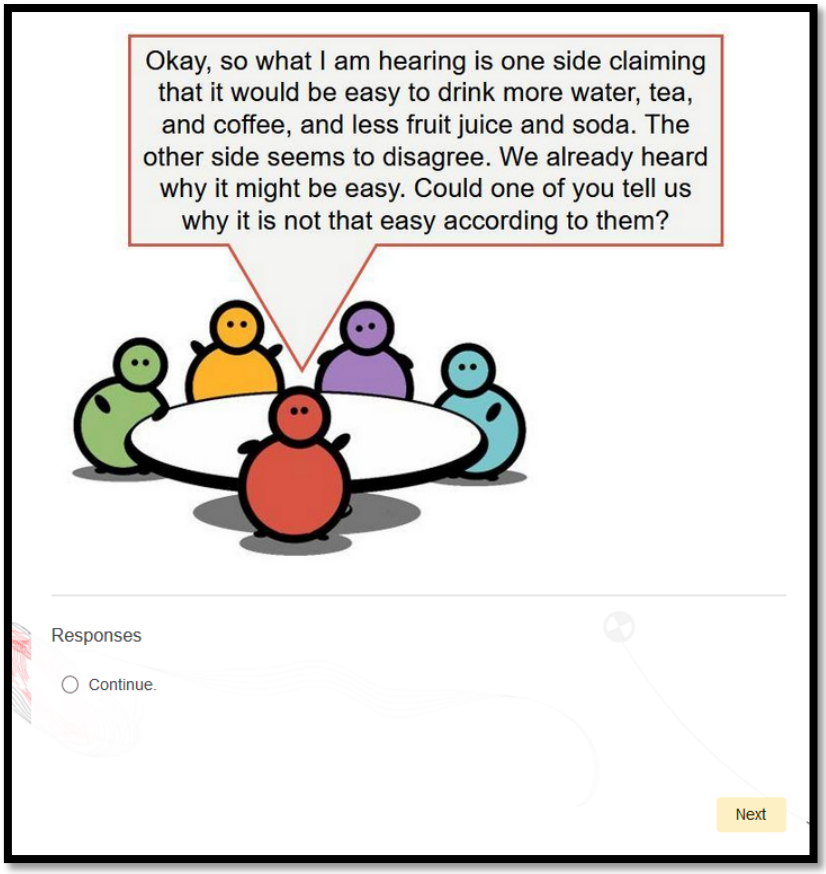

Figure 18: One of the interactions with the five coaches in the final system.

The peer mediation coach their behaviour was based off of an interesting work from the field of conflict resolution (Jacobs & Aakhus, 2003). They presented three models of rationality used in conflict mediation. We summarized one of their tables describing these models of rationality in Table 12 below.

Table 12: Models of rationality as described in (Jacobs & Aakhus, 2003).

	Critical Discussion	**Bargaining**	**Therapeutic discussion**
Source of conflict	Disagreement over facts and public values	Conflict between competing wants and interests	Failures of mutual respect and mutual understanding
Optimal solution	Claim that is most consistent with available facts and values	Proposal that maximizes gain and minimizes costs to both parties	Definition of the situation that acknowledges and affirms each party's point of view
Principle of resolution	Public justifiability	Mutual acceptability	Sincerity of openness
Process of resolution	Argumentation and refutation	Offers and concessions	Self-disclosure, explanations, and definitions
Mode of resolution	Agreement	Contract	Reciprocal affirmation

We had the peer mediator coach behave in a way that matched the description of the behaviour of a mediator in the type of conflict that they were having according to Jacobs and Aakhus. For each conflict in each dialogue we looked at what the source of conflict was, and had the peer mediator coach mediate said conflict until it led to a peaceful resolution. No conflicts ended up having a conflict between competing wants and interests, and thus no bargaining was used. All conflicts were resolved using critical discussion and therapeutic discussion. For an example of the kind of behaviour the peer mediator coach displayed, Figure 18 shows the peer mediator coach starting a critical discussion. We had all the other coaches respond reasonably to this coach, and work along with them to resolve the conflict as described in this work. This broke with their behaviour as modelled in the IPA and IPC model sometimes, but was necessary to have the conflicts be resolved peacefully within a reasonable amount of time.

4.5.4 Questionnaire

We used several questionnaires measuring different constructs. For all the questionnaires we used, we rephrased the statements and questions to be about the coaching team as a whole, instead of about the individual coaches. For each questionnaire we presented the questions in a random order.

The first set of questionnaires we used was the Godspeed Questionnaire Series (Bartneck C. , 2020). This series consists of questionnaires measuring five constructs about how agents come across to participants. These are anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety (Bartneck, Croft, & Kulic, 2009).

We measured the construct of perceived quality of the coaching given by the coaching team. To do so, we used an adjusted version of the Coaching Behavior Scale for Sport (Côte, Yardley, Hay, Sedgwick, & Baker, 1999), or CBS-S for short. The original CBS-S we found contained several items that were not relevant to the interactions in our experiment. For that reason, we did not use items on the scales of physical training and fitness, technical skills, competition strategies, personal rapport, and negative personal rapport (i.e. items 1 to 15, and items 27 to 47). On the scale of mental preparation, we did not use the item regarding performance under pressure (i.e. item 16), as the dialogues with the coaching team were about general health goals, and did not address performing under pressure. On the scale of goal settings, we did not use the items regarding monitoring of progress, identifying target dates for attaining goals, and setting long-term goals (i.e. items 22, 24, and 25), as the dialogues with the coaching team were not about progress, planning, or setting long-term goals. They were more focused more on how to behave in the short term, and on giving more general tips. Furthermore, several items relevant to the interactions in our experiment were added under a new ``coaching quality" scale. These were the following items:

1. My coaching team helps me to be motivated and inspired by others.
2. My coaching team helps me to discover which things help me to attain and maintain my (fill in health goal) better.
3. My coaching team had the right knowledge and abilities to give good coaching.
4. My coaching team gives advice of good quality.

Another construct measured was group cohesion. This was done using a selection of items from an earlier study on estimating group cohesion in small groups (Hung & Gatica-Perez, 2010). As these items were developed for analysing audio-visual nonverbal behaviour, and our material was mostly varying in what was “said” through the use of text, we had to pick items that fit our setup. Furthermore, we rephrased all of them to statements that our participants could agree or disagree with. We decided on six items from the task cohesion category, and six items from the social cohesion category. These were the following items:

Task cohesion

1. The team members share the same purpose/goal/intentions.
2. The team seems to share the responsibility for the task.
3. Overall, the team members appear to be collaborative.
4. Every team member seems to have sufficient time to make their contribution.
5. Overall, the team members give each other a lot of feedback.

6. The morale of the team is good.

Social cohesion

7. Overall, the team members appear to be in tune/in sync with each other.
8. Overall, the team members seem involved/engaged in the discussion.
9. Overall, there appears to be equal participation from the team members.
10. The team members appear to be receptive to each other.
11. Overall, the team members appear to be supportive towards each other.
12. Overall, it feels like the team operates spontaneously.

To measure the construct of persuasiveness of the team, we posed them with two items of our own. The first item they could indicate their agreement to. The second item asked them to explain why they answered the previous item the way they did. Answering that item was optional. These were the following items:

1. I would try out the behaviour the coaching team recommended, if I had (fill in health goal) as a health goal.
2. Why, or why not?

The final construct we measured was the interaction experience. We developed six items for this construct. Each of the items the participants could indicate their agreement with. These were the following items:

1. I am satisfied with the coaching conversation I had.
2. I am satisfied with the amount of freedom I had in my replies in the coaching conversation.
3. I am satisfied with the way the coaches spoke to me in the coaching conversation.
4. I would recommend having a conversation with the coaching team to my friends, family, and colleagues that could benefit from a conversation about this topic.
5. I would have a conversation with the coaching team about other topics.
6. I would have another conversation with the coaching team about this topic.

4.5.5 Design

We conducted a user study consisting of three smaller within-subject studies. For the purpose of this deliverable, we were mostly interested in the comparison with the study described in Section 4.3. Thus, we made a between-subject comparison for the dialogues with the same subject and model between the studies to report here. In each of the three studies, participants were presented in a random order with one dialogue with the 2D coaching team based on the IPA model and one based on the IPC model, both involving a peer mediator coach. These dialogues would be about the same general health topic, and contain the same general tips. The difference was in the way the coaches presented it. Each dialogue contained at several conflicts, of which at least one, which was between the coaches, was always presented. The others were presented if the participant themselves were confrontational, and were conflicts between the coaches and the participant.

4.5.6 Procedure

The entire user study consisting of three smaller studies was made available online, and was completed without interaction with the researchers. At the start of the study participants were presented with an information brochure that explained the goal of the study, the data that will collected, the kind of tasks they needed to perform and the possible risks that were involved by participating. If anything was unclear, or they had any issues, they could use the contact information provided to contact the main responsible researcher. The participant could then accept to participate in the study by signing the informed consent form.

The study then introduced the scenario, and showed how to interact with the coaches using an example. For each of the three smaller studies, the participant was then presented in a random order with the two dialogues about the same topic (one based on IPA model, one on IPC model, both involving a peer mediator coach), followed by the questionnaires about that conversation. The questionnaires were the exact same for each dialogue.

4.5.7 Results

A series of paired sample T-tests was conducted on all the scales and questionnaires that were used. This was done for all 243 participants. Even though the study was divided into three smaller studies, in each of them the comparison was made between the IPA model and IPC model. Hence, we merged the data and analyses for all participants. That way we could compare the performance of both models overall with the addition of a peer mediator coach and conflict mediation content. Furthermore, this allowed us to compare the differences between this study that included a peer mediator coach and some new conflict mediation content, and the previous study that otherwise had the same content, but did not have a peer mediator coach or conflict mediation content.

The outcomes of our within-subject analyses can be found in Table 13. For the Godspeed questionnaire series, a positive Cohen's d means a higher rating on the construct for the IPA model, as compared to the IPC model. For all the other constructs, a negative Cohen's d means a higher rating on the construct for the IPA model, as compared to the IPC model.

Table 13: Results of the peer mediator coach presence and behaviour impact on group discussion user study.

	N	Mean (IPA; IPC)	Standard Deviation (IPA; IPC)	Correlation between scores	Difference between scores (IPA – IPC)	Cohen's d
Godspeed: Anthropomorphism	243	M = 3.32 M = 3.05	SD = 1.01 SD = 1.01	r = .680, p < 0.001	.27 (95% CI [.17, .38]), p < 0.001	.27
Godspeed: Animacy	243	M = 3.55 M = 3.30	SD = .93 SD = .96	r = .646, p < 0.001	.25 (95% CI [.15, .35]), p < 0.001	.26
Godspeed: Likeability	243	M = 3.64 M = 3.00	SD = .90 SD = 1.04	r = .572, p < 0.001	.64 (95% CI [.52, .76]), p < 0.001	.66
Godspeed: Perceived Intelligence	243	M = 3.77 M = 3.28	SD = .85 SD = 1.00	r = .561, p < 0.001	.49 (95% CI [.38, .60]), p < 0.001	.53
Godspeed: Perceived safety	243	M = 3.83 M = 3.41	SD = .93 SD = 1.06	r = .527, p < 0.001	.42 (95% CI [.30, .54]), p < 0.001	.42
Adjusted CBS-S: Mental preparation	243	M = 3.01 M = 3.34	SD = 1.01 SD = 1.03	r = .561, p < 0.001	-.33 (95% CI [-.45, -.21]), p < 0.001	-.32
Adjusted CBS-S: Goal setting	243	M = 2.42 M = 2.71	SD = .93 SD = .98	r = .534, p < 0.001	-.29 (95% CI [-.41, -.17]), p < 0.001	-.30

Adjusted CBS-S: Quality of coaching	243	M = 2.58 M = 3.02	SD = .99 SD = 1.04	r = .509, p < 0.001	-.44 (95% CI [-.56, -.31]), p < 0.001	-.43
Group cohesion: Task cohesion	243	M = 2.84 M = 3.38	SD = 1.10 SD = 1.25	r = .562, p < 0.001	-.54 (95% CI [-.68, -.40]), p < 0.001	-.46
Group cohesion: Social cohesion	243	M = 3.02 M = 3.48	SD = 1.14 SD = 1.21	r = .528, p < 0.001	-.46 (95% CI [-.61, -.32]), p < 0.001	-.39
Persuasion	243	M = 2.63 M = 3.07	SD = 1.42 SD = 1.68	r = .550, p < 0.001	-.45 (95% CI [-.64, -.26]), p < 0.001	-.29
Interaction experience	243	M = 3.51 M = 4.28	SD = 1.61 SD = 1.67	r = .617, p < 0.001	-.78 (95% CI [-.96, -.60]), p < 0.001	-.47

The outcomes of our between-subject analyses can be found in Table 14. For these analyses, we called the study described in Section 4.3 study 1, and the study described in Section 4.4 study 2. We compared the outcomes for both studies for the IPA model, and for both studies for the IPC model. For the Godspeed questionnaire series, a positive Cohen's d means a higher rating on the construct for study 1, as compared to study 2. For all the other constructs, a negative Cohen's d means a higher rating on the construct for study 1, as compared to study 2.

Table 14: Results comparison study 1 (verbal conflict presentation style in group discussion) and study 2 (peer mediator coach presence and behaviour impact on group discussion).

	N (study 1; study 2	Mean (study 1; study 2)	Standard Deviation (study 1; study 2)	Difference between scores (study 1 – study 2)	Cohen's d
IPA					
Godspeed: Anthropomorphism	242 243	M = 3.33 M = 3.32	SD = .96 SD = 1.01	.00 (95% CI [-.17, .18]), p = 0.959	.00
Godspeed: Animacy	242 243	M = 3.50 M = 3.55	SD = .92 SD = .93	-.05 (95% CI [-.21, .12]), p = 0.556	-.05
Godspeed: Likeability	242 243	M = 3.64 M = 3.64	SD = .87 SD = .90	.00 (95% CI [-.16, .16]), p = 0.974	.00
Godspeed: Perceived Intelligence	242 243	M = 3.69 M = 3.77	SD = .84 SD = .85	-.08 (95% CI [-.23, .07]), p = 0.312	-.09
Godspeed: Perceived safety	242 243	M = 3.79 M = 3.83	SD = .90 SD = .93	-.03 (95% CI [-.20, .13]), p = 0.684	-.04

Adjusted CBS-S: Mental preparation	242 243	M = 2.95 M = 3.01	SD = .94 SD = 1.01	-.06 (95% CI [-.24, .11]), p = 0.485	-.06
Adjusted CBS-S: Goal setting	242 243	M = 2.39 M = 2.42	SD = .92 SD = .93	-.03 (95% CI [-.20, .14]), p = 0.722	-.03
Adjusted CBS-S: Quality of coaching	242 243	M = 2.61 M = 2.58	SD = .93 SD = .99	.03 (95% CI [-.14, .20]), p = 0.737	.03
Group cohesion: Task cohesion	242 243	M = 3.16 M = 2.84	SD = 1.20 SD = 1.10	.32 (95% CI [.12, .53]), p = 0.002	.28
Group cohesion: Social cohesion	242 243	M = 3.30 M = 3.02	SD = 1.19 SD = 1.14	.28 (95% CI [.07, .49]), p = 0.008	.24
Persuasion	82 243	M = 2.52 M = 2.63	SD = 1.42 SD = 1.42	-.10 (95% CI [-.46, .26]), p = 0.578	-.07
Interaction experience	242 243	M = 3.56 M = 3.51	SD = 1.61 SD = 1.61	.06 (95% CI [-.23, .34]), p = 0.706	.03
IPC					
Godspeed: Anthropomorphism	242 243	M = 3.05 M = 3.05	SD = .99 SD = 1.01	.00 (95% CI [-.18, .18]), p = 0.998	.00
Godspeed: Animacy	242 243	M = 3.25 M = 3.30	SD = .96 SD = .96	-.05 (95% CI [-.22, .12]), p = 0.585	-.05
Godspeed: Likeability	242 243	M = 2.89 M = 3.00	SD = 1.10 SD = 1.04	-.10 (95% CI [-.30, .09]), p = 0.279	-.10
Godspeed: Perceived Intelligence	242 243	M = 3.20 M = 3.28	SD = 1.01 SD = 1.00	-.08 (95% CI [-.26, .10]), p = 0.367	-.08
Godspeed: Perceived safety	242 243	M = 3.31 M = 3.41	SD = 1.08 SD = 1.06	-.10 (95% CI [-.29, .09]), p = 0.309	-.09
Adjusted CBS-S: Mental preparation	242 243	M = 3.25 M = 3.34	SD = 1.03 SD = 1.03	-.10 (95% CI [-.28, .09]), p = 0.303	-.09
Adjusted CBS-S: Goal setting	242 243	M = 2.84 M = 2.71	SD = 1.04 SD = .98	.13 (95% CI [-.05, .31]), p = 0.163	.13
Adjusted CBS-S: Quality of coaching	242 243	M = 3.06 M = 3.02	SD = 1.06 SD = 1.04	.04 (95% CI [-.15, .23]), p = 0.680	.04
Group cohesion: Task cohesion	242 243	M = 3.53 M = 3.38	SD = 1.22 SD = 1.25	.16 (95% CI [-.06, .38]), p = 0.161	.13
Group cohesion: Social cohesion	242 243	M = 3.70 M = 3.48	SD = 1.26 SD = 1.21	.22 (95% CI [.00, .44]), p = 0.051	.18
Persuasion	82 243	M = 3.07 M = 3.07	SD = 1.68 SD = 1.68	.00 (95% CI [-.42, .42]), p = 0.997	.00
Interaction experience	242 243	M = 4.43 M = 4.28	SD = 1.73 SD = 1.67	.14 (95% CI [-.16, .45]), p = 0.357	.08

4.5.8 Discussion

Within-subject results discussion

As can be seen in Table 14, all the scales measuring our constructs showed significant results. We will briefly discuss what all of them mean, and discuss the potential implications. We will interpret effect sizes of d = 0.2 as small, d = 0.5 as medium, and d = 0.8 as large, following the guidelines by Cohen (Cohen, 1988). As was mentioned in the results section, for the Godspeed questionnaire series, a positive Cohen's d means a higher rating on the construct for the IPA model, as compared to the IPC model. For all the other constructs, a negative Cohen's d means a higher rating on the construct for the IPA model, as compared to the IPC model.

First off, in the Godspeed questionnaire series we found that the coaching team using the IPA model came off as more anthropomorphic (d = .27), animate (d = .26), likeable (d = .66), intelligent (d = .53), and people felt safer too (d = .42) when compared to the coaching team using the IPC model. These results suggest that with regards to how the coaching team came off, the IPA model was preferred in all respects to the IPC model. Especially the likeability, and intelligence stand out for their medium-large sized effects. All in all, there is a clear preference for the coaching team using the IPA model.

Next, taking a look at the adjusted CBS-S, we found that the coaching team using the IPA model came off as better able to help with mental preparation (d = -.32), as well as with goal setting (d = -.30), and was evaluated as giving a better quality of coaching (d = -.43) when compared to the coaching team using the IPC model. These results suggest that with regards to how the coaching team came off, the IPA model was preferred in all respects to the IPC model. Especially the quality of coaching construct stands out for its small-medium sized effect. All in all, there is a clear preference for the coaching team using the IPA model.

For the group cohesion questionnaire, we found that the coaching team using the IPA model came off as more cohesive both in task cohesion (d = -.46), as well as social cohesion (d = -.39) when compared to the coaching team using the IPC model. These results suggest that with regards to how the coaching team came off, the IPA model was preferred in all respects to the IPC model. The social cohesion effect was small-medium and the task cohesion effect nearly medium, but both meaningfully show a preference for the coaching team using the IPA model regardless.

For the persuasion question we found that the coaching team using the IPA model was more persuasive (d = -.29) when compared to the coaching team using the IPC model. This result suggest that the coaching team using the IPA model was better able to persuade participants than the coaching team using the IPC model. The effect was small-medium, but meaningfully shows a preference for the coaching team using the IPA model regardless.

Finally, in our interaction experience questionnaire we found that the coaching team using the IPA model gave the participants a more positive interaction experience (d = -.47) when compared to the coaching team using the IPC model. This result suggests that the coaching team using the IPA model had an interaction with the participants that was perceived as more pleasant than the coaching team using the IPC model. The effect was of near medium size, showing a decent amount of preference for the coaching team using the IPA model.

Similarly to the previous study in Section 4.3, all the results considered, it is clear that participants preferred their dialogue with the coaching team using the IPA model. This may again have to do with the fact that the coaching team using the IPC model had two hostile coaches, whereas the coaching team using the IPA model only had one negative coach. This may have given the coaching team using the IPC model a more negative tone. Furthermore, the dialogues with the coaching team using the IPC model usually opened with these more negative coaches, which may have primed people negatively. Finally, the coaching team using the IPA model may have had a more clear coach that provides information and asks for clarification, as compared to the coaching team using the IPC model due to the task-related: giving of information, and task-related: asking for information coaching roles in the IPA model.

Though the results for the within-subject comparisons are similar in the study described in Section 4.3 and the study described in Section 4.4, there are minor differences. The effect sizes are generally slightly

larger for the study described in Section 4.3 when compared to the study described in Section 4.4. This implies that there may have been a very minor impact of the peer mediator coach and conflict mediation content to close the gap between the dialogues using the IPA model and IPC model.

Between-subject results discussion

As can be seen in Table 14, the scales measuring the Group cohesion: task cohesion, and Group cohesion: social cohesion constructs for the IPA model showed significant results, and Group cohesion: social cohesion construct for the IPC model showed a result trending towards significance. All other results were insignificant. We will briefly discuss what all of the results mean, and discuss the potential implications. We will interpret effect sizes of d = 0.2 as small, d = 0.5 as medium, and d = 0.8 as large, following the guidelines by Cohen (Cohen, 1988). As was mentioned in the results section, for the Godspeed questionnaire series, a positive Cohen's d means a higher rating on the construct for study 1 (described in Section 4.3), as compared to study 2 (described in Section 4.4). For all the other constructs, a negative Cohen's d means a higher rating on the construct for study 1 (described in Section 4.3), as compared to study 2 (described in Section 4.4).

First off, in the Godspeed questionnaire series we found that the dialogue with the coaching team in study 1 (described in Section 4.3) and study 2 (described in Section 4.4) did not significantly differ from each other for the IPA model, as well as the IPC model. This means that the addition of the peer mediator coach and peer mediation content did not significantly impact how anthropomorphic, animate, likeable, or intelligent the coaching team was seen as, nor did it impact how safe participants felt.

Next, taking a look at the adjusted CBS-S, we found that the dialogue with the coaching team in study 1 (described in Section 4.3) and study 2 (described in Section 4.4) did not significantly differ from each other for the IPA model, as well as the IPC model. This means that the addition of the peer mediator coach and peer mediation content did not significantly impact how the coaching team came off with regards to being able to help with mental preparation, as well as with goal setting, and did not significantly impact how the quality of coaching was evaluated.

For the group cohesion questionnaire, we found that the coaching team in study 1 (described in Section 4.3) using the IPA model came off as less cohesive both in task cohesion (d = .28), as well as social cohesion (d = .24) when compared to the coaching team in study 2 (described in Section 4.4) using the IPA model. These results suggest that with regards to how the coaching team came off, for the IPA model the presence of the peer mediator coach and conflict mediation content was preferred in all respects to the peer mediator coach and peer mediation content not being there and instead leaving less well resolved conflicts. Furthermore, we found that the coaching team in study 1 (described in Section 4.3) using the IPC model came off as trending towards significantly less cohesive in social cohesion (d = .18, p = .051) when compared to the coaching team in study 2 (described in Section 4.4) using the IPC model. These results suggest that with regards to how the coaching team came off, for the IPC model the presence of the peer mediator coach and conflict mediation content might be preferred with respects to social cohesion to the peer mediator coach and peer mediation content not being there and instead leaving less well resolved conflicts. The group cohesion effects were small-medium for the IPA model, and very small and only trending towards significance in the case of social cohesion for the IPC model. This shows a meaningful, but small preference for the coaching team including a peer mediator coach and conflict mediation content. Furthermore, it shows that the impact was more clearly made in the dialogue with the coaching team using the IPA model than the dialogue with the coaching team using the IPC model.

For the persuasion question we found that the dialogue with the coaching team in study 1 (described in Section 4.3) and study 2 (described in Section 4.4) did not significantly differ from each other for the IPA model, as well as the IPC model. This means that the addition of the peer mediator coach and peer mediation content did not significantly impact how persuasive the coaching team was.

Finally, in our interaction experience questionnaire we found that the dialogue with the coaching team in study 1 (described in Section 4.3) and study 2 (described in Section 4.4) did not significantly differ from each other for the IPA model, as well as the IPC model. This means that the addition of the peer mediator coach and peer mediation content did not significantly impact how pleasant the interaction with the coaching team was perceived as.

The impact we see being mostly made on group cohesion makes sense. Conflicts can logically be seen as making the group seems less cohesive. Resolving these conflicts in a respectful manner together would logically lead to a team being seen as less at war with itself, and thus more cohesive. Thus, this result makes sense. It is interesting to note that this did not show as much for the coaching team that used the IPC model for their dialogue. Perhaps the two hostile coaches combined had too much of a negative impact on the perceived group cohesion to be rectified by the peer mediator coach and their conflict mediation.

4.6 Gesture generation evaluation

To evaluate our gesture generation model, we performed two types of evaluation studies: one objective study where we conducted 5 experiments including ablation studies and one perceptive study where we ask participants to rate videos of a virtual agent gesturing in two different conditions.

Table 15: Summary table for the study: "Gesture generation evaluation".

Study	Method	Setting	*N*	Participants <50	Participants >50	Participants with health conditions (DM-II, CP)
Gesture generation evaluation	Online followed by questionnaires (no interaction with researchers)	Online study	28	-	28	-

4.6.1 Objective Evaluation Study

We randomly split our data with the proportion of 64% training data, 16% validation data, and 20% testing data. This is chosen according to the common 80/20 rule. 80% of the data is for both training and validation and 20% of the data is for testing. The 80% is then split again 80% × 80% = 64% for training and 80% × 20% = 16% for validation. Each of the training, validation, and testing dataset contains a mix of samples from both speakers and different dialogues.

We perform five experiments. Experiment 1 is for obtaining the baseline performance by generating random outputs according to the data distribution. Experiment 2 is for obtaining the performance of the network by performing training and testing with our entire dataset. Experiment 3 is an ablation study to find out which features are more pertinent. A presence of pertinent features enables the model to perform prediction with a good performance. We replace some features with random values and retain the rest to find out how the performance of the model is being affected. Experiment 4 is to find out whether a model trained with one speaker only is generalizable to the dialogue counterpart. Experiment 5 is for finding out whether including eyebrow movements will lead to a higher performance on the "Beat" class.

To optimise the model, we vary the dimensions of the encoder and the decoder. The dimensions of the encoder and decoder are varied from 1 to 3, because our input data has three features. A challenge we face is that the loss function used in the training concerns only the matches at the same timestep, therefore ignoring the possibilities of shifts or dilations, which means that the network is not completely optimized for our objective. Therefore, we have to rely on the stochasticity of the neural network. In practice, it means we have to train the model many times to get a good result.

In Experiment 1 (random output), we generate random outputs according to the probability distribution of the gesture classes, while completely ignoring the prosody input. Specifically, we measure two sets of probabilities, namely the probabilities that a sample is started by a particular class and the probabilities that a class follows another (or the same) class. This is done because our data consist of sequences, where each element affects the next element. We match this result against the output from our ground truth. We do this 50 times and we measure the mean of their performances. This can be seen as an extremely simple predictor and thus can be seen as the baseline result. The result is shown in Table 16, Result 1.

In Experiment 2 (training and testing with the entire data), we train and test the neural network with the entire data with the 64%, 16%, and 20% split mentioned earlier. Note that in this data, we have two speakers performing several dialogues. We mix and shuffle the data, and then split them into training, validation, and testing data. The result is reported in Table 16, Result 2.

In Experiment 3 (ablation study), we want to observe how much the model we obtain in Experiment 2 learns about the structure of the data and how each feature affects the performance of the model. In order to do that, we use the model and data used in Experiment 2, but we replace some or all input features (intensity, F0, and F0 direction score) with random values. First, in order to observe how much the model learns the structure of the data, we randomise all input features (Table 16, Result 3). This way, we force the model to make "educated guesses" about the outputs without seeing the inputs. Unlike in Experiment 1 where the random outputs are generated based on two explicitly-set probability distributions, here we use a model whose prediction ability comes only from the training. Subsequently, we keep some features while randomizing the others in order to find which features are tied to gesture classes. In Table 16, Result 4), we keep only the intensity. In the Table 16, Result 5, we flip the condition, so we keep the F0 and F0 direction score. After that, to isolate the individual effect of the F0 and the F0 direction score, we keep the F0 only (Table 16, Result 6) and F0 direction score only (Table 16, Result 7).

In Experiment 4 (trained with one speaker, tested on the other speaker), we train the model with the first speaker and test it on the second speaker, and then we do the reverse. The results of both sub-experiments are in Table 16, Results 8 and 9. It should be noted that one speaker is a man and the other one is a woman.

In Experiment 5 (inclusion of eyebrow movements), in order to find out whether inclusion of eyebrow movements helps on predicting beat class, we compare the performance of the network when the data ignores the eyebrow movements, when the data considers upward eyebrow movements (Action Unit 1 or 2), and when the data considers both upward and downward eyebrow movements (Action Unit 1 or 2 or 4). We make 35 random permutations of our samples. For each sample in our dataset, we make three variations, namely the one which ignores the facial movements, the one where the presence of upward eyebrow movement is marked as a potentially-beat gesture, and the one where the presence of either upward or downward eyebrow movements is marked as a potentially-beat gesture. Then, we split them into training, validation, and testing datasets. Therefore, we have 35 × 3 = 105 unique training/validation/testing datasets. For each of them, we train and test the network 6 times and choose the one with the highest alignment score. For each variation of the three variations, we calculate the average alignment, insertion, and deletion scores of the 35 permutations. The results are shown in Table 17.

Table 16: Alignment, Insertion, and Deletion scores.

Exp 1: Random output result (Result 1)			
	Alignment	**Insertion**	**Deletion**
Beat	0.0	0.506	1.0
NonBeatStroke	0.123	0.389	0.854
NonBeatNonStroke	0.118	0.448	0.870
NoGesture	0.513	1.152	0.469
Exp 2: Trained and tested with the entire data (Result 2)			
	Alignment	**Insertion**	**Deletion**
Beat	0.213	3.863	0.763
NonBeatStroke	0.551	0.404	0.493
NonBeatNonStroke	0.234	0.141	0.730

NoGesture	0.558	0.507	0.405
Exp 3: All input features are randomised (Result 3)			
	Alignment	**Insertion**	**Deletion**
Beat	0.0	0.0	1.0
NonBeatStroke	0.098	0.851	0.914
NonBeatNonStroke	0.035	0.358	0.959
NoGesture	0.260	0.871	0.713
Exp 3: Using intensity only (Result 4)			
	Alignment	**Insertion**	**Deletion**
Beat	0.0	0.0	1.0
NonBeatStroke	0.170	1.280	0.864
NonBeatNonStroke	0.039	0.450	0.952
NoGesture	0.461	1.284	0.550
Exp 3: Using F_0 and the F_0 direction score only (Result 5)			
	Alignment	**Insertion**	**Deletion**
Beat	0.0	3.631	1.0
NonBeatStroke	0.564	0.802	0.529
NonBeatNonStroke	0.201	0.345	0.793
NoGesture	0.558	0.635	0.480
Exp 3: Using F_0 only (Result 6)			
	Alignment	**Insertion**	**Deletion**
Beat	0.116	4.706	0.913
NonBeatStroke	0.579	0.552	0.521
NonBeatNonStroke	0.202	0.286	0.782
NoGesture	0.547	0.589	0.496
Exp 3: Using F_0 direction score (Result 7)			
	Alignment	**Insertion**	**Deletion**
Beat	0.0	0.0	1.0
NonBeatStroke	0.091	0.966	0.918

NonBeatNonStroke	0.041	0.359	0.969
NoGesture	0.292	0.797	0.693
Exp 4: Trained with 1st speaker tested on the 2nd speaker (Result 8)			
	Alignment	Insertion	Deletion
Beat	0.086	2.582	0.910
NonBeatStroke	0.604	0.479	0.486
NonBeatNonStroke	0.274	0.237	0.688
NoGesture	0.538	0.371	0.474
Exp 4: Trained with 2nd speaker tested on the 1st speaker (Result 9)			
	Alignment	**Insertion**	**Deletion**
Beat	0.111	3.690	0.841
NonBeatStroke	0.432	0.633	0.626
NonBeatNonStroke	0.298	0.326	0.715
NoGesture	0.478	0.405	0.480

Table 17: The effect of inclusion of eyebrow movements on the beat performance (experiment 5).

	Alignment	Insertion	Deletion
No eyebrow information	0.257	2.620	0.745
With upward eyebrow movement	0.246	1.448	0.742
With upward/downward eyebrow movement	0.415	0.298	0.579

4.6.2 Perceptive Evaluation Study

In the subjective study, we ask 28 respondents to watch 12 videos. The 12 videos consist of 6 pairs. The sequence of the videos is shuffled, therefore the respondent cannot easily guess the pairs. Each pair itself contains a video whose timing is based on the output of the gesture generation model and a baseline video. The baseline video is taken from the Gest-IS English corpus. It serves as ground truth. The gestures displayed by the humans in the corpus are reproduced by the virtual agent but the sequence of the gesture is shuffled. Our objective is to compare the respondent's perception differences between the videos based on the output of the gesture generation model and the baseline videos. An example of the video is shown in Figure 19. We compared the naturalness, the time consistency, and the semantic consistency of the videos. For each of those factors, we measure it by having three questions. Each question asks the user to give a rating from one to five. We sum the respondent's scores on the three questions to get the score of the quality we want to measure.

We get the survey respondents from Prolific. We show the box plots of the results in Figure 20, Figure 21, and Figure 22. In all the three measured qualities, namely naturalness, time consistency, and semantic consistency, the videos created based on the output of the gesture generation model have higher average scores (one-way ANOVA test). However, the differences between the two groups of videos, namely the ones created from the output of the gesture generation model and the ones created by shuffling the sequence of gestures of the ground truth, are only significant on one quality, namely time consistency.

Figure 19: A video used in the subjective study of the gesture generation work.

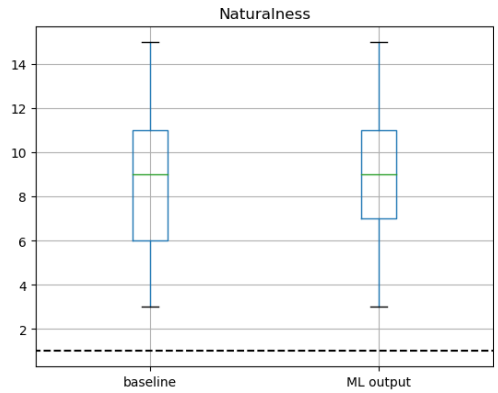

Figure 20: The box plot of the naturalness scores obtained from the gesture generation's perceptive study.

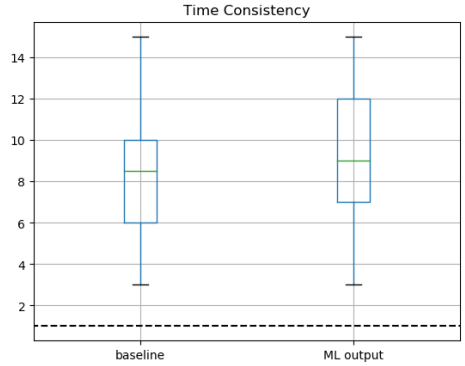

Figure 21: The box plot of the time-consistency scores obtained from the gesture generation's perceptive study.

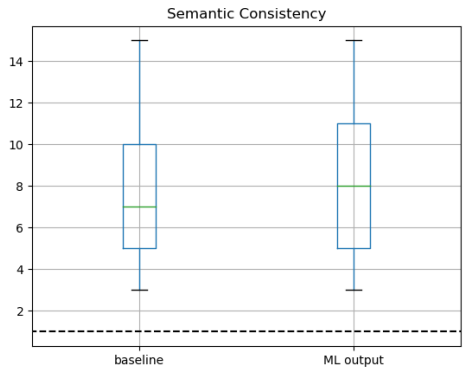

Figure 22: The box plot of the semantic-consistency scores obtained from the gesture generation's perceptive study.

4.7 Cohesive group evaluation

The cohesive group evaluation is a two condition between-subject study. The scenario consists of four coaches (virtual agents) interacting with each other and the user. There are three different dialogues that have been developed based on two different models (Bales and Leary) on weight loss, stress management and sleep as mentioned in the previous evaluation study. We make use of the same dialogue to evaluate our model. For this evaluation study we choose only one topic, i.e., stress management based on Bales' model.

This study was initially planned to take place in a laboratory setting with the participants interacting with the coaches in real-time. However due to the recent health regulations that have been placed to limit the number of people in the lab and 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 in a way that it feels 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 was played. In the next section we provide details of the study and discuss the results.

Table 18: Summary table for the study: "Cohesive group evaluation".

Study	Method	Setting	*N*	Participants <50	Participants >50	Participants with health conditions (DM-II, CP)
Cohesive group evaluation	Online interaction followed by questionnaires (no interaction with researchers)	Online study	32	-	32	32

4.7.1 Objectives

The main objective of this evaluation study is to understand the perception on the Council of 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.

4.7.2 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.

4.7.3 Questionnaire

We make use of two pre-study questionnaires and three post-study questionnaires. The pre-study questionnaires measure the NARS (a priori) and how easy the user is persuaded. The post-questionnaire measures the cohesiveness, the credibility and the persuasiveness of the group. Following are the list of questions used.

NARS

1. I would hate the idea that virtual agents or artificial intelligence was making judgements about things.
2. I feel that if I depend on virtual agents too much, something bad might happen.
3. I would feel paranoid talking with a virtual agent.
4. I feel that in the future society will be dominated by virtual agents.

Persuadability

1. I always follow advice from my general practitioner.
2. I am very inclined to listen to authority figures.
3. I always obey directions from my superiors.
4. I am more inclined to listen to authority figure than to a peer.
5. I am inclined to follow the advice I read on trusted websites.

Cohesiveness

1. Overall, the team members appear to be collaborative.
2. I feel that team members share the same purpose/goal/intentions.
3. Overall, the team members seem to be supportive towards each other.
4. Overall, the work group appears to be in tune/in sync with each other.

Credibility

1. The group of coaches are trustworthy.
2. The group of coaches are reliable.
3. The group of coaches show expertise.

Persuasiveness

1. The group provides me with the means to decide what I need to do.
2. The group helps me decide how to approach the problem.
3. The group had an influence on me.

4.7.4 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 of users interacted with agents that display behaviours generated by a random behaviour generator. The second group of participants interact with agents that display cohesive group behaviours generated by our model. The behaviours we focus on are gaze, smile and head nods. The agent appearance and dialogue content remain the same for both the groups. We made use of the four agents developed for the Council of Coaches project for this study as shown in Figure 23 below. Based on the results from our previous study on persuasiveness, we gave the role of providing advice to an older authoritative agent while the supportive coach was assigned to a younger peer coach. 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 in total lasted for about 3 minutes.

Figure 23: A screenshot of the interaction with the coaches.

4.7.5 Procedure

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

4.7.6 Results and discussion

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 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.

5 Conclusion

In this document, we report the final adjustments we have integrated into the technical prototype developed for the Council of Coaches project. The overall architecture is composed of four layers which include “Sense”, “Remember”, “Think” and “Act” layers. We also present the latest developments of the Greta and ASAP agent platform.

We have presented a further analysis of cohesion in multi-party interactions which focuses on verbal and non-verbal behaviours during conversation. The results show that certain verbal, non-verbal social cues and interruptions have an impact on level of cohesion. The results from this work have contributed towards developing a computational model to simulate a cohesive group of virtual agents.

We described six evaluation studies in the context of Council of Coaches system. The studies focus on the evaluation of different aspects of the final Council of Coaches Technical Prototype such as the user interface usability evaluation, multi-device interaction, impact of verbal conflict presentation styles, gesture generation, and cohesive group evaluation, in the context of Council of Coaches system.

6 Bibliography

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. Language Resources and Evaluation.

Baltrusaitis, T., Zadeh, A., Lim, Y., & Morency, L. (2018). Openface 2.0: Facial behaviour analysis toolkit. IEEE Face and Gesture Recognition, pp. 59-66.

Bangor, A. K. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. Journal of usability studies, 114-123.

Bosdriesz, L. (2020). Opportunities and Challenges for Adding Speech to Dialogues with a Council of Coaches. Enschede, The Netherlands: University of Twente - Technical report (available on request).

Brooke, J. (1986). System usability scale (SUS): a quick-and-dirty method of system evaluation user information. Reading, UK: Digital Equipment.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., . . . others. (2005). The AMI meeting corpus: A pre-announcement. International workshop on machine learning for multimodal interaction, pp. 28-39.

Casey-Campbell, M., & Martens, M. (2009). Sticking it all together: A critical assessment of the group cohesion performance literature. International Journal of Management Reviews, 223-246.

Dohsaka, K., Asai, R., Higashinaka, R., Minami, Y., & Maeda, E. (2009). Effects of conversational agents on human communication in thought-evoking multi-party dialogues. Proceedings of the {SIGDIAL} 2009 Conference (pp. 217-224). London, UK: Association for Computational Linguistics.

Hung, H., & Gatica-Perez, D. (2010). Estimating cohesion in small groups using audio-visual nonverbal behavior. IEEE Transactions on Multimedia, 563-575.

Kantharaju, R., De Franco, D., Pease, A., & Pelachaud, C. (2018). Is Two Better than One?: Effects of Multiple Agents on User Persuasion. Proc. of the 18th International Conference on Intelligent Virtual Agents (pp. 255--262). ACM.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. The relation between verbal and nonverbal communication.

Niebuhr, O., & Pfitzinger, H. (2010). . On pitch-accent identification-The role of syllableduration and intensity. 5th International conference on Speech Prosody.

O'Brien, H., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. International Journal of Human-Computer Studies, 112(1071-5819), 28-39.

Petersen, T. (2020). Council of Coaches in Virtual Reality. Enschede, The Netherlands: University of Twente - Bachelor's thesis (available on request).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., . . . Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society.

Ravenet, B., Cafaro, A., Biancardi, B., Ochs, M., & Pelachaud, C. (2015). Conversational Behavior Reflecting Interpersonal Attitudes in Small Group Interactions. Proc. International Conference on Intelligent Virtual Agents (pp. 375--388). Springer.

Saint-Amand, K. (2018). Gest-IS: Multi-lingual Corpus of Gesture and Information.

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 (IVA '18). Association for Computing Machinery, New York, NY, USA, 43–50. DOI: https://doi.org/10.1145/3267851.3267882

van Welbergen H., Yaghoubzadeh R., Kopp S. (2014) AsapRealizer 2.0: The Next Steps in Fluent Behavior Realization for ECAs. In: Bickmore T., Marsella S., Sidner C. (eds) Intelligent Virtual Agents. IVA 2014. Lecture Notes in Computer Science, vol 8637. Springer, Cham. https://doi.org/10.1007/978-3-319-09767-1_56

van der Werff, R. (2020). Implementing Virtual Reality in the Council of Coaches system. Enschede, The Netherlands: University of Twente - Bachelor's thesis (available on request).

Voigt, R., Podesva, R., & Jurafsky, D. (2014). Speaker movement correlates with prosodic indicators of engagement. Speech Prosody.

Acknowledgements

The Council of Coaches project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement #769553. This result only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains.

Headings and titles in this document, as well as the Council of Coaches logo use the Comfortaa font, designed by Johan Aakerlund and Cyreal and licensed under the Open Font License[1].

Additional text in this document uses the Roboto font, designed by Christian Robertson and licensed under the Apache License, Version 2.0[2].

The Council of Coaches logo and Blobmen graphics were *drawn freely* in Inkscape, licensed under the GNU General Public License[3].

[1] Open Font License: http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=OFL_web
[2] Apache License, Version 2.0: http://www.apache.org/licenses/LICENSE-2.0
[3] Inkscape License Information: https://inkscape.org/about/license/