On-the-Fly Detection of User Engagement Decrease in Spontaneous Human–Robot Interaction Using Recurrent and Deep Neural Networks

Atef Ben-Youssef, Giovanna Varni, Slim Essid & Chloé Clavel

International Journal of Social Robotics

ISSN 1875-4791

Int J of Soc Robotics
DOI 10.1007/s12369-019-00591-2

Springer

International Journal of Social Robotics
https://doi.org/10.1007/s12369-019-00591-2

On-the-Fly Detection of User Engagement Decrease in Spontaneous Human–Robot Interaction Using Recurrent and Deep Neural Networks

Atef Ben-Youssef[1] · **Giovanna Varni**[1] · **Slim Essid**[1] · **Chloé Clavel**[1]

Accepted: 11 September 2019

Abstract
In this paper we consider the detection of a decrease of engagement by users spontaneously interacting with a socially assistive robot in a public space. We first describe the UE-HRI dataset that collects spontaneous human–robot interactions following the guidelines provided by the affective computing research community to collect data "in-the-wild". We then analyze the users' behaviors, focusing on proxemics, gaze, head motion, facial expressions and speech during interactions with the robot. Finally, we investigate the use of deep leaning techniques (recurrent and deep neural networks) to detect user engagement decrease in real-time. The results of this work highlight, in particular, the relevance of taking into account the temporal dynamics of a user's behavior. Allowing 1–2 s as buffer delay improves the performance of taking a decision on user engagement.

Keywords User engagement decrease · Socially assistive robot · HRI in public space · Real-time detection

1 Introduction

Socially assistive robots (SAR) should be able to communicate and cooperate with humans in order to provide assistance, coaching, companionship, support for convalescence, rehabilitation, learning, or therapeutic aid, etc. (e.g. [23,50]). SAR deployed in public spaces have considerable potential for providing the humans with whom they engage, with a multitude of services: welcoming them, giving them recommendations or interacting in a personalized way [9,10,24,46]. These types of robot employ short-term adaptation in order to keep the user's attention and achieve their goal of assisting them through social interaction. They are equipped with sensors combined with software modules to track humans and inform the interaction process. These modules can for instance track faces, recognize speech, and synthesize speech synchronized with animation. Extracting basic information such as facial expressions, gaze, and head motions allows the robots to better understand the person. Processing this information serves more sophisticated modules that analyze emotions, mood, affective state, and user's engagement in order to give appropriate responses.

This study focuses on real-time detection of user's engagement decrease during a social interaction with a robot in a public space. In public space settings, it is not easy for the robot to achieve its goal in spontaneous social interaction, where participants are free to treat the robot as they like and leave the interaction when they wish [5]. Recognizing user's engagement state represents a key issue in socially assistive robotics.

For this study, we recorded a multimodal dataset of spontaneous interactions with the humanoid robot Pepper[1] [6]. In keeping with the current emerging trend in Affective Computing, this dataset consists of data collected "in-the-wild" [47]. It comprises 278 interactions where the users were free to participate in the interaction if they wished to and free to leave it when they wanted to, and where they were left to behave without unconstraints. Multimodal information describing the user's behavior (i.e. distance to the robot, gaze and head motion as well as facial expressions and speech features) was thus synchronously recorded. We analyze the

✉ Atef Ben-Youssef
atef.benyoussef@telecom-paristech.fr

Giovanna Varni
giovanna.varni@telecom-paristech.fr

Slim Essid
slim.essid@telecom-paristech.fr

Chloé Clavel
chloe.clavel@telecom-paristech.fr

[1] LTCI, Télécom Paris, Institut polytechnique de Paris, 75013 Paris, France

[1] https://www.softbankrobotics.com/emea/en/pepper.

Published online: 28 September 2019

dataset, focusing on the non-verbal behavior displayed by the users. We then make use of data-driven methods for detecting engagement decrease. Such methods rely on a ground-truth obtained by manual annotation of the engagement. Perceived engagement can be a subjective observation. For this reason, each interaction was annotated independently by two annotators: a researcher who knew the purpose of the work and an uninformed one who did not.

This paper is organized as follows. Section 2 presents the related work on user engagement in human–robot interaction (HRI). Section 3 describes the dataset of spontaneous HRI. Section 4 focuses on the analysis of user engagement decrease. Section 5 describes our approach to detect the decrease of user engagement. A discussion is presented in Sect. 6. Finally, conclusions are drawn in Sect. 7.

2 Related Work

2.1 Socially Assistive Robots in Public Space

SAR are robots providing assistance to human users through social interaction [50]. These robots are designed e assist users by creating effective interactions [23]. SAR deployed in real world settings need to secure and maintain the user's engagement. Pitsch et al. [42] analyzed interactions between a robot deployed as a guide in a museum, and visitors. They found that the first 5 s of the interaction had a relevant impact on the user's engagement during the interaction (e.g. leaving/staying, responsiveness, exchanging rituals). Gehle et al. [26] likewise analyzed the interaction opening strategies of a robot playing the role of a museum's guide, in its interaction with visitors. Hayashi et al. [29] proposed to use robots in train stations to assist passengers. Their goal was to identify the best way to provide users with travel information. They compared the use of one versus two robots. The findings of this study showed that the most effective way of attracting people's interest was by presenting information using two humanoid robots rather than one. They reported also that the interactivity was useful in giving the feeling of talking with robots. Another interesting scenario is the use of SAR to provide shopping information to customers [24,34]. The MuMMER project aims to develop a socially intelligent humanoid robot that is able to operate in a public shopping mall [24]. In [34], SAR were designed to naturally interact with customers and to provide shopping information. In public spaces, SAR could inspire the design of hotel-assistive robots [23].

In such application contexts, robots are expected to respond appropriately to the users' behavior and engage them in stimulating experiences [18,21]. In particular, they should be able to monitor a user's state of engagement in order to be able to react to possible signs of disengagement in such a way as to maintain their interest. In real world settings, one of the challenges is to deal with the dynamic and flexible nature of human behavior in order to secure and maintain users' engagement in their interaction with SAR.

Tackling these challenges, our research aims to detect user engagement states in real-time in order to assist humans for the purpose of providing such public services. The proposed detection model integrates data on the temporal dynamics of engagement behavior, with the multimodal data collected *in-the-wild*.

2.2 Engagement and Disengagement in HRI

The engagement was defined in human–computer interaction by Sidner [49] as "the process by which individuals in an interaction start, maintain and end their perceived connection to one another"; and by Poggi [43] as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction". This concept of engagement has been explored from different perspectives with regard to humans interacting with social robots or virtual agents [19]. More specifically, a focus has been put on user engagement prediction [14,25], the analysis of the emergence of engagement [42,52], the identification of the addressees to interact with [37], and the study of the relationship between personality and engagement [15,32] and engagement perception [28]. Similarly, disengagement has been tackled in many studies by analyzing interaction problems, the time of their occurrence and their causes [1,51], the dynamics of affective states [12,20] and the prediction of disengagement [11,36]. The most crucial causes of interaction problems are found to be the limitations of the systems used to detect social signals and of the interaction models. For example, it was reported that the most frequent causes were the engagement model, face tracker, turn-taking model, or speech recognition issues, misunderstanding, lack of adaptation, repetition and long pauses, over-fragmentation, over-clarity, over-coordination, over-directedness, insufficient or exaggerated state-of-mind updates and repair requests [1,11,39].

Humans behave differently during social norm violations and technical errors in HRI [51]. It was shown that the automatic detection of these errors based on human behavior works to some extent. The performance of error detection is better when the robot knows the human with whom it is interacting . Detecting social norm violation is harder than detecting technical failures. We conclude from the work of Trung et al. [51] that detecting disengagement in social interaction with a robot is difficult.

The most common features used in these studies to asses engagement and disengagement were, among others, gaze [2,32,36,45,49], head motion [2,36], face [11,36,38], posture [2,36], speech [32,36], and distance [52]. Other,

more subtle, features were also included: semantics, attention, emotions and affects [11,12,20,36]. In a previous study, we show that the use of combined multimodal features effectively improves the performance of a user engagement breakdown system [5]. Combining features from two or more modalities allows one to achieve better results in engagement detection/prediction, compared to the use of features from only one modality. Kendon [35] analyzed gaze and speech. He found that the speakers looks at each other during fluent speech and at the end of sentences, but look away during hesitations or unfluent speech. This type of social signal is probably relevant information to evaluate the engagement level during the interaction. Prosody, articulation, voice-quality related features, linguistic analysis as well as facial expressions and gaze were used to detect interest in [48].

To model user engagement in HRI, researchers have considered a subset of systems going from rule-based to machine-learning-based. Machine learning approaches have been compared to rule-based approaches in [25]. It has been shown that the rule-based classifiers have a competitive performance compared to the set of supervised classifiers trained on a small labeled corpus. The authors found that conditional random fields (CRF), which give an accuracy of 61.5% and F1-score of 0.61, is a much more stable classifier than others. Machine learning approaches are the most commonly used for automatically predicting engagement in HRI. By comparing logistic regression and boosted decision tree models in [11], the logistic regression model was selected for managing disengagement decisions. In [8], Bohus et al. used a frame-by-frame binary classification scheme using a maximum entropy model to predict engagement intentions. Leave-one-out cross-validation using support vector machines (SVMs) was used in [14,36,48]. SVMs with a polynomial kernel were successfully used to recognize the interest in [48]. The problem to address the engagement of only one user or more than one in interaction was studied by Leite et al. [36]. They found that the disengagement model trained in the single-user condition might not be appropriate for the group condition, but the group condition model generalizes better to the single-user condition. A mixed model combining both conditions is a good compromise, but it does not achieve the performance levels of the models trained for a specific type of interaction. Their best models give an accuracy of 63% and AUC of 0.61 for the single-user condition and an accuracy of 73% and AUC of 0.62 for the group condition. This finding has encouraged us to work with mixed conditions. Liu et al. applied the Echo State Networks (ESNs) architecture, a variant of Recurrent Neural Networks, to a real-world dataset and showed that these networks are able to predict engagement breakdown behavior using 30 s of facial expression features [38].

Our positioning in relation to these previous studies is as follows. First, our collected dataset targets the diverse social signals that are involved in user engagement, considering a wide range of heterogeneous sensors: microphone array, cameras, sonars, lasers, along with user tracked variables (i.e. face features, head angles, eye gaze and position toward the robot). To the best of our knowledge, none of the existing datasets provide such a thorough coverage of signals amenable to exploitation for user engagement analysis. This is also the first significant dataset offering a large amount of data collected by the robot "Pepper". Pepper offers a large combination of features compared to the other robots used in the literature (NAO, iCub, MyKeepon, and so on). Second, the use of such a large and real-world dataset allows us to investigate deep learning approaches such as recurrent neural networks for the multimodal detection of user engagement decrease. This "into-the wild" dataset is here used to model the temporal user behavior in order to make decision in real-time about engagement decrease. It follows the work of : (i) [38] that uses such neural networks with facial expression alone on a reduced set of our dataset that has already been made public; (ii) [5] that shows the superiority of multimodality for a related but close task which is the prediction of engagement breakdown using task-designed logistic regression. This could lead to the development of lifelike humanoid SAR that could better understand the behavior of the humans they are interacting with, and therefore respond more appropriately in order to increase their engagement.

3 Spontaneous Human–Humanoid Interactions

3.1 Experimental Design

The experiments were conducted in a public space at Telecom ParisTech over 17 months. The recordings consisted of interactions between humans and the robot Pepper (see Fig. 1). The collected data constitute the UE-HRI dataset[2] described in [6]. It includes all data streams available on Pepper, packaged in the open-source Robot Operating System (ROS) framework.[3] Each stream is translated into a message (called ROS topic) and packaged together into a ROSbag file. In order to keep all the streams synchronized, they were indexed using the robot timestamps. The recorded data is split into ROSbag files of 100 Mb in order to quickly move them from the robot to a storage server over Ethernet. ROSbag files are then merged together into one ROSbag file in order to get one file per interaction. Figure 2 shows the experimental setup of the interaction.

[2] https://www.tsi.telecom-paristech.fr/aao/en/2017/05/18/ue-hri-dataset/.

[3] http://wiki.ros.org/naoqi_driver.

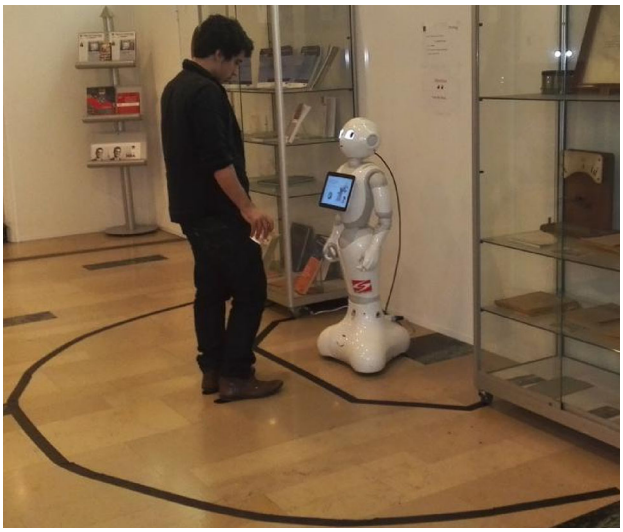


Fig. 1 Participant in the first engagement zone (less than 1.5 m from the robot) interacting with Pepper

Pepper automatically starts the interaction when it detects movement, and focuses on the participant in front of it, who is in the interaction zone (i.e. a distance of less than 1.5-m from the robot, indicated by means of black tape stuck on the floor in Fig. 1).

First, the robot asks the user to sign the *agreement* form displayed on its embedded tablet, authorizing researchers to use her collected data for further analysis. After validation of the agreement, the robot enters the *welcome* phase by introducing itself through very lively animations and providing the user with the following instructions: "speak loud and be alone in the 1st engagement zone". It then enters the *dialog* phase. This includes a set of open-ended questions where the robot asks the participant to introduce herself and to talk about her favorite restaurants and films. The next phase is the *cucumber* phase, when the robot presents its vision technology to the user in a humoristic way by showing that, from its viewpoint, the difference between a cucumber and a human is the face. Finally, the robot enters the *survey* phase, during which the user is asked to assess her satisfaction with the interaction with Pepper, by answering 15 questions on a 5-level Likert scale (from disagree "1" to agree "5") [15,27] (see Appendix A).

3.2 Participants

The recordings involved 278 users (182 males, 96 females) , whose average age was 25 ($\pm$ 9.5) years. This was estimated using an ad hoc software module embedded in Pepper [31]. The users were students, teachers, researchers, visitors and other staff of Telecom ParisTech. A poster on the wall warned users that they were being recorded during the interaction with the robot. The contact information of the main researcher was also made available on the poster. This was done to allow the users to contact the researcher, should they have concerns about the exploitation of their data, and to be able to ask to have it deleted if they so wished. No instructions were given to the user except those provided by the robot in the welcome phase. Users were free to participate in the interaction and free to leave when they wished. The interaction was unsupervised, so the number of users simultaneously involved in it was not controlled. Even though the robot warned that only one user was to be in the first engagement zone at a time, the collected data included 209 interactions featuring a single user, and 69 multiparty interactions (32 started as multiparty and ended as single-user). Note that only 46 users stayed until the end of the scenario and the remaining 72, 84, 70 and 6 users left the interaction at the welcome, dialogue, cucumber and survey phase, respectively.

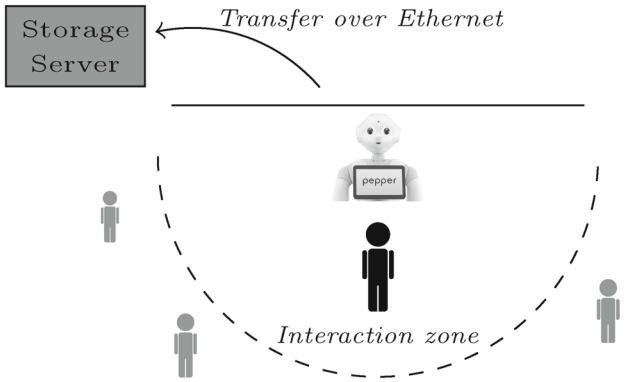


Fig. 2 Technical setup

3.3 Social Signals on the Robot Pepper

Pepper can record a large variety of data streams ranging from raw signals (audio, video, sonar and laser) to face tracking and estimation of gaze direction, head motion and facial expression. In this work, features were extracted by using the available trackers of NAOqi-SDK as they are integrated in the robot.

Distance The distance between the user and the robot was computed using measured raw signals (i.e. sonar) and tracked variables as described below. More specifically, the front sonar[4] (i.e. ultrasonic sensor) was used. The NAOqi People Perception[5] module was also used to extract the distance of the participant's face from the robot camera as well as her 3D head position in relation to the robot's torso reference. The space in front of the robot was divided into three

[4] http://doc.aldebaran.com/2-7/family/pepper_technical/sonar_pep.html.

[5] http://doc.aldebaran.com/2-7/naoqi/peopleperception/.

configurable zones using the ALEngagementZones module. The default configuration was used here. The first engagement zone is the area about 1.5 m away from the robot. In this work, this was used as the interaction zone. The second zone is the area between 1.5 and 2.5 m away. The third zone is the area more than 2.5 m away from the robot. The participant's position was classified to be in one of these three spaces (or 0 if unknown) using the 3D coordinates of the user's head in the robot frame.[6]

Gaze Pepper's ALGazeAnalysis module gives information about the user face orientation in order to detect whether the user is looking at the robot or not. OpenFace 2.0 [4] was used to compute gaze direction in relation to the plane of the face [54].

Head and face OpenFace 2.0 [4] was also used to compute the head pose of the user along the three axes (yaw, pitch, roll). Moreover, it was also used to recognize the occurrence and intensity of each facial Action Unit (AU) [3].

Speech The audio signal was recorded at a sampling frequency of 48 KHz using 4 microphones that are available inside the head of the robot. The audio signal contains the speech of both the participant and the robot as well as noise in the environment. In order to simplify the analysis of the audio, we selected the first channel (i.e. first microphone) to extract speech features.[7] Speech features included: the fundamental frequency (F0) (extracted via an autocorrelation and cepstrum based method), log-energy, loudness contours, voicing probability and the first 12 MFCCs excluding the 0th MFCC. All these features were computed from the audio signal over 50-ms windows at a frame rate of 100 Hz with openSMILE [22]. Features indicating if the robot is speaking or not, as well as the robot's and the user's speech duration, were computed from the dialog (text-to-speech and automatic speech recognition) ROSbag topics.

3.4 Annotation of Engagement

We developed a script that extracts synchronized front and bottom images[8] and audio from the corresponding ROSbag topics and merges them into a video using ffmpeg.[9] Two annotators with different scientific backgrounds annotated the dataset: a researcher who knew the purpose of the work and an uninformed one who did not. The ELAN annotation tool [53] was used to annotate the videos. On all recordings, the annotator indicates the start and the end of the interaction as well as the number of participants (i.e. mono-user or multi-users). In order to characterize engagement, annotators were asked to annotate the interaction video segment by segment based on verbal and non-verbal behaviors expressed by the user that exhibits an engagement decrease, with the following label "Sign of Engagement Decrease (SED)".

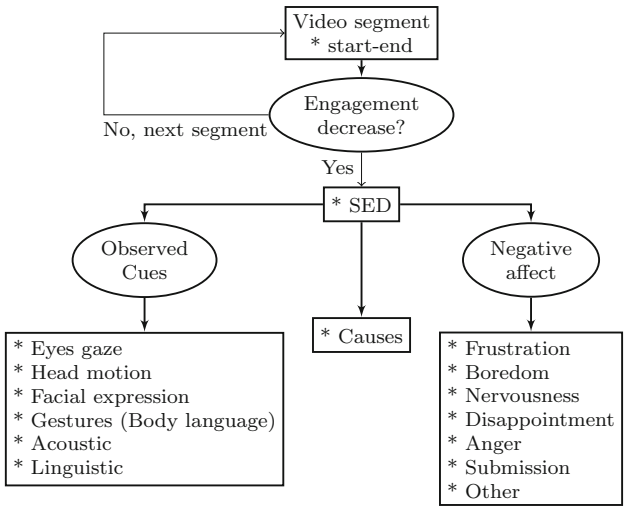


Fig. 3 Flow-chart of the different annotation levels. The '*' shows what the annotator has to select

A sign of engagement decrease (SED) reflects any cue exhibited by the user showing any form of disinterest in the robot. It could occur any time during the interaction. This cue may correspond to verbal or nonverbal behaviors of the participant. SED could represent an early sign of future engagement breakdown, that is, a sign that leaving the interaction will occur in the near future and before the end of the scenario.

Figure 3 shows a flow-chart that summarizes the annotation process described above. A video tutorial was created to explain the annotation process and how to annotate the interaction using ELAN. The annotator defines the start and the end segment as well as the corresponding label, observed cues and negative affect of that segment. For each defined segment, the annotator assigns the corresponding observed cues of that decrease, in order of importance. This part could be sub-segmented. For example, if the participant says:"I'm bored", with a corresponding facial expression, the annotator indicates in the "Cues 1" track: "speech linguistic" and in "Cues 2" track: "face". The annotator decides which one is more visible in the segment to appear in "Cues 1". If these two cues are successive in time, both should appear in "Cues 1" with a sub-segmentation of the start and end of each one. The annotator also assigns the corresponding negative affect of that segment (if relevant) of that decrease. Negative affects (frustration, boredom, nervousness, disappointment, anger, submission) are based on verbal and nonverbal behavior while interacting with Pepper. Annotators are free to add more information concerning this segment. We recommend

[6] http://doc.aldebaran.com/2-7/glossary.html#term-frame-robot.

[7] Beamforming would be a better alternative that will be considered in future work.

[8] http://doc.aldebaran.com/2-7/family/pepper_technical/video_2D_pep.html.

[9] https://www.ffmpeg.org/.

Table 1 Extracted stream feature

Stream	Feature	Description
Distance	Front sonar	1 feature (m)
	Face distance	1 feature (m)
	3D head position	3 features [x, y, z] (in torso frame)
	Engagement zone	1 feature $\in \{0, 1, 2, 3\}$
Gaze	Gaze direction	2 features [yaw, pitch] (in radians)
	Is looking at robot	1 feature $\in \{0, 1\}$
Head	Head angles	3 features [yaw, pitch, roll] (in radian)
Face	Action Unit	17 features $\in [0, 1]$
Speech	Voicing probability	1 feature
	F0	1 feature
	Loudness	1 feature
	Log(energy)	1 feature
	12 MFCCs	12 features
	Is Robot Speaking	1 feature $\in \{0, 1\}$
	Robot speech duration	1 feature (s)
	User speech duration	1 feature (s)

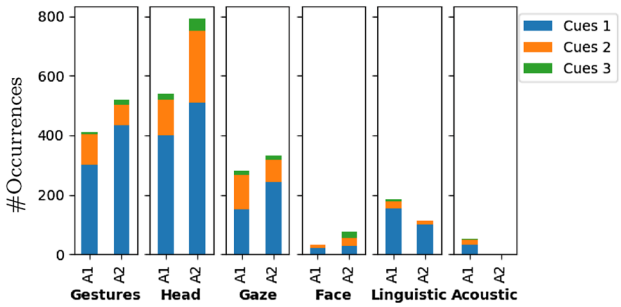
that they add information about the causes in the "Causes" track.

The overall *Cohen kappa* agreement score on annotated recordings for SED annotation is $\kappa = 0.73$ (substantial agreement) (see Fig. 9 in Appendix B). If we automatically correct the annotations by merging together the "engaged" segment located between 2 SED segments and inferior to 1 s in duration, to get 1 large SED segment instead of 2 separated by 1 or 2 frames of "engaged", the Kappa increases slightly to $\kappa = 0.74$ (see Fig. 9).

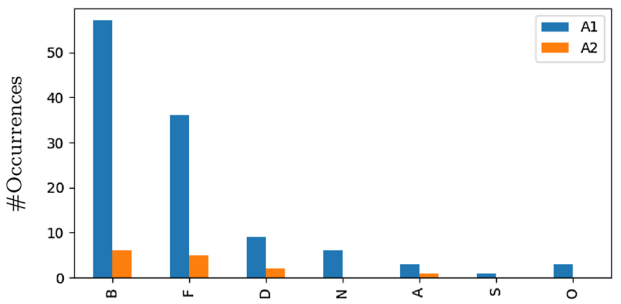
4 User Engagement Decrease Analysis

According to both annotators, the average duration of the interaction is 7 ($\pm$ 5) min. During interactions, users displayed SED in around 6 segments lasting in average 6 ($\pm$ 9) s. Note that for participants who left the interaction by the second phase, the intervals between SED were shorter compared to those who stayed till the end of the interaction. Note also that the last segment where SED were shown is generally longer. In average, its duration is around 9 ($\pm$ 15) s. In 90% of the interaction duration, the users are engaged. For the reaming 10%, the users exhibit SED.

Figure 4 displays the number of occurrences of the behavior exhibited by the users when their engagement decreases, as perceived by the annotators. Figure 4a confirms that the non-verbal behaviors play a special role to point the engagement level. Head motion, gesture (i.e. posture, hand waving, and so on) and eye gaze are the most recurrent features to identify a decrease of engagement in our dataset. Figure 4b shows that annotators disagreed on selecting the appropriate affects related to the SED segments. This showed that the annotation of affects was more subjective here than the annotations of the SED category and their cues.



(a) Cues



(b) Negative affects. The meaning of B: Boredom, F: Frustration, D: Disappointment , N: Nervousness, A: Anger, S: Submission, O: Other

Fig. 4 Cues and affects distribution of signs of engagement decrease (SED) by each annotator. A1: denotes the first annotator 1, A2: denotes the second annotator

Table 2 Engagement decrease causes

Causes of SED	Rate (%)
User interrupted by another person	39
Robot error (long pauses, misunderstood)	17
User uses his phone	10
Robot focus on another person	5
User time constraint	2
User missed robot's request	2

Due to the wide variety of possible factors that can cause engagement decrease in spontaneous interactions, it is difficult to determine the exact cause for each SED segment. However, we asked the annotators to try to mention any information related to the cause of that decrease. Table 2 presents the main causes of the engagement decrease detected by the two annotators, with their percentage of occurrence. We individuated two principal sources that lead to the decrease of engagement: the first is due to a social norm violation (e.g. another person interrupts the interaction while the robot is talking; user time constraint; user is using her phone); the second cause is due to robot's technical issues (e.g. robot makes long pauses or misunderstands the user).

We compared users' behaviors when they were engaged with the robot versus when they showed signs of engagement decrease based on the annotations. Figure 5 presents the results of the comparison for the different configurations: when both annotators perceived the user as being engaged (denoted by "Engagement agreed"), when both annotators agreed about the user engagement decrease (denoted by "SED agreed") and when both annotators disagreed about the engagement state (denoted by "SED: Ax" when a decrease of engagement is perceived by one annotator x and not by the other one). Figure 5a shows the average distance between the user and the robot. The users were closer to the robot when they were fully engaged than when their engagement decreased. Regarding gaze, when users were engaged they looked more at the robot than when their engagement decreased (Fig. 5e "1" when the user looks at the robot, "0" otherwise). This could be confirmed with vertical gaze direction around pitch axis (i.e. angle x) in Fig. 5b. Head motion (i.e. shaking, tilting and nodding) were displayed in Fig. 5c. Users move their head more when their engagement decreases. Concerning action units (AU) [44] (see Fig. 5d), we found that users have the appearance of being happier (where happiness involves AU06 and AU12) when they are engaged, compared to when their engagement decreases. Similarly, for sadness, which is the combination of AU01, AU04, AU15, anger (the combination of AU04, AU05, AU07, AU23) and disgust (the combination of AU09, AU15, AU16), it appears that users express these negative emotions when their engagement decreases, compared to when they are engaged. Figure 5f shows that the users are more engaged when the robot is speaking. This could be confirmed with Fig. 5e (i.e. "1" when the robot is speaking, "0" when the robot is listening).

In the next section, for training and testing of engagement decrease detection, we consider only the segments where both annotators agreed on the engagement category.

5 Detection of User Engagement Decrease

5.1 User Engagement Modeling

We modeled the task of user engagement decrease detection as a binary classification, where the goal is to predict, in real-time, whether the user is engaged or not with the robot, based on the user's behavior analysis.

Our SED detection approach is illustrated in Fig. 6. We define *observation window* as a window of $[t-\tau, t]$, that is, a window that ends at time t and takes into account the last τ seconds of user behavior. We use $[x_{t-\tau}, \ldots, x_{t-1}, x_t]$ as a feature vector computed over the frames of the observation window as input for the classifier. As for the output, each observation window is labeled as either engaged or not.

At running time t, we build a model that classifies the observed behavior over $[t-\tau, t]$ as *user engaged* or *user not engaged*. Let $X = [x_1, x_2, \ldots, x_T]$ denote the sequence of multimodal user-behavior feature vectors and $Y^\eta = [y_1^\eta, y_2^\eta, \ldots, y_T^\eta]$ denote the corresponding sequence of (binary valued) output labels, where η is the duration of the buffer for holding more observations and

$$y_t^\eta = \mathcal{C}([x_{t-\tau}, \ldots, x_{t-1}, x_t]) \; with \; \tau \geq \eta \tag{1}$$

where $\mathcal{C}(.)$ is the classifier decision function and

$$\begin{cases} y_t^\eta = 1, \; SED \; perceived \; at \; time \; t-\eta \\ y_t^\eta = 0, \; otherwise \end{cases}$$

5.2 Deep Networks

In this study, a sequential modeling approach is proposed to detect SED using deep learning techniques [7].

Remembering information for long periods of time is the default behavior of Long-Short Term Memory (LSTM) [30]. LSTM uses a memory unit that can remember information/context from the beginning of the input sequence (i.e. $t-\tau$). Gated Recurrent Unit (GRU) networks [16] are similar to the LSTM, but use a simplified structure. Both LSTM and GRU can be used for modeling temporal sequences. However, GRU involves less computation units than LSTM, since they do not have an output gate. Therefore LSTM are usually preferred if trained on very large datasets (big data).

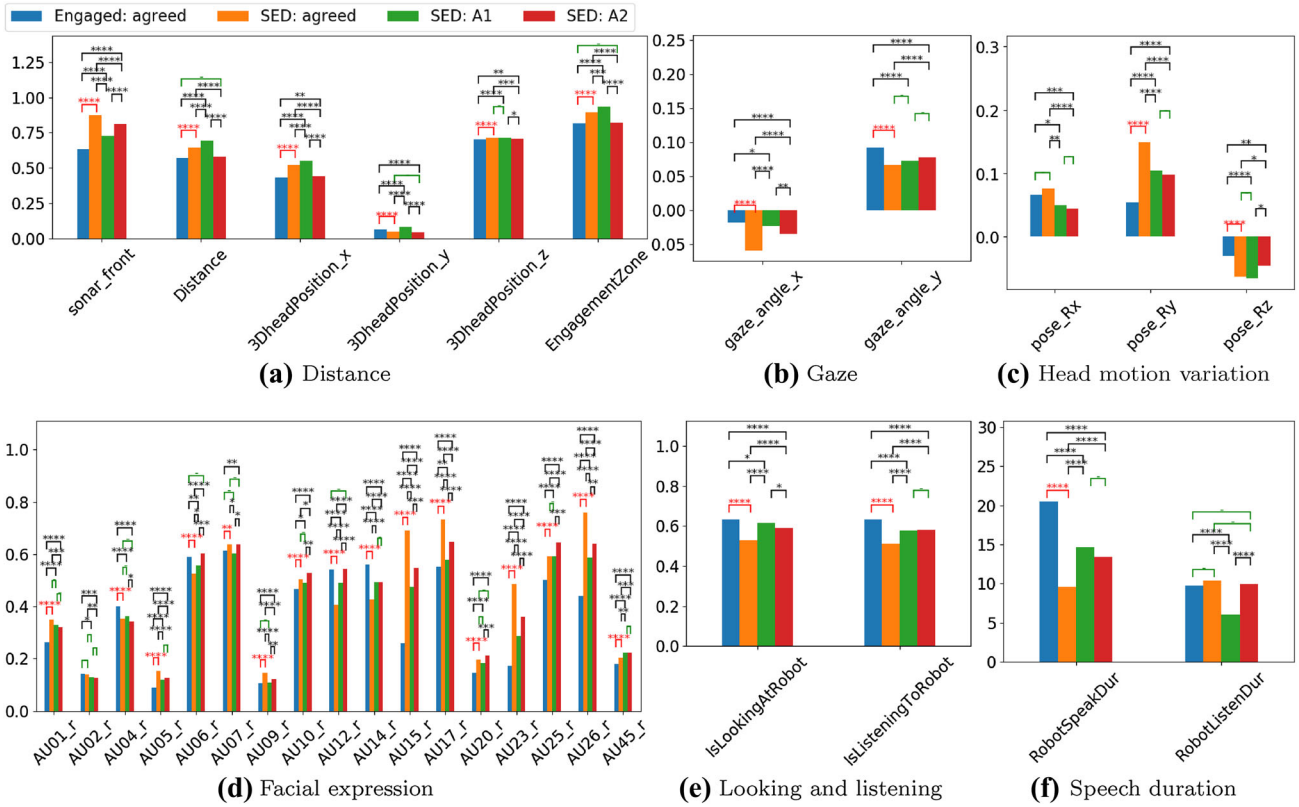


Fig. 5 Selected features of users' behavior when the two annotators (A1 and A2) agreed on their engagement as well as when they disagreed. Paired *t* tests were calculated: red was used when the annotators agreed, ****means $p < 0.0001$, ***means $p < 0.001$, **means $p < 0.01$, *means $p < 0.05$ and green "-" means $p \geq 0.05$. (Color figure online)

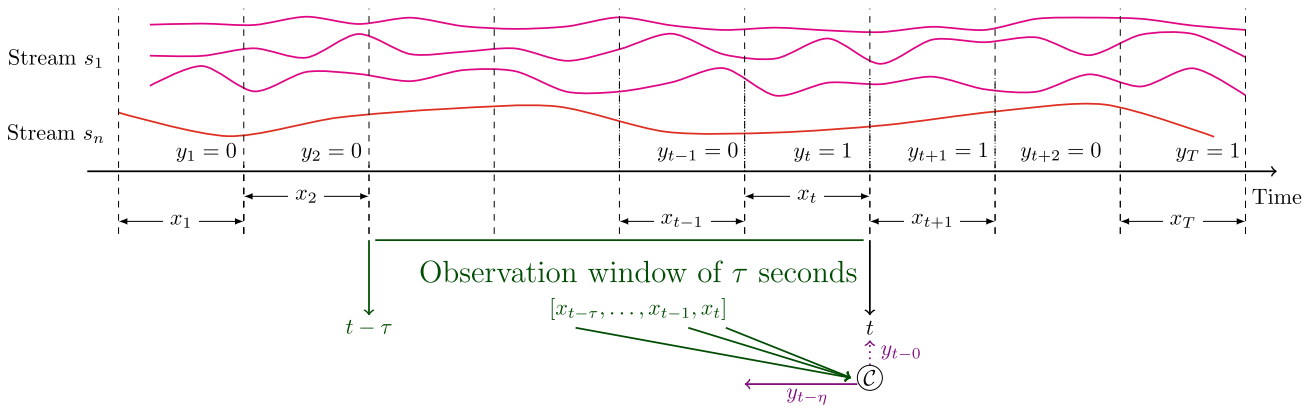


Fig. 6 Illustration of the detection approach. Input: observation window of user behavior is shown in green. Output: buffer duration is shown in violet. (Color figure online)

5.3 Experiments

The data streams of different sampling frequencies were indexed using the robot's timestamps. To obtain synchronized feature vectors, temporal integration [33] (a.k.a temporal pooling), is performed over all feature streams using common *integration windows*. The integrated features are obtained by applying an integration function f over sliding (possibly overlapping) integration windows of length L seconds. The functions f used in this study are statistics, namely the mean and variance. Also, we fix the integration window length L to 500 ms. No overlapping was used. It was shown that combining multiple features gives the highest performance in disengagement prediction (c.f. Sect. 2.2).

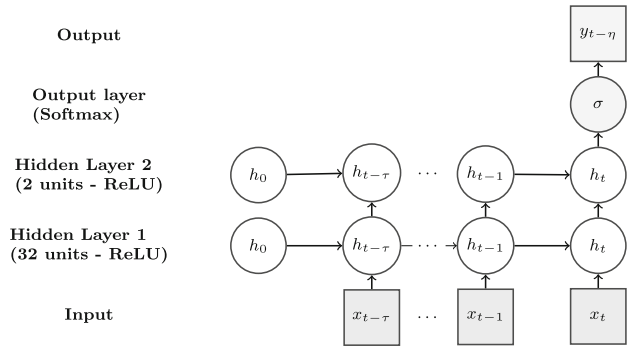


Fig. 7 Many-to-One deep architecture with 2 layers

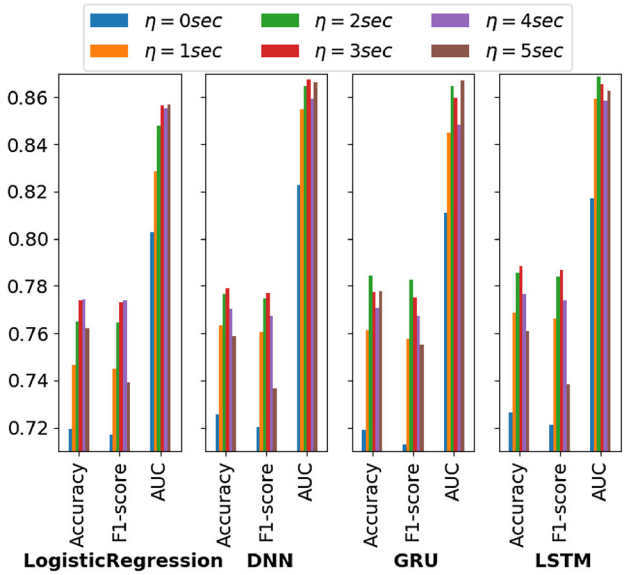


Fig. 8 Performance of a set of classifiers using an observation window of $\tau = 5$ s and a buffer of $\eta \in [0, 5]$ sec, respectively

Therefore, the synchronized texture-window level feature vectors of Distance, Gaze, Head, Face and Speech Streams shown in Table 1 were concatenated together to describe users' behavior and were employed as the input features for the SED detection model. Further details are given in our previous work [5].

Our dataset contains missing values. For example, we have missing values on the face features (i.e. head motion, gaze, AU) when some occlusion occur. This happens for instance when the robot's head is moving, causing the user's face to go out of the cameras' field of view. We chose to replace the missing values by means of the corresponding feature from the training data. We then normalized the data by subtracting the mean value and dividing by the standard deviation of each feature, using the training data.

The whole dataset using both single user and multiparty interactions was used, since this was reported to be a good compromise in [36]. We used threefold cross-validation to train and test a set of SED classifiers. The split of train and test sets was done at the interaction-level. Hence, the users of the test set (i.e. all observations of the user) were not seen during the training phase, which resulted in a user-independent detection model.

We used scikit-learn's [41] implementation of logistic regression as a baseline and Keras's [17] implementation for DNN, GRU and LSTM. We leave the further optimization of the classifiers' hyper-parameters for future work and focus here on the validation of the usefulness of the recurrent network architectures considered.

Following preliminary experiments, we used 2 layers with 32 units followed by 2 units, ReLU activation, dropout with probability of 0.1 and the RMSprop algorithm as optimizer to train the deep networks (see Fig. 7). We used 10% of the training data as a validation set. We trained each model with 100 epochs, using an early stopping callback to stop the training once the validation accuracy started to decrease. In general, the models converge after a maximum of 35 epochs. For logistic regression, we used ℓ_2 regularization and the inverse of the regularization strength C set to 1. To deal with the imbalanced data distribution, the weights for each class were computed and used for training the models.

5.4 Evaluation Criteria

Traditionally, the accuracy rate and F1-score have been the most commonly used evaluation criteria. However, they are not well suited to our study because the dataset is unbalanced. We have around 90% of the data labeled as engaged and only 10% of SED. In case of imbalanced data, the accuracy reflects only the underlying class distribution, not the prediction performance of the minority class. In order to compute meaningful accuracy and F1-score, the test set should represent the true distribution of both classes. Therefore, the test set is resampled to be the average over all the samples of the minority class and the n-differing samples of the majority class selected from the available samples. We also computed the area under the receiver operating characteristic curve (AUC) in order to determine which of the models used predicts the classes best. The AUC corresponds to the probability of correctly identifying the SED class [13]. The closer the AUC comes to 1, the more accurate it is.

5.5 Results

The performance of the different sets of classifiers was compared (see Fig. 8). We found that deep learning techniques are better than conventional machine learning techniques (i.e. Logistic Regression). With the chosen hyper-parameter values the best results were obtained with LSTM for all tested buffer durations η. This is because they better model the tem-

poral dynamics through connections between hidden units in the same layer.

When we use a buffer delay in the range of [1, 3] s, the performance of all the classifiers increases. This could be explained by the fact that using more information about the user's behavior plays an important role in inferring the state of their engagement. A buffer longer than 3 s does not give a better performance. In addition, a buffer of 3 s is already large for real-time detection [40].

To better understand how performance is affected by the size τ of the observation window of the user behavior, we varied it from 0 to 6 s. Table 3 shows this variation for each buffer η. For real-time operation using $\eta = 0$, the best results were found using short observation windows of $\tau = 1$ s for detecting SED. Increasing the buffer duration up to 3 s improves the performance of the SED detector. The best performance was found using an observation window of $\tau = 5$ s for a buffer η of 3 s and at approximately the same performance for a buffer of 2 s. We note that taking a buffer duration to make a decision approximately in the middle of the observation window is the best strategy to detect SED, and the optimal size of the observation window is inferior to the average duration of SED segments (i.e. 6 s). Table 4 shows that logistic regression presents 30% more false alerts than LSTM and 7% fewer undetected engagement decreases.

6 Discussion

In order to develop lifelike humanoid robots that understand better the behavior of the humans with whom they interact and can respond appropriately to increase user engagement, we investigated the use of deep networks to successfully detect SED. We achieved good performance: 78% of accuracy, 0.78 of F1-score and 0.87 of AUC. Note that in other related studies, the performances of engagement detection systems, using different datasets, were 62% accuracy and 0.61 F1-score in [25] and 73% accuracy and 0.62 AUC in [36]. Thus we are using a bigger data-set with different annotation schema. But, we achieve promising results that could be improved and integrated in the robot architecture to detect SED with real-time capability.

The classifiers provide not only the class of user engagement, but also the estimated confidence that could be used as additional information, representing the system's uncertainty, in real-world human–robot applications.

In preliminary experiments using less data (e.g. 195 interactions), the best performances for GRU/LSTM using a buffer of $\eta = 1$ s and an observation widow of $\tau = 2$ s were 76%/75%, 0.75/0.75 and 0.84/0.84 for accuracy, F1-score and AUC, respectively. Thus, the GRU gives a slightly better performance.

Table 3 LSTM performance for different observation windows τ (s) for each buffer η (s) with $\tau \geq \eta$

τ	Accuracy						F1-score						AUC					
	η																	
	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
0	73.42	–	—	–	–	–	0.732	–	–	–	–	–	0.820	–	–	–	–	–
1	**74.03**	76.97	–	–	–	–	**0.738**	0.768	–	–	–	–	**0.827**	0.851	–	–	–	–
2	73.70	76.73	77.26	–	–	–	0.734	0.765	0.771	–	–	–	0.823	0.850	0.849	–	–	–
3	73.74	**77.25**	78.06	78.32	–	–	0.734	**0.770**	0.779	0.782	–	–	0.826	**0.860**	0.863	0.862	–	–
4	72.11	75.60	77.43	77.97	77.02	–	0.716	0.752	0.772	0.777	0.767	–	0.815	0.844	0.860	0.861	0.850	–
5	72.62	76.88	**78.56**	**78.83**	77.65	76.07	0.721	0.766	**0.784**	**0.787**	0.774	0.739	0.817	0.859	**0.869**	**0.865**	0.859	0.863
6	71.52	76.28	74.95	78.23	77.67	75.75	0.708	0.759	0.744	0.780	0.774	0.732	0.818	0.847	0.847	0.865	0.861	0.856

Bold refers to the best results

Table 4 Confusion matrix of LSTM versus logistic regression using an observation window of $\tau = 5$ s and a buffer of $\eta = 2$ s

Classes	LSTM		Logistic Regression	
	Engaged	SED	Engaged	SED
Engaged	17,745	2656	16,479	3922
SED	6072	14,329	5616	14,785

Table 5 Results using LSTM with an observation window $\tau = 5$ s and a buffer of $\eta = 2$ s with gaze direction, head motion, and facial expression extracted using OpenFace versus Pepper's OKAO^{TM} Vision software, combined with speech and distance streams

Tracker	Accuracy	F1-score	AUC
OpenFace [4]	78.56	0.784	0.869
Pepper OKAO software [31]	76.33	0.762	0.849

We evaluate the impact of two different extractors: OpenFace [4], and Pepper's OKAO^{TM} Vision software [31] tracker of gaze direction, head motion and facial expression/AU. Table 5 compares the performance of these extractors on the task of detecting SED using LSTM with an observation window of $\tau = 5$ s and a buffer of $\eta = 2$ s. We found that OpenFace performs better than Pepper's tracker. Note that when features are missing (e.g. when the robot's head is moving and user's facial features cannot be determined), we focus on the other modality (i.e. distance, speech) to detect SED.

In spontaneous HRI, finding the exact moment of SED is a hard decision. It depends on the head motion, the looking away, the spoken word, getting away from the robot, etc. The annotated start and end of this segment is flexible and could vary by $\pm n$ frames (see Fig. 9b in Appendix B). It would be interesting to take into account this flexibility both in the training and in the testing phases instead of using it only when the annotators agree and ignoring the parts where they disagree.

Future work should also investigate whether the SED detection model generalizes well to other interaction settings (i.e. other scenarios, multiparty).

7 Conclusion

We analyzed users' behavior in two engagement states where they exhibited engaged behavior or, alternatively, signs of engagement decrease. We found significant differences in their behavior that allowed us to develop a real-time detector of engagement decrease during a spontaneous interaction with a humanoid robot.

We then studied the use of deep learning techniques with multimodal data for real-time detection of user engagement decrease. Our engagement classification results show that the real-time detector taking into account the past user behavior without any buffer performs well. Using the temporal dynamics of user behavior improves the results as well. The optimal size of the observation window of user behavior is found to be smaller than the average duration of SED segments (i.e. 6 s). Moreover, by using a delay of 1 or 2 s, we improved the performance of the detector. Depending on the application context, these delays could be reasonably suitable to improve the experience quality of interacting with the robot in-the-wild.

Finally, we believe that the publicly available dataset that we have collected [6], presents a high potential for other tasks in human–robot interaction (e.g. analysis of the social relationship between the user and the robot).

Acknowledgements This work was supported by European projects H2020 ANIMATAS (ITN 7659552) and a grant overseen by the French National Research Agency (ANR-17-MAOI). The authors would like to thank Nicolas Rollet and Christian Licoppe for useful discussions on pre-closing and Rodolphe Gelin, Angelica Lim, Marine Chanoux and Myriam Bilac from Softbank robotics for their help in the recording of UE-HRI dataset.

Funding This work was supported by SoftBank Robotics.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix A

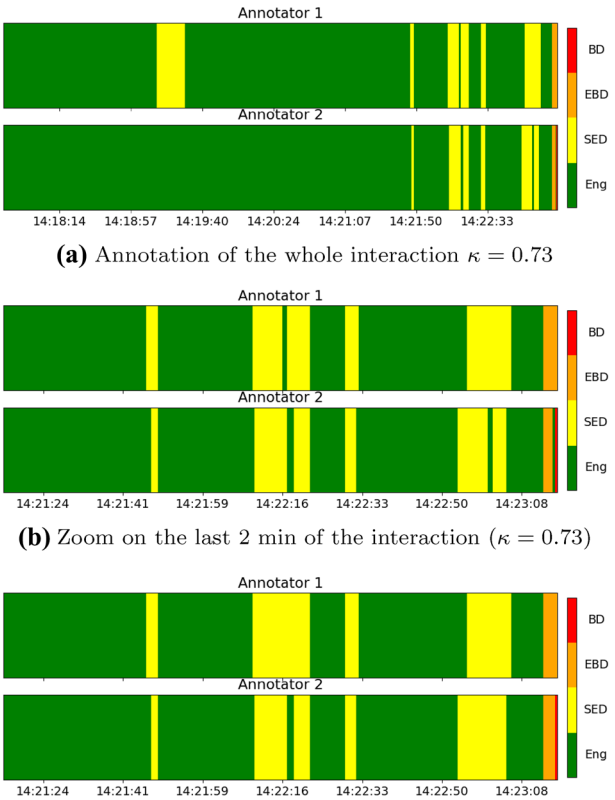
Survey of satisfaction presented as the final phase of the scenario. The participant was asked to indicate:

1. his satisfaction with the interaction,
2. his involvement in the interaction,
3. his desire to leave the interaction,
4. his desire to continue the interaction during the welcome phase,
5. his desire to continue the interaction during the dialog phase,
6. his desire to continue the interaction during the cucumber phase,
7. his desire to continue the interaction during the survey phase,
8. his desire to stay during the interaction,
9. if he believes that the robot wanted to stay during the interaction,
10. his desire to continue the conversation,
11. if he believes that the robot wanted to continue the conversation,

12. his feeling about his involvement in the interaction,
13. if he finds that the interaction was boring or fun,
14. if he finds that the information was interesting,
15. if he liked the interaction.

Appendix B

See Fig. 9.



(a) Annotation of the whole interaction $\kappa = 0.73$

(b) Zoom on the last 2 min of the interaction ($\kappa = 0.73$)

(c) Zoom on the last 2 min of the interaction with ignoring the "Engaged" segment less than 1 second located between 2 SED segments. $\kappa = 0.74$

Fig. 9 Example of annotation. BD: Engagement BreakDown i.e. leaving before the end of the interaction scenario. EBD: early sign of engagement breakdown (EBD) (i.e. the last SED of the interaction that BD will occur just after)

References

1. Andrist S, Bohus D, Kamar E, Horvitz E (2017) What went wrong and why? Diagnosing situated interaction failures in the wild. In: 9th international conference on social robotics (ICSR), Tsukuba, Japan
2. Anzalone SM, Varni G, Zibetti E, Ivaldi S, Chetouani M (2015) Automated prediction of extraversion during human–robot interaction. In: Finzi A, Alberto and Mastrogiovanni, Fulvio and Orlandini, Andrea and Sgorbissa (ed) AIRO@AI*IA, vol 1544, pp 29–39
3. Baltrusaitis T, Mahmoud M, Robinson P (2015) Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG). IEEE, pp 1–6
4. Baltrusaitis T, Zadeh A, Lim YC, Morency LP (2018) OpenFace 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face and gesture recognition (FG 2018). IEEE, pp 59–66
5. Ben-Youssef A, Clavel C, Essid S (2019) Early detection of user engagement breakdown in spontaneous human–humanoid interaction. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2019.2898399
6. Ben-Youssef A, Clavel C, Essid S, Bilac M, Chamoux M, Lim A (2017) UE-HRI: a new dataset for the study of user engagement in spontaneous human–robot interactions. In: Proceedings of the 19th ACM international conference on multimodal interaction, ICMI 2017. ACM, New York, pp 464–472
7. Bengio Y (2009) Learning deep architectures for AI. Found Trends Mach Learn 2(1):1–127
8. Bohus D, Horvitz E (2009) Learning to predict engagement with a spoken dialog system in open-world settings. In: Proceedings of the SIGDIAL 2009 conference on the 10th annual meeting of the special interest group on discourse and dialogue—SIGDIAL '09, September, pp 244–252
9. Bohus D, Horvitz E (2009) Models for multiparty engagement in open-world dialog. In: Proceedings of the SIGDIAL 2009 conference: the 10th annual meeting of the special interest group on discourse and dialogue, SIGDIAL '09. Association for Computational Linguistics, Stroudsburg, pp 225–234
10. Bohus D, Horvitz E (2009) Open-world dialog: challenges, directions, and a prototype. In: Proceedings of the IJCAI'2009 workshop on knowledge and reasoning in practical dialogue systems, Pasadena, California, USA, pp 34–45
11. Bohus D, Horvitz E (2014) Managing human–robot engagement with forecasts and...*um*...hesitations. In: Proceedings of the 16th international conference on multimodal interaction—ICMI '14. ACM Press, New York, pp 2–9
12. Bosch N, D'Mello S (2015) The affective experience of novice computer programmers. Int J Artif Intell Educ 27(1):181–206
13. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30(7):1145–1159
14. Castellano G, Leite I, Pereira A, Martinho C, Paiva A, McOwan PW (2012) Detecting engagement in HRI: an exploration of social and task-based context. In: 2012 international conference on privacy, security, risk and trust and 2012 international conference on social computing. IEEE, pp 421–428
15. Celiktutan O, Skordos E, Gunes H (2017) Multimodal human–human–robot interactions (MHHRI) dataset for studying personality and engagement. IEEE Trans Affect Comput
16. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation
17. Chollet F (2015) keras. https://github.com/fchollet/keras. Accessed 05 Feb 2018
18. Clavel C, Cafaro A, Campano S, Pelachaud C (2016) Fostering user engagement in face-to-face human–agent interactions: a survey. Springer, Cham, pp 93–120
19. Corrigan LJ, Peters C, Küster D, Castellano G (2016) Engagement perception and generation for social robots and virtual agents. Springer, Cham, pp 29–51
20. D'Mello S, Graesser A (2012) Dynamics of affective states during complex learning. Learn Instr 22(2):145–157

21. Dominey P, Metta G, Nori F, Natale L (2008) Anticipation and initiative in human-humanoid interaction. In: Humanoids 2008—8th IEEE-RAS international conference on humanoid robots. IEEE, pp 693–699
22. Eyben F, Wöllmer M, Schuller B (2010) Opensmile: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the international conference on multimedia—MM '10. ACM Press, New York, pp 1459–1462
23. Feil-Seifer D, Mataric M (2005) Defining socially assistive robotics. In: 9th international conference on rehabilitation robotics, 2005. ICORR 2005. IEEE, pp 465–468
24. Foster ME, Alami R, Gestranius O, Lemon O, Niemelä M, Odobez JM, Pandey AK (2016) The MuMMER project: engaging human–robot interaction in real-world public spaces. Springer, Cham, pp 753–763
25. Foster ME, Gaschler A, Giuliani M (2017) Automatically classifying user engagement for dynamic multi-party human-robot interaction. Int J Soc Robot 9(5):659–674
26. Gehle R, Pitsch K, Dankert T, Wrede S (2017) How to open an interaction between robot and museum visitor? Strategies to establish a focused encounter in HRI. In: Proceedings of the 2017 ACM/IEEE international conference on human–robot interaction—HRI '17. ACM Press, New York, pp 187–195
27. Glas N, Pelachaud C (2015) User engagement and preferences in information-giving chat with virtual agents, pp 33–40
28. Hall J, Tritton T, Rowe A, Pipe A, Melhuish C, Leonards U (2014) Perception of own and robot engagement in human–robot interactions and their dependence on robotics knowledge. Robot Autonom Syst 62(3):392–399
29. Hayashi K, Sakamoto D, Kanda T, Shiomi M, KoizumiS, Ishiguro H, Ogasawara T, Hagita N (2007) Humanoid robots as a passive-social medium. In: Proceedings of the ACM/IEEE international conference on human–robot interaction—HRI '07. ACM Press, New York, p 137
30. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–80
31. Human Vision Components (HVC-P2) B5T-007001 Command Specifications. Technical report, OMRON Corporation Electronic and Mechanical Components Company, Japan (2016)
32. Ivaldi S, Lefort S, Peters J, Chetouani M, Provasi J, Zibetti E (2017) Towards engagement models that consider individual factors in HRI: on the relation of extroversion and negative attitude towards robots to gaze and speech during a human-robot assembly task. Int J Soc Robot 9(1):63–86
33. Joder C, Essid S, Richard G (2009) Temporal integration for audio classification with application to musical instrument classification. IEEE Trans Audio Speech Lang Process 17(1):174–186
34. Kanda T, Shiomi M, Miyashita Z, Ishiguro H, Hagita N (2009) An affective guide robot in a shopping mall. In: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction—HRI '09. ACM Press, New York, p 173
35. Kendon A (1967) Some functions of gaze-direction in social interaction. Acta Psychol 26:22–63
36. Leite I, McCoy M, Ullman D, Salomons N, Scassellati B (2015) Comparing models of disengagement in individual and group interactions. In: Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction—HRI '15. ACM Press, New York, pp 99–105
37. Li L, Xu Q, Tan YK (2012) Attention-based addressee selection for service and social robots to interact with multiple persons. In: Proceedings of the workshop at SIGGRAPH Asia, WASA '12. ACM, New York, pp 131–136
38. Liu T, Kappas A (2018) Predicting engagement breakdown in HRI using thin-slices of facial expressions. In: Workshops at the thirty-second AAAI conference on artificial intelligence, pp 37–43
39. Martinovski B, Traum D (2003) The error is the clue: breakdown in human–machine interaction. In: Proceedings of the ISCA workshop on error handling in spoken dialogue systems, pp 11–17
40. Miller RB (1968) Response time in man-computer conversational transactions. In: Proceedings of the December 9–11, 1968, fall joint computer conference, part I on–AFIPS '68 (Fall, part I). ACM Press, New York, p 267
41. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
42. Pitsch K, Kuzuoka H, Suzuki Y, Sussenbach L, Luff P, Heath C (2009) "The first five seconds": contingent stepwise entry into an interaction as a means to secure sustained engagement in HRI. In: RO-MAN 2009—the 18th IEEE international symposium on robot and human interactive communication. IEEE, Toyama, pp 985–991
43. Poggi I (2007) Mind, hands, face and body: a goal and belief view of multimodal communication. Weidler Buchverlag, Berlin
44. Rawassizadeh R, Momeni E, Dobbins C, Gharibshah J, Pazzani M (2016) Scalable daily human behavioral pattern mining from multivariate temporal data. IEEE Trans Knowl Data Eng 28(11):3098–3112
45. Rich C, Ponsler B, Holroyd A, Sidner CL (2010) Recognizing engagement in human–robot interaction. In: 2010 5th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 375–382
46. Robots in public spaces (2013) towards multi-party, short-term, dynamic human-robot interaction. In: Giuliani M, Petrick R (eds) International conference on social robotics (ICSR 2013), Bristol, UK
47. Schuller B, Ganascia JG, Devillers L (2016) Multimodal sentiment analysis in the wild: ethical considerations on data collection, annotation, and exploitation. In: Actes du workshop on ethics in corpus collection, annotation and application (ETHI-CA2), LREC, Portoroz, Slovénie
48. Schuller B, Müeller R, Höernler B, Höethker A, Konosu H, Rigoll G (2007) Audiovisual recognition of spontaneous interest within conversations. In: Proceedings of the ninth international conference on multimodal interfaces—ICMI '07. ACM Press, New York, p 30
49. Sidner CL, Lee C, Kidd CD, Lesh N, Rich C (2005) Explorations in engagement for humans and robots. Artif Intell 166(1–2):140–164
50. Tapus A Mataric MJ (2008) Socially assistive robots: the link between personality, empathy, physiological signals, and task performance. Undefined
51. Trung P, Giuliani M, Miksch M, Stollnberger G, Stadler S, Mirnig N, Tscheligi M (2017) Head and shoulders: automatic error detection in human–robot interaction. In: Proceedings of the 19th ACM international conference on multimodal interaction—ICMI 2017. ACM Press, New York, pp 181–188
52. Vaufreydaz D, Johal W, Combe C (2016) Starting engagement detection towards a companion robot using multimodal features. Robot Autonom Syst 75:4–16
53. Wittenburg P, Brugman H, Russel A, Klassmann A, Sloetjes H (2006) ELAN: a professional framework for multimodality research. In: LREC 2006, pp 1556–1559
54. Wood E, Baltruaitis T, Zhang X, Sugano Y, Robinson P, Bulling A (2015) Rendering of eyes for eye-shape registration and gaze estimation. In: 2015 IEEE international conference on computer vision (ICCV). IEEE, pp 3756–3764

Atef Ben-Youssef is the head of Research and Development at Ludo-Vic SAS. He received his Ph.D. degree in signal, image, speech, telecommunication from Grenoble-Alpes University in 2011. After his Ph.D. at GIPSA-Lab, he worked as post-doc at the CSTR, University of Edinburgh, from 2012 to 2014. Then, he was at the LIMSI-CNRS, from 2014 to 2015. In 2015–2016, he worked as a teacher and researcher attached to Grenoble-Alpes University. From September 2016 to December 2018, he joined the LTCI, Telecom-ParisTech. He is interested in machine learning applied for social computing including the analysis, the recognition, the interpretation and the synthesis of social behavior.

Giovanna Varni is an Associate Professor at LTCI, Télécom Paris, Institut polytechnique de Paris, France. Previously, She was post-doc at the Institute for Intelligent Systems and Robotics (ISIR), Pierre et Marie Curie University, in Paris (France), and at InfoMus Lab, DIBRIS, Università degli Studi di Genova, in Genova (Italy). She is an interdisciplinary researcher mainly investigating on Social Signal Processing (SSP), Affective Computing, and Human-Computer Interaction.

Slim Essid is a Full Professor at Telecom Paris' department of Images, Data & Signals and the coordinator of the Audio Data Analysis and Signal Processing team. His research interests are in machine learning for audio and multimodal data analysis. He received the M.Sc. (D.E.A.) degree in digital communication systems from the École Nationale Supérieure des Télécommunications, Paris, France, in 2002; the Ph.D. degree from the Université Pierre et Marie Curie (UPMC), in 2005; and the habilitation (HDR) degree from UPMC in 2015. He has been involved in various collaborative French and European research projects, among them Quaero, Networks of Excellence FP6-Kspace and FP7-3DLife, and collaborative projects FP7-REVERIE and FP7-LASIE. He has published over 100 peer-reviewed conference and journal papers with more than 100 distinct co-authors. On a regular basis he serves as a reviewer for various machine learning, signal processing, audio and multimedia conferences and journals, for instance various IEEE transactions, and as an expert for research funding agencies.

Chloé Clavel has been Associate Professor at Télécom ParisTech since 2013 at the Laboratoire de Traitement et Communication de l'Information (LTCI). Her research work deals with interactions between humans and virtual agents, from user's socio-emotional behavior analysis to socio-affective interaction strategies with a focus on speech and language processing. Her research is integrated into a broader topic of social computing that she is leading within the LTCI. She is currently working on interactions between humans and virtual agents, from the analysis of the user's socio-emotional behaviour (verbal and non-verbal) to socio-emotional interaction strategies. She has participated in several European and national collaborative projects around Social Computing (e.g. H2020 ITN ANIMATAS, aria-valuspa UE-TIC, Labex smart). She recently obtained a French funding for Young Researchers on the themes of opinion analysis in interactions (ANR MAOI). Her research has resulted in 61 publications including 16 journal papers with a total of 1019 citations. She is Associate Editor of IEEE Transactions on Affective Computing, acts as reviewers for various journals (ex: International Journal of Human-Computer Studies, Computer Speech and Language etc) and is part of program and organization committees of national and international conferences.