Longitudinal Evolution of Coachees' Behavioural Responses to Interaction Ruptures in Robotic Positive Psychology Coaching

Micol Spitale[1]*, Minja Axelsson[1]*, Neval Kara[2] and Hatice Gunes[1]

***Abstract*— Robotic mental well-being coaches could be used to help people maintain their well-being, and improve access to mental healthcare. In coaching, the alliance between the coach and coachee is important for the success of the practice. However, this alliance might be negatively affected by *interaction ruptures* (e.g., the robot making mistakes and the user feeling awkward) that still commonly occur in human-robot interactions. Therefore, robotic coaches should be able to recognize ruptures occurring during their interactions with human users to guarantee the success of the well-being practice. To this aim, we analyse coachee behavioural responses to interaction ruptures during a robotic positive psychology coaching practice and how these behavioural cues evolve over time. We focus our analysis on a dataset we collected in a previous work, where 26 participants interacted with either a QTrobot or a Misty II robot at their workplace over 4 weeks. We undertake a longitudinal analysis of coachees' multimodal non-verbal cues (i.e., facial expressions, vocal acoustic features, and body pose features) to investigate the contribution of individual modalities for detecting interaction ruptures. Our results show that coachees: i) displayed facial cues of rupture (e.g, laughing at the robot) and suspicion more in the first week than in the last week; ii) talked more and were less silent in the last week than in the previous weeks; and iii) exhibited a higher number of hand-over-face gestures (a cue for self-disclosure) in the last week than in the previous weeks. Our findings aim to inform the development of AI models for multi-modal detection of interaction ruptures which can be used to improve the effectiveness and the success of robotic well-being coaching.**

I. INTRODUCTION

The prevalence of mental health issues such as depression and anxiety has been increasing [1], and COVID-19 has exacerbated this [2]. One of the World Health Organization's [3] main goals is to improve the mental health of individuals and society at large. This includes the promotion of mental well-being, the prevention of mental issues, and a higher number of efforts to increase access to quality mental health care. However, these objectives are yet to be accomplished in our society because of the wide gap between those who require care and those who have access to it. Robotic coaches could help with alleviating this problem by promoting mental well-being and providing affordable and easy access to mental well-being-related practices [4]–[6].

In coaching, the alliance between the coach and coachee is important for the success of coaching over time [7]. Safran et al. [8] defined the *rupture* of the therapeutic alliance as the "deterioration in the quality of the relationship between patient and therapist". In robotic coaching, such *ruptures* can occur when, for example, the robot makes mistakes during the coaching session which might negatively impact the alliance and trust perceived by the user towards the robot, which may lead to an unsuccessful well-being practice. Recent studies [9], [10] reported the occurrence of interaction ruptures when deploying robotic coaches in the workplace. Spitale et al. [9] showed that participants got frustrated when the robot interrupted them (e.g., when the robot mistakenly detected that the user's speech was done when the person was still talking) or when the robot took a very long time to respond (e.g., due to issues in internet connectivity). Axelsson et al. [10] highlighted that the participants expected the robot to be more interactive and responsive in order to be perceived as a mindfulness coach. These findings suggest that a smooth conversation and interaction flow is critical to the success of a mental well-being practice. However, interaction ruptures are still known to be very common during human-robot interactions [11], and they can negatively impact the user's trust towards the robot [12]. Furthermore, previous studies [10], [13], [14] have shown that participants adapt to the robot behavior over time. For example, Axelsson et al. [10] showed that over time participants experienced the robot-led mindfulness practice to be more helpful. Hence, the first step towards solving the above-mentioned issues to ultimately help improve the coach-coachee alliance is to better understand the behavioural responses of the coachees over time in robotic well-being coaching sessions.

This paper presents a longitudinal analysis of the coachees' behavioural responses to interaction ruptures occurring in robotic well-being coaching. The analysis was conducted on the data acquired in the study reported in [9], where 26 participants interacted with either a QTrobot or a Misty II robot over 4 weeks. The robotic well-being coaches delivered a positive psychology exercise (e.g., savouring, gratitude) on a weekly basis for about 10 minutes. In this paper, we focus on understanding *how the interaction ruptures and the coachees' behavioural responses evolve over these 4 weeks*.

The main contribution of this paper is three-fold:

1) analysing the evolution of coachees' behaviours across four weeks to evaluate how coachees adapt to robotic well-being coaching;
2) identifying the coachees' behavioural cues of interaction ruptures that can inform the development of AI models for automatic detection;
3) investigating the contribution of individual modalities

*Both authors equally contributed to this work. N. Kara contributed to this work while undertaking undergraduate research with the AFAR Lab at the Dep. of Computer Science & Technology, University of Cambridge.

[1]University of Cambridge, [2]Cankaya University
ms2871@cam.ac.uk, minja.axelsson@cl.cam.ac.uk, karaaneval@gmail.com, hatice.gunes@cl.cam.ac.uk

for detecting interaction ruptures to inform the design of multi-modal AI-driven models.

II. BACKGROUND AND RELATED WORK

A. Coaching and Ruptures in Coaching

Coaching for mental well-being aims to support the coachee in flourishing in their life [15] (cf. psychological therapies which aim to treat mental illness). Positive psychology coaching aims to encourage the coachee to focus on the positive aspects of their life [16], in order to improve life satisfaction and positive affect [17]. Positive psychology practices include, e.g., savouring, where the coachee is encouraged to recall positive memories and their positive emotions. One of the aspects important for successful coaching is the alliance between the coach and coachee [7]. Difficulties in the coach-coachee relationship has been found as one of the barriers to the effectiveness of coaching [18]. Such difficulties may include the coach struggling with the concepts of coaching, not being sensitive enough, being vague, being too focused on one area and not being flexible, and not being involved or supportive in the coaching session [18]. The relationship between the coach and coachee can also be negatively impacted if the coachee perceives the coach as not involved or supportive [18]. Resolving these issues can help improve the coach-coachee relationship and alliance.

Interaction ruptures, i.e., difficulties in the coaching dynamics during a coaching session, might negatively affect the coach-coachee relationship. *Interaction ruptures* might occur when the coachee perceives the coach to be *off* in their responses and the timing of their responses. In such awkward social interactions (i.e., interactions where ruptures occur), coachees may feel nervous, confused, embarrassed, uncertain or self-conscious [19]. Embarrassment can occur when interaction partners do not behave according to a social script, rules/norms, or roles [19], [20]. Behavioural expressions of awkwardness in social situations may include smiling and smile control, laughter, gaze shifting, and fidgeting [19], [20].

B. Robotic Well-being Coaches

The design and investigation of robotic coaches for well-being have seen an increasing trend [4], [21]–[23]. The main reason for such an interest is the various advantages they provide - e.g. physical presence (in comparison to a mobile app), easily accessible with a consistent behaviour (as compared to a human coach) [22]. A disadvantage may be unwanted robot behaviour such as misunderstandings in the communication (in comparison to a human coach) [22]. Robotic positive psychology coaches have been tested with emotional adaptation [24], [25], and deployed with students [6], [26], [27] and in the workplace [9]. Robotic mindfulness meditation coaches have been deployed in lab settings [14], [28], [29], at a wellness centre [30], and at a public cafe [10]. Robots have also been used to assess the mental well-being of children [5], [31]. Furthermore, recent studies have explored the use of robotic mental well-being coaches over time [9], [10], [32]. Their results showed that participants undertaking the well-being coaching with the robot adapt to the robotic coach's behaviour over time. For example, Axelsson et al. [10] deployed two robotic well-being coaches in a public cafe over 4 weeks to deliver mindfulness sessions. Participants reported getting used to the robotic coach over time and found it more helpful over time. However, none of these works have explored how the *interaction ruptures* during robotic well-being coaching affected coachees' behaviours.

C. Robot Failures and Repair Strategies in HRI

Within the Human-Robot Interaction (HRI) context, interaction ruptures can be caused by robot mistakes (i.e., robot failures), which could jeopardize the well-being coaching. Many HRI studies have explored *robot failures*, providing taxonomies, e.g., [11], [33], and demonstrating that robots could use different strategies to repair the perceived trust in the robot [34]. For example, Sebo et al. examined different trust repair strategies when a robot intentionally broke a human player's trust during a game [35]. They found that an apology was a more effective trust repair strategy than denial. Esterwood and Robert also examined how a robot's different trust repair strategies changed people's trust in a robot when it made mistakes during a box sorting task [36], [37], finding that explanations were the most effective strategy for repairing trust after multiple violations. Kontogiorgos et al. examined human non-verbal behaviour reactions to conversational failures, where the Furhat robot was simulating errors during a cooking instruction class [38], [39]. They found that severe errors may decrease users' trust in the robot [40]. Lindgren examined conversational failures with a robot in a Wizard-of-Oz study and determined a taxonomy of errors and potential mitigation strategies based on a thematic analysis [41]. These works contributed greatly to our understanding of robot errors, however they did not investigate robot failures in social interactions, and in particular robot failures during *well-being coaching*.

D. Participant Behaviour Analysis in HRI

Interaction ruptures brought on by either the robot or the human frequently affect interactions between humans and robots. Past studies analysed human behaviours to recognize these ruptures. Giuliani et al. [42] investigated the verbal and non-verbal social signals that humans show when error situations occur in HRI experiments. Alghowinem et al. examined how robots could sense participants' engagement through non-verbal behaviours during positive psychology coaching [27]. Data of self-disclosure and non-self-disclosure was manually annotated, and non-verbal behaviour features relating to body gestures, acoustic signals and head orientation were extracted. The most salient features (e.g., body features related to hand-over-face gestures) were then selected based on the Feature Selection Framework. The study found that non-verbal cues alone (without linguistic features) can be used to detect self-disclosure. Kontogiorgos et al. examined non-verbal behavioural cues in relation to predetermined conversational breakdowns in human-robot inter-

Fig. 1. Coachee interacting with the Misty II robot. The data was collected via the Jabra microphone placed on the table, a video camera on the side of the robot, and a GoPro on a side table to capture the lateral view of the interaction.

action [39]. These works helped us understand the relevant factors for investigating users' non-verbal behaviours during HRI. However, no past work has investigated the *longitudinal* evolution of robot failures together with participants' *multimodal* behavioural responses. In this paper we focus on this particular research problem in the context of robotic well-being coaching.

III. METHODOLOGY

This section describes the methodology for the dataset collection, the definition of variables for interaction ruptures, the annotation process, the extraction of behavioural features, and the longitudinal analysis.

A. Dataset Collection

In our previous study [9], we deployed a robotic positive psychology coach at a workplace, Cambridge Consultants Inc., over four weeks. Coachees ($n = 26$) were Cambridge Consultants Inc.'s employees, and we screened them to exclude higher levels of anxiety and depression before the study, in order to recruit healthy coachees. Coachees interacted with the robot once a week, with the first group interacting with the child-like QTrobot, and the second group interacting with the toy-like Misty II robot. The robotic coaches conducted the following positive psychology exercises, which were adapted from existing interventions: (1) savouring [43], (2) gratitude [44], (3) accomplishments [44], and (4) optimism about the future [45]. The robot interaction was pre-scripted and the follow-up questions that it asked to the coachees were independent from their responses (the robot was only able to detect when the coachee stopped speaking to continue the conversation). During the interaction, we collected video recordings (coachee's face and a side view of the interaction) and audio recordings (both the coachee's and robot's speech) using two cameras (a frontal video camera and a lateral GoPro) and a Jabra microphone as shown in Figure 1. The study was approved by the Ethics Committee of the Computer Lab, University of Cambridge.

B. Interaction Ruptures and Expressions of Awkwardness

Repairing interaction ruptures during coaching could help with maintaining a successful coach-coachee relationship. A robotic coach should be able to recognize interaction ruptures, for example, when a user is expressing awkwardness, or if the robot has made a mistake, and then it should attempt to repair these. In order to analyse the occurrence of interaction ruptures in robotic well-being coaching, we define the following measures, based on a data-driven approach, that relies on manual annotation of the videos of the robotic coaches' interactions with coachees:

User Awkwardness: The coachee displays behaviours that express that the interaction is awkward, and they may look confused, uncertain, distressed or uncomfortable.

Robot Mistake: The robot makes a mistake such as interrupting the coachee, not responding to the coachee, or responding with an utterance that is not appropriate for what the coachee has just said.

Interaction Rupture: We define an interaction rupture as either the presence of user awkwardness or a robot mistake, or both.

1) Annotation Process: We annotated the videos using the ELAN[1] video annotation tool, as it has been commonly used in other human behaviour annotation tasks (e.g., in [46]). We marked instances of user awkwardness and robot mistakes with binary labels (i.e., 1: present, or 0: absent), marking the time when the displays of user awkwardness or robot mistakes start and end. We observed that each coachee expressed awkwardness differently, therefore the label of 'presence' or 'absence' was determined by the researcher following the abovementioned definitions.

2) Measures: We defined three measures as indexes of an interaction rupture. These measures were defined in order to examine the behaviours coachees displayed when the well-being coaching was disrupted and the behaviours they displayed when the robot made mistakes, together with how often these two measures co-occurred (i.e. how related they are). The three measures are defined as follows:

- **User Awkwardness Index (UAI)**: This variable refers to the percentage of the sum (K) of the j number of occurrences, within a time interval (t_j), during which a user (u) displayed cues of awkwardness, with respect to the duration of the whole interaction (T):

$$UAI_u = \frac{\sum_{j=1}^{K} (t_j)^u}{T^u} * 100\% \quad (1)$$

- **Robot Mistake Index (RMI)**: This variable refers to the percentage of the sum (M) of the j number of occurrences, within a time interval (t_j), during which the robot made mistakes while interacting with a user (u), with respect to the duration of the whole interaction (T):

$$RMI_u = \frac{\sum_{j=1}^{M} (t_j)^u}{T^u} * 100\% \quad (2)$$

- **Interaction Rupture Index (IRI)**: This variable refers to the union of the occurrences of user awkwardness

[1] https://archive.mpi.nl/tla/elan

(UAI) and robot mistake (RMI) instances during an interaction with the user (u):

$$IRI_u = UAI_u \vee RMI_u \tag{3}$$

C. Behavioural Feature Extraction

We used off-the-shelf state-of-the-art methods to extract behavioural features from the audio-visual data collected in the study and this section describes the methods used for facial, audio, and body feature extraction. We explored facial and audio features because they were shown to be markers of robot failures [38], and we investigated the body cues related to self-touching behaviours or hand-over-face gestures [47] because these were shown to be a marker for self-disclosure [27].

1) Facial Features: We processed the video recordings using the OpenFace 2.2.0 toolkit [48] and extracted the presence and intensity of the 17 facial action units (AUs) provided by OpenFace, to measure the facial cues of the coachees, namely AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU5 (upper lid raiser), AU6 (cheek raiser), AU7 (lid tightener), AU9 (nose wrinkler), AU10 (upper lip raiser), AU12 (lip corner puller), AU14 (dimpler), AU15 (lip corner depressor), AU17 (chin raiser), AU20 (lip stretcher), AU23 (lip tightener), AU25 (lips parted), AU26 (jaw drop) and AU45 (blink) resulting in a total of 34 facial features. We then computed the average, median and standard deviation of each raw facial feature, which resulted in a (26x4)x102 facial feature vector.

2) Audio Features: We analysed the audio recordings using the openSMILE toolbox [49] and we extracted interpretable speech features, namely loudness and pitch. Using these, we computed other high-level features – such as the length of the coachees' silence and speech – by processing the audio recordings via HuggingFace library[2]. We diarized the speech and computed the duration of the voice detected by the coachee and the robot separately. We then computed the average and standard deviation of all the speech features, which resulted in a (26x4)x18 speech feature vector. We reduced the audio feature set using a PCA which found the most contributing sound indicators of IRI.

3) Body Features: We processed the video recordings using the OpenPose toolbox [50] and extracted the 25-2D body key points to estimate the movement of the torso, hands, arms, and head. Specifically, we extracted touch-behaviour-related features by computing the Euclidean distance between the two hands, and their distances to key points on the face, chest, and shoulders. We extracted touch-behaviour related features specifically due to previous literature reporting these to be relevant for self-disclosure during robotic coaching [27]. Additionally, we computed the velocity of touch behaviours by computing the differences in position in a frame-by-frame manner. We extracted 36 proximity and velocity features for touch behaviours, which resulted in a (26x4)x36 body feature vector.

[2]https://huggingface.co/pyannote/speaker-diarization

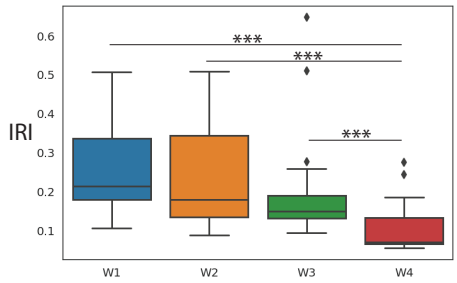

Fig. 2. Interaction rupture index (IRI) during the robotic well-being coaching across the four weeks (W1, W2, W3, and W4), *** $p < .0001$.

D. Longitudinal Analysis

As our original study was on longitudinal robotic well-being coaching, we focused our analysis on understanding the evolution of coachees' behavioural responses over time. We first checked the normality assumption for our distribution (conducting Kolmogorov–Smirnov test). Our results showed that the sample was not normally distributed, therefore we adopted non-parametric statistical tests for our analysis.

Given the fact that the subjects interacted with two different robots (either QTrobot or Misty II robot), we checked whether the robot form has an impact on the interaction ruptures. Our results did not show any statically significant difference between the robot groups except for week four where we found that the occurrences of the interaction ruptures in the group interacting with Misty II were significantly ($Z = 104$, $p < .05$) higher than in the group interacting with QTrobot. Qualitative data collected from participant interviews during the study [9] indicated that participants found the exercise in week four (optimism about the future) to be the most challenging. This could have influenced participants' behaviours during the robotic well-being coaching. Participants may also have expressed more awkward behaviours with the QTrobot due to having higher expectations because of its human-like form (cf. [9]).

Based on this preliminary analysis, we did not consider the robot form as a condition for the rest of the analyses in this paper. We first conducted Friedman tests and the post-hoc Wilcoxon signed-ranked tests to compare the differences in coachees' behavioural responses across the weeks for the interaction rupture measures (dependent variables). In this case, the dependent variables were the facial, body, and audio features while the independent variable was the time (weeks). We conducted the statistical analysis using the Python library Stats[3].

IV. RESULTS

This section reports the longitudinal evolution of the interaction ruptures and the coachees' behavioural responses to those ruptures, across the four weeks of the study.

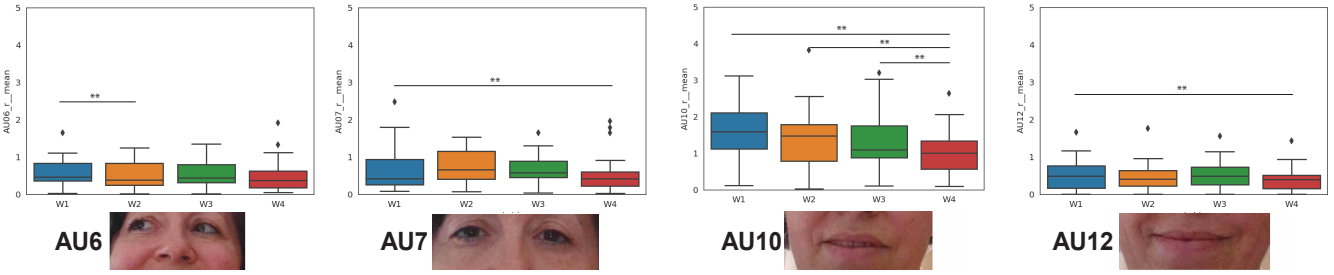

Fig. 3. Facial AU6, AU7, AU10, and AU12 mean differences across four weeks (W1, W2, W3, and W4). $^{**}p < .01$

A. Longitudinal Evolution of Interaction Ruptures

We conducted Friedman tests for the UAI, RMI, and IRI measures to evaluate differences in 23 coachees' interactions over four weeks. We discarded the data of 3 subjects because of missing video recordings due to technical errors. The results showed significant differences for UAI ($\chi^2 = 23.71$, $p < .001$), RMI ($\chi^2 = 40.56$, $p < .001$), and IRI ($\chi^2 = 33.31$, $p < .001$), see Figure 2. We therefore ran a pair-wise comparison using Wilcoxon signed-rank tests with a Bonferroni correction ($.05/4$). The post-hoc analysis results

TABLE I
POST-HOC ANALYSIS RESULTS FOR THE UAI, RMI, AND IRI DEPENDENT VARIABLES.

Variable	Pair-wise (a, b)	Z	p	Mdn (a)	Mdn (b)
UAI	(W1, W4)	7	<.001	0.14	0.03
UAI	(W2, W4)	34	<.005	0.08	-
UAI	(W3, W4)	37	<.005	0.06	-
RMI	(W1, W4)	0	<.0001	0.16	0.06
RMI	(W2, W4)	4	<.0001	0.13	-
RMI	(W3, W4)	21	<.0001	0.10	-
IRI	(W1, W4)	0	<.0001	0.20	0.07
IRI	(W2, W4)	7	<.0001	0.18	-
IRI	(W3, W4)	13	<.0001	0.15	-

are reported in Table I. In summary, the weekly occurrences of interaction ruptures (IRI) decreased significantly from W1 to W4.

B. Longitudinal Evolution of Coachees' Behavioural Responses

1) Facial Cues: To evaluate the differences in facial cues of coachees over four weeks, we conducted Friedman tests for the facial AUs extracted. The results showed significant differences for the mean intensity of AU1, i.e., inner brow raise ($\chi^2 = 16.98$, $p < .001$), for the mean intensity of AU2, i.e., outer brow raiser ($\chi^2 = 12.03$, $p < .01$), for the mean intensity of AU5, i.e., upper lid raiser ($\chi^2 = 17.76$, $p < .001$), for the mean intensity of AU6, i.e., cheek raiser ($\chi^2 = 8.63$, $p < .05$), for the mean intensity of AU7, i.e., lid tightener ($\chi^2 = 12.55$, $p < .01$), for the mean intensity of AU10, i.e., upper lip raiser ($\chi^2 = 13.59$, $p < .01$), for the mean intensity of AU12, i.e., lip corner puller ($\chi^2 = 10.93$, $p < .05$), for the median intensity of AU15, i.e., lip corner depressor ($\chi^2 = 23.27$, $p < .0001$), and for the mean intensity of AU23, i.e., lip tightener ($\chi^2 = 9.10$, $p < .05$). We, therefore, ran the pair-wise comparison using Wilcoxon signed-rank tests with a Bonferroni correction ($.05/4$) only for the facial AU features with intensity median levels higher than 0.5. The post-hoc analysis results are reported in Table

TABLE II
POST-HOC ANALYSIS RESULTS FOR THE FACIAL AUS THAT SHOWED STASTICALLY SIGNIFICANT DIFFERENCES.

Variable	Pair-wise (a, b)	Z	p	Mdn (a)	Mdn (b)
AU6	(W1, W2)	53	<.01	0.47	0.39
AU7	(W1, W4)	42	<.01	0.41	0.43
AU10	(W1, W4)	45	<.01	1.59	1.00
AU12	(W1, W4)	41	<.01	0.49	0.02
AU15	(W1, W4)	29	<.01	0.04	0.02
AU15	(W2, W4)	16	<.01	0.06	-
AU15	(W3, W4)	33.5	<.01	0.05	-

II. In summary, the lid tightener (AU7), the upper lip raiser (AU10), the cheek raiser (AU6) and lip corner puller (AU12) were displayed significantly less in W4 with respect to W1.

2) Audio Cues: To evaluate the differences in audio cues of coachees over the four weeks, we conducted Friedman tests for the audio features extracted (see Figure 4). The results showed significant differences for the speaking length ($\chi^2 = 27.21$, $p < .0001$), for the mean of the speaking length ($\chi^2 = 20.48$, $p < .001$), for the standard deviation of the speaking length ($\chi^2 = 14.32$, $p < .01$), for the silence length ($\chi^2 = 29.97$, $p < .0001$), for the mean of the silence length ($\chi^2 = 34.56$, $p < .0001$), for the mean of the loudness ($\chi^2 = 33.21$, $p < .0001$), and for the standard deviation of the loudness ($\chi^2 = 29.14$, $p < .0001$). Again, we run a pair-wise comparison using Wilcoxon signed-rank tests with a Bonferroni correction ($.05/4$). The post-hoc analysis results are reported in Table III. In summary, coachees spoke for significantly greater lengths of time in W4, and there was less number of silent segments, with respect to the previous weeks.

3) Body Cues: We compared the self-touching and hand-over-face gestures calculated from the body cues, over four weeks, by running Friedman tests for the body features extracted (see Figure 5). The results showed significant differences for the mean of the distance between keypoints 4 (left hand) and 18 (right side of face) ($\chi^2 = 8.43$, $p < .05$), for the median of the distance between keypoints 4 and 18 ($\chi^2 = 14.99$, $p < .01$), for the median of the distance between keypoints 7 (right hand) and 17 (left side of face)

[3]https://docs.scipy.org/doc/scipy/tutorial/stats.html

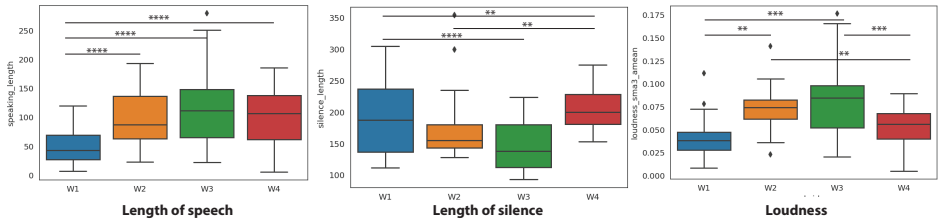

Fig. 4. Audio feature mean differences across four weeks (W1, W2, W3, and W4). $^{**}p < .01$, $^{***}p < .001$, $^{****}p < .0001$

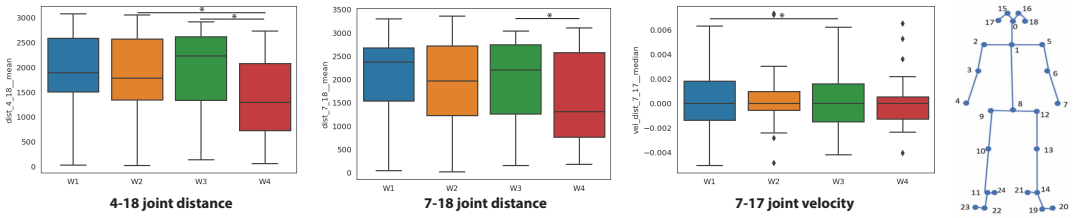

Fig. 5. Hand-over-face gesture feature mean differences across four weeks (W1, W2, W3, and W4). $^{*}p < .05$, $^{**}p < .01$

TABLE III

POST-HOC ANALYSIS RESULTS FOR THE AUDIO FEATURES THAT SHOWED STASTICALLY SIGNIFICANT DIFFERENCES, WHERE M CORRESPENDS TO THE MEAN, AND SD TO THE STANDARD DEVIATION.

Variable	Pair-wise (a, b)	*Z*	*p*	Mdn (a)	Mdn (b)
Speaking length	(W1, W2)	1	<.0001	42.27	86.72
Speaking length	(W1, W3)	8	<.0001	-	111.44
Speaking length	(W1, W4)	14	<.0001	-	106.14
M Speaking length	(W1, W2)	34	<.001	2.43	3.94
M Speaking length	(W1, W3)	52	<.01	-	4.36
M Speaking length	(W1, W4)	40	<.01	-	5.57
SD Speaking length	(W1, W4)	39	<.01	2.45	5.18
Silence length	(W1, W3)	23	<.001	186.91	137.76
Silence length	(W2, W4)	23	<.01	155.13	199.53
Silence length	(W3, W4)	0	<.0001	137.76	-
M Silence length	(W1, W3)	3	<.0001	5.10	3.72
M Silence length	(W2, W3)	20	<.0001	5.23	-
M Silence length	(W2, W4)	48	<.01	-	5.77
M Silence length	(W3, W4)	0	<.0001	3.72	-
SD Silence length	(W1, W2)	56	<.05	6.60	9.80
SD Silence length	(W1, W4)	15	<.0001	-	11.52
SD Silence length	(W2, W4)	24	<.01	9.80	-
SD Silence length	(W3, W4)	24	<.0001	6.11	-
SD Silence length	(W2, W3)	36	<.01	9.80	6.11
M Loudness	(W1, W2)	0	<.0001	0.04	0.07
M Loudness	(W1, W3)	12	<.0001	-	0.08
M Loudness	(W2, W4)	40	<.01	0.07	0.05
M Loudness	(W3, W4)	23	<.001	0.08	-
SD Loudness	(W1, W2)	0	<.0001	3.37	2.16
SD Loudness	(W1, W3)	20	<.0001	-	1.83
SD Loudness	(W1, W4)	35	<.001	-	2.07

($\chi^2 = 8.38$, $p < .05$), for the mean of the distance between keypoints 7 and 18 ($\chi^2 = 8.43$, $p < .05$), and for the median of the distance between keypoints 7 and 18 ($\chi^2 = 8.71$, $p < .05$). Again, we ran a pair-wise comparison using Wilcoxon signed-rank tests with a Bonferroni correction ($.05/4$).

The post-hoc analysis results are reported in Table IV. In summary, coachees displayed significantly more hand-over-face gestures in W4 with respect to the previous weeks.

TABLE IV

POST-HOC ANALYSIS RESULTS FOR THE BODY FEATURES THAT SHOWED STASTICALLY SIGNIFICANT DIFFERENCES, WHERE MD REFERS TO THE MEAN DISTANCE AND MDND REFERS TO MEDIAN DISTANCE.

Variable	Pair-wise (a, b)	*Z*	*p*	Mdn (a)	Mdn (b)
MD 4-18	(W2, W4)	55	<.05	2401.29	1283.92
MD 4-18	(W3, W4)	55	<.05	2448.99	-
MdnD 4-18	(W2, W4)	42	<.01	2556.93	1180.23
MdnD 4-18	(W3, W4)	44	<.05	2518.85	-
MD 7-18	(W3, W4)	48	<.05	1449.41	2579.54
MdnD 7-18	(W2, W4)	56	<.05	2667.50	1220.21

V. DISCUSSION

Our results showed that occurrences of Interaction Ruptures (IRI) throughout the study decreased significantly from week 1 to week 4. IRIs consisted of instances of User Awkwardness (UAI) and Robot Mistakes (RMI), which were evaluated from coachee videos by a human annotator (one of the research team members). Both UAI and RMI decreased throughout the four weeks.

The computationally extracted behavioural cues of the coachees reflect this trend. Through facial cues, coachees express less suspicion (as expressed through the lid tightener AU7, and the upper lip raiser AU10), and less embarrassment-related laughter cues (as expressed through the cheek raiser AU6 and lip corner puller AU12 [51]). While laughter can be an expression of enjoyment [51], in our observations we noticed that laughter was more often an expression of being uncomfortable. Additionally, we found that coachees spoke for greater lengths of time toward week 4, and there were less instances of silence. This indicates that coachees may have become more comfortable and confident in speaking to the robotic well-being coach.

Regarding body cues, we found that coachees showed more self-touch behaviours in week 4, specifically touching their face and displaying hand-over-face gestures. Such behaviours have been found to be more present when

Fig. 6. Coachees' behavioural signals: a) Suspicion expressed through lid tightener AU7, b) Laughter expressed through cheek raiser AU6 and lip corner puller AU12, c) Contemplation - where awkwardness behaviour markers are not present, d) Suspicion expressed through lid tightener AU7 and upper lip raiser AU10, e) Laughter expressed through cheek raiser AU6 and lip corner puller AU16, f) Contemplation expressed through hand-over-face gesture.

coachees self-disclose during coaching with a robot [27]. Our qualitative results [9] indicated that coachees found the exercise in week 4 (optimism about the future) to be the most challenging. The hand-over-face gestures in this week may be an indicator of coachees thinking more reflectively, and self-disclosing more. This is supported by the increase in silence periods in their speech in week 4 (which was otherwise decreasing from W1 to W2 and W3), as well as the decrease in loudness in their speech (which was otherwise increasing throughout the weeks).

This behaviour analysis indicates that overall, coachees got used to the positive psychology coaching exercises with the robot, they became more comfortable and confident and expressed less awkwardness over time. The trend observed for these behaviours was consistent over time, with the exception of week 4, where the more challenging exercise was introduced. This has implications for the order of exercises to be conducted with a robotic coach. When a robotic well-being coach is deployed longitudinally, it could be useful to introduce less challenging exercises first, in order to help coachees get used to the robot. More challenging exercises (such as optimism about the future) could be introduced at a later stage, as users would be already used to the robot and can focus more on the reflection called for by the exercise.

According to our behaviour observations from the videos (see Figure 6), in general, coachees expressed less awkwardness over time and expressed their thoughts in more detail. For example, C2 expressed their emotions openly through a wide range of facial expressions and body gestures. This coachee's communication with the robot improved throughout the sessions since they expressed less awkwardness and shared their thoughts in more detail. The coachee appeared to get used to the robot, in particular its slightly delayed response timing.

There were some notable exceptions where coachees' behaviour differed from the general trend. According to our observations, there were instances where the robot's mistakes caused coachees to limit their self-disclosure. For example, C20 did not engage with the robot in weeks 3 and 4 and did not answer its questions. Additionally, C3 appeared to become more reluctant to self-disclose after the robot's interruptions. This reduction in self-disclosures persisted throughout the four weeks. In contrast, C13 restricted self-disclosure during the first week, but shared their thoughts in greater detail over the following three weeks. In particular, they appeared to be most thoughtful in week four, during the optimism exercise, which was noted to be the most challenging. These differences across coachees show that different coachees express different levels of awkwardness and discomfort, which the future robotic coaches should be able to recognize and adapt to. Note that, when interpreting these results one needs to bear in mind the fact that the videos were coded by a single annotator. Due to this, we are unable to report inter-observer agreement levels.

VI. CONCLUSIONS

This study examined the behavioural responses of coachees to interaction ruptures when interacting with a robotic well-being coach. We found that throughout the four weeks of interacting with the robotic coach, interaction ruptures decreased. Throughout the study, coachees displayed a lower number of behavioural cues related to awkwardness and a higher number of cues related to self-disclosure. Coachees displayed less facial expressions related to suspicion (e.g., lid tightener (AU7) and upper lip raiser (AU10)), increased the duration and the loudness of their speech, and displayed more hand-over-face gestures related to self-disclosure. The findings of this work will inform the development of AI models for the multi-modal detection of interaction ruptures in well-being coaching, a research topic that has not been examined to date. Interaction ruptures can be detrimental to the effectiveness of coaching. Therefore, future work should focus on not only detection but also the design of repair strategies for the occurrence of such ruptures.

ACKNOWLEDGMENT

We thank Cambridge Consultants Inc. and its employees for participating in this study. **Funding:** M. Spitale and H. Gunes are supported by the EPSRC/UKRI under grant ref. EP/R030782/1 (ARoEQ) and EP/R511675/1. M. Axelsson is funded by the Osk. Huttunen foundation and the EPSRC under grant EP/T517847/1. **Open Access:** For open access purposes, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. **Data access:** Raw data related to this publication cannot be openly released due to anonymity and privacy issues.

REFERENCES

[1] C. Baker, "Mental health statistics for england: prevalence, services and funding," 2020.

[2] R. C. O'Connor, K. Wetherall, S. Cleare, H. McClelland, A. J. Melson, C. L. Niedzwiedz, R. E. O'Carroll, D. B. O'Connor, S. Platt, E. Scowcroft, *et al.*, "Mental health and well-being during the covid-19 pandemic: longitudinal analyses of adults in the uk covid-19 mental health & wellbeing study," *BJPsych*, vol. 218, no. 6, pp. 326–333, 2021.

[3] W. H. Organization. (2023) Mental health at work. [Online]. Available: https://www.who.int/health-topics/mental-health

[4] M. Spitale and H. Gunes, "Affective robotics for wellbeing: A scoping review," in *IEEE ACII-W 2022*. IEEE, 2022.

[5] N. I. Abbasi, M. Spitale, J. Anderson, T. Ford, P. B. Jones, and H. Gunes, "Can robots help in the evaluation of mental wellbeing in children? an empirical study," in *IEEE RO-MAN 2022*. IEEE, 2022, pp. 1459–1466.

[6] S. Jeong, S. Alghowinem, L. Aymerich-Franch, K. Arias, A. Lapedriza, R. Picard, H. W. Park, and C. Breazeal, "A robotic positive psychology coach to improve college students' wellbeing," in *IEEE RO-MAN 2020*. IEEE, 2020, pp. 187–194.

[7] E. De Haan and J. Gannon, "The coaching relationship," *The SAGE handbook of coaching*, pp. 195–217, 2017.

[8] J. D. Safran and J. C. Muran, "The resolution of ruptures in the therapeutic alliance." *J. Consult. Psychol.*, vol. 64, no. 3, p. 447, 1996.

[9] M. Spitale, M. Axelsson, and H. Gunes, "Robotic mental well-being coaches for the workplace: An in-the-wild study on form," in *ACM/IEEE HRI 2023*. IEEE, 2023.

[10] M. Axelsson, M. Spitale, and H. Gunes, "Robotic coaches delivering group mindfulness practice at a public cafe," in *Companion ACM/IEEE HRI 2023*, 2023.

[11] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Frontiers in psychology*, vol. 9, p. 861, 2018.

[12] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *ACM/IEEE HRI 2013*. IEEE, 2013, pp. 251–258.

[13] S. Serholt and W. Barendregt, "Robots tutoring children: Longitudinal evaluation of social engagement in child-robot interaction," in *NordiCHI 2016*, 2016, pp. 1–10.

[14] I. P. Bodala, N. Churamani, and H. Gunes, "Teleoperated robot coaching for mindfulness training: A longitudinal study," in *IEEE RO-MAN 2021*. IEEE, 2021, pp. 939–944.

[15] V. Hart, J. Blattner, and S. Leipsic, "Coaching versus therapy: A perspective." *Consult. Psychol. J.*, vol. 53, no. 4, p. 229, 2001.

[16] M. E. Seligman, "Coaching and positive psychology," *Australian Psychologist*, vol. 42, no. 4, pp. 266–267, 2007.

[17] L. L. D'raven and N. Pasha-Zaidi, "Positive psychology interventions: A review for counselling practitioners," *Can. J. Couns.*, vol. 48, no. 4, 2014.

[18] A. Carter, A. Blackman, B. Hicks, M. Williams, and R. Hay, "Perspectives on effective coaching by those who have been coached," *Int. J. Train. Dev.*, vol. 21, no. 2, pp. 73–91, 2017.

[19] D. Keltner and B. N. Buswell, "Embarrassment: its distinct form and appeasement functions." *Psychological bulletin*, vol. 122, no. 3, p. 250, 1997.

[20] R. J. Edelmann, "Social embarrassment: An analysis of the process," *J. Soc. Pers. Relatsh.*, vol. 2, no. 2, pp. 195–213, 1985.

[21] M. Axelsson, M. Spitale, and H. Gunes, "Adaptive robotic mental well-being coaches," in *Companion ACM/IEEE HRI 2023*, 2023.

[22] M. Axelsson, I. P. Bodala, and H. Gunes, "Participatory design of a robotic mental well-being coach," in *IEEE RO-MAN 2021)*. IEEE, 2021, pp. 1081–1088.

[23] M. Axelsson, M. Spitale, and H. Gunes, "Robots as mental well-being coaches: Design and ethical recommendations," *arXiv preprint arXiv:2208.14874*, 2022.

[24] N. Churamani, M. Axelsson, A. Caldir, and H. Gunes, "Continual learning for affective robotics: A proof of concept for wellbeing," in *IEEE ACII-W 2022*. IEEE, 2022.

[25] M. Axelsson, N. Churamani, A. Caldir, and H. Gunes, "Participant perceptions of a robotic coach conducting positive psychology exercises: A systematic analysis," *arXiv preprint arXiv:2209.03827*, 2022.

[26] S. Jeong, L. Aymerich-Franch, K. Arias, S. Alghowinem, A. Lapedriza, R. Picard, H. W. Park, and C. Breazeal, "Deploying a robotic positive psychology coach to improve college students' psychological well-being," *UMUAI*, pp. 1–45, 2022.

[27] S. Alghowinem, S. Jeong, K. Arias, R. Picard, C. Breazeal, and H. W. Park, "Beyond the words: Analysis and detection of self-disclosure behavior during robot positive psychology interaction," in *IEEE FG 2021*. IEEE, 2021, pp. 01–08.

[28] M. Alimardani, L. Kemmeren, K. Okumura, and K. Hiraki, "Robot-assisted mindfulness practice: Analysis of neurophysiological responses and affective state change," in *IEEE RO-MAN 2020*. IEEE, 2020, pp. 683–689.

[29] S. Yoon, M. Alimardani, and K. Hiraki, "The effect of robot-guided meditation on intra-brain eeg phase synchronization," in *Companion ACM/IEEE HRI 2021*, 2021, pp. 318–322.

[30] K. Matheus, M. Vázquez, and B. Scassellati, "A social robot for anxiety reduction via deep breathing," in *IEEE RO-MAN 2022*. IEEE, 2022, pp. 89–94.

[31] N. I. Abbasi, M. Spitale, J. Anderson, T. Ford, P. B. Jones, and H. Gunes, "Computational audio modelling for robot-assisted assessment of children's mental wellbeing," in *ICSR 2022*. Springer, 2023, pp. 23–35.

[32] I. P. Bodala, N. Churamani, and H. Gunes, "Creating a robot coach for mindfulness and wellbeing: A longitudinal study," *arXiv preprint arXiv:2006.05289*, 2020.

[33] L. Tian and S. Oviatt, "A taxonomy of social errors in human-robot interaction," *ACM THRI*, vol. 10, no. 2, pp. 1–32, 2021.

[34] A. L. Baker, E. K. Phillips, D. Ullman, and J. R. Keebler, "Toward an understanding of trust repair in human-robot interaction: Current research and future directions," *ACM TiiS*, vol. 8, no. 4, pp. 1–30, 2018.

[35] S. S. Sebo, P. Krishnamurthi, and B. Scassellati, ""i don't believe you": Investigating the effects of robot trust violation and repair," in *ACM/IEEE HRI 2019*. IEEE, 2019, pp. 57–65.

[36] C. Esterwood and L. P. Robert, "Do you still trust me? human-robot trust repair strategies," in *IEEE RO-MAN 2021*. IEEE, 2021, pp. 183–188.

[37] C. Esterwood and L. P. Robert Jr, "Three strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness," *Computers in Human Behavior*, vol. 142, p. 107658, 2023.

[38] D. Kontogiorgos, A. Pereira, B. Sahindal, S. van Waveren, and J. Gustafson, "Behavioural responses to robot conversational failures," in *ACM/IEEE HRI 2020*, 2020, pp. 53–62.

[39] D. Kontogiorgos, M. Tran, J. Gustafson, and M. Soleymani, "A systematic cross-corpus analysis of human reactions to robot conversational failures," in *ACM ICMI 2021*, 2021, pp. 112–120.

[40] D. Kontogiorgos, S. Van Waveren, O. Wallberg, A. Pereira, I. Leite, and J. Gustafson, "Embodiment effects in interactions with failing robots," in *ACM CHI 2020*, 2020, pp. 1–14.

[41] H. Lindgren, "Younger and older adults' perceptions on role, behavior, goal and recovery strategies for managing breakdown situations in human-robot dialogues," in *HAI 2021*, 2021, pp. 433–437.

[42] M. Giuliani, N. Mirnig, G. Stollnberger, S. Stadler, R. Buchner, and M. Tscheligi, "Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations," *Frontiers in psychology*, vol. 6, p. 931, 2015.

[43] J. L. Smith and A. A. Hanni, "Effects of a savoring intervention on resilience and well-being of older adults," *J Appl Gerontol*, vol. 38, no. 1, pp. 137–152, 2019.

[44] T. Gregersen, P. D. MacIntyre, K. H. Finegan, K. Talbot, and S. Claman, "Examining emotional intelligence within the context of positive psychology interventions," 2014.

[45] L. B. Shapira and M. Mongrain, "The benefits of self-compassion and optimism exercises for individuals vulnerable to depression," *J. Posit. Psychol.*, vol. 5, no. 5, pp. 377–389, 2010.

[46] J. Novikova, L. Watts, and T. Inamura, "Emotionally expressive robot behavior improves human-robot collaboration," in *IEEE RO-MAN 2015*. IEEE, 2015, pp. 7–12.

[47] M. Mahmoud and P. Robinson, "Interpreting hand-over-face gestures," in *IEEE ACII 2011*. Springer, 2011, pp. 248–255.

[48] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *IEEE FG 2018*, 2018, pp. 59–66.

[49] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *ACM ICMI 2010*, 2010, pp. 1459–1462.

[50] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[51] M. Pantic and L. J. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image Vis. Comput.*, vol. 18, no. 11, pp. 881–905, 2000.