UE-HRI: A New Dataset for the Study of User Engagement in Spontaneous Human-Robot Interactions

Atef Ben-Youssef
Chloé Clavel
Slim Essid
Télécom ParisTech, Université Paris-Saclay
Paris, France
{atef.benyoussef,chloe.clavel,slim.essid}@
telecom-paristech.fr

Miriam Bilac
Marine Chamoux
Angelica Lim
SoftBank Robotics Europe
Paris, France
{mbilac,mchamoux,alim}@softbankrobotics.com

ABSTRACT

In this paper, we present a new dataset of spontaneous interactions between a robot and humans, of which 54 interactions (between 4 and 15-minute duration each) are freely available for download and use. Participants were recorded while holding spontaneous conversations with the robot Pepper. The conversations started automatically when the robot detected the presence of a participant and kept the recording if he/she accepted the agreement (*i.e.* to be recorded). Pepper was in a public space where the participants were free to start and end the interaction when they wished. The dataset provides rich streams of data that could be used by research and development groups in a variety of areas.

CCS CONCEPTS

• **Human-centered computing → User studies**;

KEYWORDS

HRI dataset, user engagement, in-the-wild

ACM Reference Format:
Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. UE-HRI: A New Dataset for the Study of User Engagement in Spontaneous Human-Robot Interactions. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3136755.3136814

1 INTRODUCTION

Social robots should be able to communicate and cooperate with humans as naturally as possible and build a relationship with them. One of the key challenges for social robots is to maintain user engagement and avoid engagement breakdown where users cannot proceed with the conversation [12]. We define *user engagement breakdown* in human-robot interaction (HRI) as a situation where the user prematurely ends the interaction, *i.e.* before the robot has had the chance to receive the complete feedback it expects from the user, as part of a running scenario.

In fact, robots should be able to recognise the level of engagement of humans in order to adapt their behavior during the interaction and predict disengagement to make correct decisions at the end of the interaction. Engagement in human-computer interaction has been addressed from different angles. Peters *et al.* [18] distinguished the two following components underlying the engagement process: *attentional involvement* and *emotional involvement*. The former is well captured by the definitions proposed by Sidner [23]: *"the process by which individuals in an interaction start, maintain and end their perceived connection to one another"*; and by Poggi [19]: *"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"*. Whereas the latter (*i.e. Emotional involvement*) was studied in [10, 11]. The authors showed that the more emotions users show, the higher their level of engagement in the interaction. They also distinguished positive and negative engagement and associated them with positive and negative emotions, respectively.

The design of engaging robots is paramount, whether in short-term human-robot interactions, or for building long-term relationships between the user and the robot. The challenge of modelling user engagement in HRI is to develop robots that respond appropriately to the user's behaviour and engage them in stimulating experiences.

In order to address user engagement in HRI, we have assembled a real dataset involving the humanoid robot Pepper. This dataset consists of 195 interactions where participants were free to enter into the interaction if they wished, free to leave when they wanted, and were expected to behave in an unconstrained way.

The dataset contains 54 manually annotated recordings freely available to the research community. They are comprised of synchronized multimodal data recorded with Pepper's sensors (microphones, cameras, sonars, laser). The annotations are related to cues of engagement decrease as well as negative affects.

The present paper provides a review of the various works addressing the common objective of fostering user engagement (Section 2). It is structured as follows. The collection of user engagement data is detailed in Section 3. Section 4 focuses on the properties on the recorded data as well as the analysis of users' engagement. Section 5 outlines possible fields of application for the current dataset. Finally, we present the conclusion in Section 6.

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3136814

2 STATE OF THE ART

In this section, we discuss state-of-the-art techniques and strategies for designing engaging robots in face-to-face or multi-party interactions with users. We also briefly review some studies aimed at evaluating the user's behaviour to improve the level of engagement with the robot as well as the available datasets.

In general, user engagement can be measured via i) user self-reports, *i.e.* in a subjective fashion; ii) by monitoring the user's responses, tracking the user's body postures, intonations, head movements and facial expressions during the interaction, *i.e.* objectively; or iii) by manually logging behavioural responses of user experience. This reflects a common categorization in experimental design [8]. A researcher could attempt to adopt any of the three above-mentioned approaches (or even combinations of those) to capture engagement.

Engagement recognition: In these studies [2, 20], analyses were performed with a robot trying to engage a human in a specific task. *Objective studies* relied on recognising engagement between humans and humanoid robots [20]. Social robots, iCub and NAO, were used in different human-robot interaction experiments to confirm the importance of non-verbal cues in improving the interactions between humans and robots [2]. *Subjective assessments* of engagement have been obtained through two questionnaires: 48 questions related to the extroversion dimension from the Revised Personality Inventory (NEO-PIR) [7] and 14 questions on a 7-point Likert-type scale from the Negative Attitude towards Robots Scale (NARS) [17]. In [13], the authors found that extroversion had a significant correlation with speech frequency and duration (the more extrovert an individual, the more often and longer they tend to address the robot) and negative attitude towards robots tends to be related to the gaze (time spent looking at the robot). Another interesting finding is that engagement models used in human-robot interaction should take into account attitudes and personality traits.

Disengagement prediction: Objective measures were taken in different conditions to predict disengagement [4, 14]. The most closely related work to our study is that of Bohus *et al.* [4], which presents a self-supervised forecasting disengagement approach used to anticipate within some small time interval when participants are about to finish their interactions. Such a disengagement forecast can better inform decision-making. The authors deployed the robot NAO "in-the-wild" to interact with one or multiple participants. They investigated methods for managing disengagement decisions that address a tradeoff between making early, incorrect disengagement decisions, versus being more conservative, and decide too late. In [14], the authors analysed two different disengagement prediction models by changing the group size of users in child-robot interactions. Their dataset consists of children in two different conditions (single and group conditions) interacting with two MyKeepon social robots. Using the exact same set of data (*i.e.*, where the group model does not encode features from the other participants around the robots), they show that the appropriate selection of multimodal features (gaze, head rolling, annotated affect signs, robot behaviour) can be used to predict disengagement.

Affect-related phenomena and engagement: In [1], negative affect leading to an engagement breakdown was studied *objectively*. Ang *et al.* [1] use prosodic features, language modelling and speaking style to detect user frustration with a telephone-based dialog system interface. They show that a prosodic decision trees can predict whether an utterance is neutral or "annoyed or frustrated". They also find that using speaking style as an indicator, increases the performance. However, their language model features were found to be poor predictors of frustration.

Subjective studies on dynamics of affective states were analysed in [6, 9]. The authors proposed a model of affective state transitions during learning activities. After interacting with AutoTutor in a tutorial session, participants were asked to provide judgments of the affective states they experienced by viewing their face and screen videos recorded during the tutorial session. The major dynamics of affective state predictions of the model were: Engagement to/from confusion, confusion to/from frustration, and frustration to/from boredom. A crucial point at which the learner disengages is when they persist in frustration, leading into boredom.

2.1 Social Signals Used for (Dis)Engagement

Table 1 summarises the features used in previous studies addressing engagement and disengagement in HRI. The most common cue appears to be gaze. Both static and dynamic human posture and gaze have been considered in [2], as they can be used to extract different pieces of meaningful information related to how humans respond to the robot as well as the synchrony of movements.

A professional tagger performed the annotation based on video and audio of the interactions from the robot's viewpoint in [4]. The annotator also identified *early disengagements*, *i.e.*, situations where the system stopped the conversation early, before the participants actually disengaged.

Annotation was performed in [14] by coding the start and end times of each participant's vocalisations, backchannel sounds, body posture (leaning forward/backward, arms on the table), gestures (smiles, mimicking robots, excitable bouncing and strong emotional reactions), concentration and boredom signs and off task behaviours. Head orientation features, looking at the robots, looking up, looking down and rolling head were extracted using the recorded video. Contextual features (robots speaking, robots bouncing, and participant choosing an action) were extracted from the interaction logs.

2.2 Methods Used for (Dis)Engagement Characterisation

Machine learning based approaches have been largely used for modelling user engagement. Logistic regression and boosted decision tree models were compared in [4]. Based on the analysis of early-detection-time versus the false-positive rate, Bohus *et al.* [4] selected the 5-second lookahead logistic regression model for managing disengagement decisions. Leave-one-out cross-validation using Support Vector Machines (SVMs) based method was used in [14]. Leite *et al.* [14] found that the disengagement model trained in the single condition might not be appropriate for the group condition, but the group condition model generalizes better to single condition. A mixed model combining both conditions is a good compromise, but it does not achieve the performance levels of

Table 1: Used streams for (dis)engagement in HRI

Streams	Associated phenomena	Study
Gaze	Engagement recognition	[2]
- Amount of shared looking		[13]
- Looking at the robot		[20]
- Looking at face and hands		[23]
- Mutual gaze	Disengagement prediction	[14]
- Directed: pointing an object		
Body posture	Engagement recognition	[2]
	Disengagement prediction	[14]
Head motion	Engagement recognition	[2]
- Nodding		[20]
- Shaking		
Speech	Extroversion and negative attitude towards robots	[13]
- Frequency		
- Duration	Disengagement prediction	[14]
Face	Forecasting disengagement	[4]
- Horizontal location	Disengagement prediction	[14]
- Size of the tracked face		
- Tracking confidence score		
- Smiles, emotional reaction*		
Dialog state	Forecasting disengagement	[4]
- Semantic output		
Output of other models	Forecasting disengagement	[4]
- Attention inference output		
Manual annotation	Disengagement prediction	[14]
- Concentration		
- Boredom signs		
- Off task behaviours		

the models trained for a specific type of interaction. In the dialogue breakdown detection challenge organised by [12] where the task was to detect a system's inappropriate utterances that lead to dialogue breakdowns, the best performance was found using LSTM-RNN (Long Short-Term Memory - Recurrent Neural Network) based methods used by two teams from the six who participated in this challenge.

Rule-based approaches were also used. Heuristic rules were used as a baseline system in [4] whereas Rich *et al.* [20] used state machines for recognising engagement.

2.3 Available Datasets

In social signal processing, the available multimodal datasets essentially feature face-to-face interactions with virtual agents rather than robots. The SEMAINE dataset proposed by SSPNet[1] involves several persons recorded in a face-to-face interaction [16]. Another dataset, called Cam3D, offers annotated multimodal recordings [15]. This dataset uses three different sensors (Kinect, cameras, and microphones) to record spontaneous speech, facial expressions and hand gestures in a desktop environment setting. Unfortunately, these datasets are less suited to human-robot interaction, especially as the social signals involved in user engagement characterisation are not restricted to facial expression and speech. For instance, other relevant cues include distance to the robot.

As far as interactions involving robots are concerned, a closely related dataset has been presented in [4]. Bohus *et al.* [4] collected data using the robot NAO. Their recordings consist of 133 interactions with 158 users (more users than interactions due to a multiparty interaction setting). In [14], two MyKeepon Robots were used to play out interactive stories around emotional words (e.g., frustration, inclusion, cooperation). The dataset consists of 40 annotated children-robots interactions.

Vaufreydaz *et al.* [24] collected data of mono-user and multi-users interactions with the robot Kompaï in a home-like environment. In their multimodal dataset collection, they used a Kinect in addition to the robot sensors: laser, ultrasound, infrared telemeters and a webcam.

The Vernissage dataset [25] contains 13 sessions of NAO interacting with two persons. These are multi-party interactions which were manually annotated with several nonverbal cues. Table 2 lists several HRI datasets.

Our collected dataset targets the diverse social signals that are involved in user engagement, considering a wide range of heterogeneous sensors: a microphone array, cameras, depth sensors, sonars, lasers, along with user feedback captured through Pepper's touch screen. To the best of our knowledge, none of the existing datasets provides such a thorough coverage of signals amenable to exploitation for user engagement analysis. This is also the first significant dataset offering data collected by the largely used robot "Pepper"[2].

3 RECORDING DATA

3.1 Data Acquisition

Our dataset consists of recordings of humans interacting with the social robot Pepper. The robot is set up in a public space,where the data collection is conducted. Participants are free to enter into the interaction and free to leave when they want, thus behaving in a spontaneous way. They know before starting the interaction that their video and audio are recorded, as a poster explicitly states they are being filmed. Yet, they do not know the actual purpose of the study. Figure 1 shows the setting of the interaction. *Engagement*

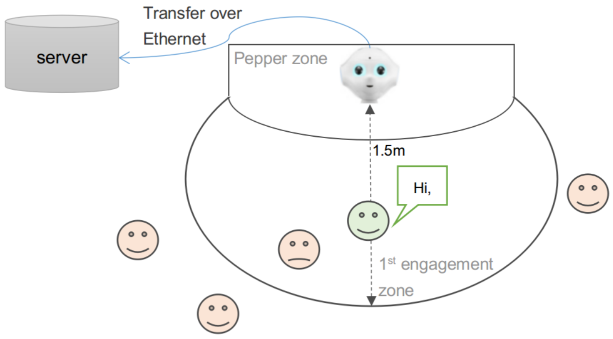


Figure 1: Recording setting.

[1] http://sspnet.eu/

[2] https://www.ald.softbankrobotics.com/en/cool-robots/pepper

Table 2: Overview of HRI dataset: F: Face, S: Speech, G: Gestures, D: Dialog

dataset	Modalities	Spontaneity	In-the-wild	#Interactions	Annotation	Publicly available
UE-HRI	F/S/G/D	Yes	Yes	195	Yes	Yes (54)
Bohus *et al.* [4]	F/S/G/D	Yes	Yes	133	Yes	No
Leite *et al.* [14]	F/S/G	Yes	No	40	Yes	No
Vaufreydaz *et al.* [24]	F/S/G	Yes	No	29	Yes	No
Vernissage [25]	F/G	Yes	No	13	Yes	Yes

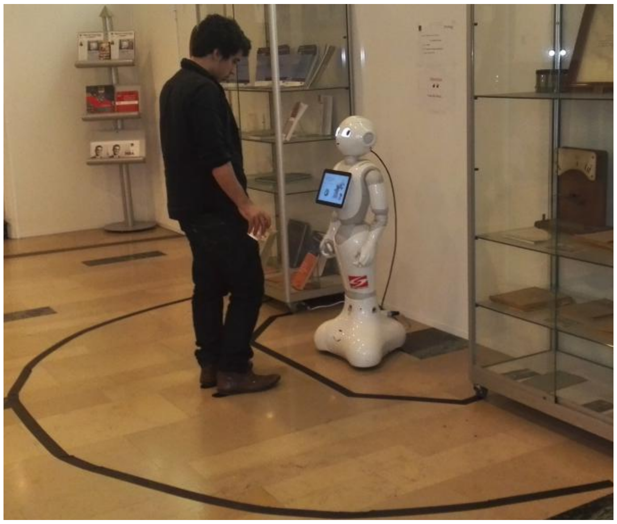


Figure 2: Participant in the first engagement zone (less than 1.5 meter from the robot) interacting with Pepper.

zones refer to the positions of detected people in space in three loci in the neighborhoud of the robot. The robot zone as well as its first engagement zone (*i.e.* distance of less than 1.5 meter from the robot) are indicated using black tape stuck on the floor (see Figure 2).

All data streams available on Pepper are packaged in the open-source Robot Operating System (ROS) framework[3]. Each stream is passed into a message (called ROS topic) and packaged together into a ROSbag file. All streams are indexed using the robot timestamps to avoid synchronisation issues. 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 and avoid data loss (even if the application were to crash, a part of the data would still be usable). ROSbag files are then merged together into one ROSbag file in order to get one file per interaction.

The recording starts automatically when the robot detects the presence of a person. If the participant validates the agreement displayed on the robot's screen, the recorded data is kept and the transfer to the server started. Otherwise, it is deleted.

3.2 Interaction Scenario

The robot starts the interaction automatically when it detects movement, and focuses on the participant who is in the first engagement zone and in front of it. The robot starts by showing an ***agreement*** and asking the participant

- to enter an email address (mandatory);
- to confirm that they are over 18 (mandatory);
- to decide whether their data could be used by us for improving the robot capacities (mandatory);
- to decide whether their data could be used for dissemination purposes (*i.e.* at academic conferences, publications or as part of teaching material) (optional);
- to decide whether their data could be shared with other researchers (for research purposes and not only for improving the robot) (optional).

After the validation of the agreement, the robot starts the ***welcome*** phase by presenting itself using very lively animations, and gives instructions ("speak loud and be alone in the 1^{st} engagement zone"), followed by the ***dialog*** phase that includes a set of open-ended questions where the robot asks the participant to introduce him/her-self, and to talk about their favourite books and films. The ***cucumber*** phase represents a modified version of a popular NAOqi application. In this phase, the robot presents its vision technology to the participant by showing that, from its viewpoint, the difference between a cucumber and human is the face. The final phase was a ***survey*** of satisfaction where 15 questions related to the interaction were asked. Figure 3 summarises with a time-line the progression of the scenario.

3.3 Recorded Streams

Table 3 presents the data streams that Pepper can provide. All streams are recorded synchronously in ROSbag format v2.0. The data can be described with a hierarchical structure. The *low-level features* represent the raw signals recorded using the different sensors on Pepper: microphones (4 channels, 48000 Hz, 16 bit signed), video cameras (up to 640*480 at 30 fps), depth sensor (up to 320x240 at 20 fps), sonars (up to 42 KHz), laser (up to 6.25 Hz). Tracked variables (*i.e.* face features, head angles, eye gaze and human position toward the robot) are *higher-level features* computed using available Pepper trackers via the NAOqi SDK. The output of available modules gives the highest-level features. Figure 4 shows Pepper

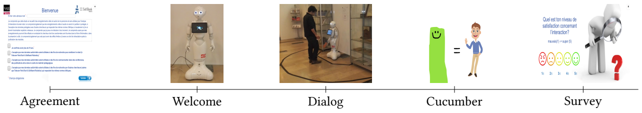


Figure 3: Phases of the scenario in chronological order.

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

Table 3: Pepper data streams

Low level	Intermediate level	High level
Measured signals	Tracked variables	Output of others modules
Videos	Face (Location of eyes, nose, mouth)	Smile degree
- Two 2D Cameras (front and bottom)	Head motion (rotation angles)	Facial expressions (neutral, happy, surprised, angry or sad)
- One 3D Sensor (depth)	Eye gaze (direction, opening degree)	Looking at robot
Audio (4 directional microphones)	User distance	User dialog input (ASR)
Sonars	User position	Robot Text(-To-Speech)
Laser	Engagement Zone	Age estimation
	Robot behaviour (joint_states)	Gender estimation
		Dialog (user last input, Text-To-Speech Word)

sensors position as well as a 3D visualisation of a frame, including sonars, laser, front, bottom and depth cameras using rviz[4].

3.4 Annotation Protocol

We use the ELAN[5] software to annotate the dataset. We developed a script that extracts synchronised front and bottom images and audio from the corresponding ROSbag topics and merges them into a video using ffmpeg[6].

The annotators are asked to listen/watch the video (recorded using Pepper [front and bottom] cameras) and to annotate the start of the interaction. When they finish watching the entire video, they indicate the end of the interaction as well as the number of participants (*i.e.* mono-user or multi-users). In order to characterise engagement, annotators are asked to annotate the interaction video segment by segment based on verbal and non-verbal behaviour expressed by the user that demonstrate an engagement decrease, with one of the following four labels:

- Sign of Engagement Decrease (SED) observed during the interaction (None of the 3 next labels).
- Early sign of future engagement BreakDown (EBD) *i.e.* first noticeable clue that an engagement breakdown will occur in the remainder of the interaction
- engagement BreakDown (BD) *i.e.* leaving before the end of the interaction.
- Temporary disengagement (TD) *i.e.* leaving for some time and coming back to the interaction.

A sign of engagement decrease (SED) reflects any clue of engagement decrease exhibited by the participant during the interaction. It could occur any time during the interaction. For each sign of engagement decrease, the annotator is asked to decide if it is an early sign of engagement breakdown, temporary disengagement, engagement breakdown or none of them (simply sign of engagement decrease).

An early sign of engagement breakdown (EBD) is considered to be the first noticeable clue shown by the participant that annotator detects as indicating a future breakdown. This clue may correspond to verbal or nonverbal behaviors of the participant.

By the end of the interaction, the annotators are also asked to annotate whether the departure of the participant corresponds to an engagement breakdown (BD) or not. The end of the interaction is defined by leaving the first engagement zone (*i.e.* leaving the 1.5-meter wide area around the robot).

If the participant leaves the engagement zone and comes back to the interaction, the annotator indicates that it is a temporary disengagement (TD). For example, if the participant goes away to say "hello" to a friend and comes back to continue the interaction with the robot. This kind of sequence thus corresponds to a "temporary leave". For example, when the participant looks away and discusses with another person, if (s)he returns back to the interaction with the robot, that is considered as a sign of engagement decrease (SED), if not (and (s)he leaves after that), that is a sign of early engagement breakdown (EBD). When the participant leaves the engagement zone for any reason and comes back to continue the interaction, it will be a temporary leave (TD).

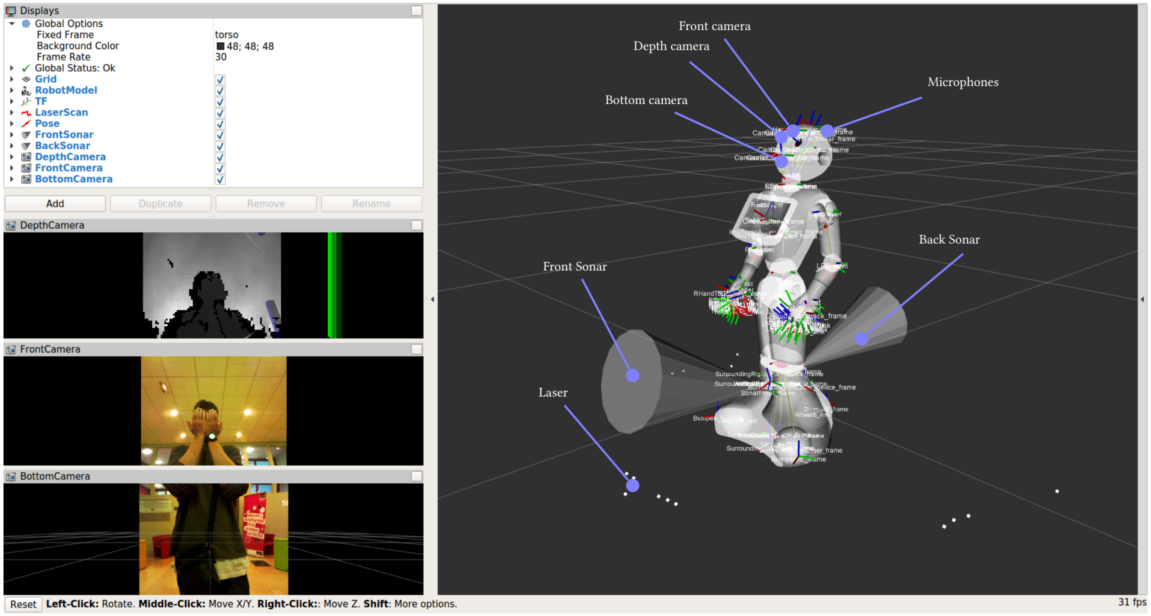
The annotator defines the start and the end segment as well as the corresponding label, observed cues and negative affect of that segment. For each defined segment of engagement decrease, the annotator assigns the corresponding observed cues of that decrease in importance order. This part could be sub-segmented. For example, if the participant says:"I'm bored", accompanied with 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 they add information about the causes of the engagement breakdown in the "Causes" track. Researchers with different scientific background participated in the process.

4 DATASET OVERVIEW

The data was collected for a period of 56 days in a hallway university. The total number of participants is so far 195 (125 males, 70 females). The total size of the data is 1.42 Tb. Since we are collecting spontaneous and realistic interactions and no instructions are given

[4] http://wiki.ros.org/rviz
[5] https://tla.mpi.nl/tools/tla-tools/elan/
[6] https://ffmpeg.org/



Displays
- Global Options
  - Fixed Frame: torso
  - Background Color: ■ 48; 48; 48
  - Frame Rate: 30
- ✓ Global Status: Ok
- Grid ☑
- RobotModel ☑
- TF ☑
- LaserScan ☑
- Pose ☑
- FrontSonar ☑
- BackSonar ☑
- DepthCamera ☑
- FrontCamera ☑
- BottomCamera ☑

Reset Left-Click: Rotate. Middle-Click: Move X/Y. Right-Click:: Move Z. Shift: More options.

31 fps

Figure 4: Low-level Pepper data streams recorded in ROS.

Table 4: dataset statistics. BD: Leaving before the end, NBD: Staying till the end

	BD	NBD
#Interaction	148	47
Average Duration (min)	6	13
Size (Gb)	1069,	382,5
#Accept to shared with research community	41	13
Size (Gb) of the 54 shared interactions	287,6	90,4

to the participants, we cannot control the number of users deciding to engage simultaneously in an interaction with the robot, though there was an instruction that only one user had to be in the first engagement zone at a time. We have thus found 124 interactions to feature a single user, and obtained 71 multiparty interactions (40 started as multiparty and ended as mono-user).

- The number of participants who accepted that their data could be used for dissemination purposes is 74.
- The number of participants who accepted that their data could be shared with other researchers is 54.

The number of participants who stayed until the survey phase is 53, and 47 stayed until the end of the interaction. Figure 5 presents the number of departures per phase. Table 4 presents statistics of the dataset with details on who left before the end (denoted by BD) and who stayed until the end (denoted by NBD). The data of 54 interactions (36 males, 18 females), where 32 are mono-user and 22 are multiparty interactions (12 start as multiparty and end as mono-user)) will be made publicly available through the website[7].

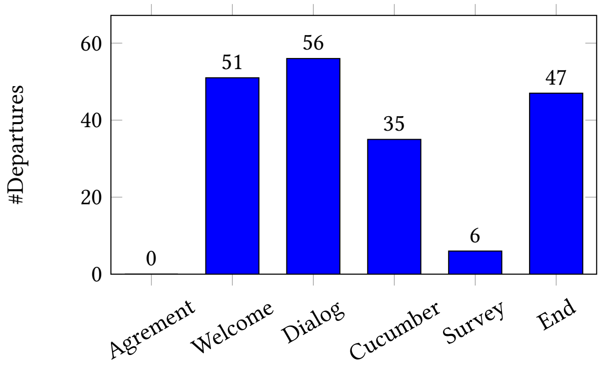
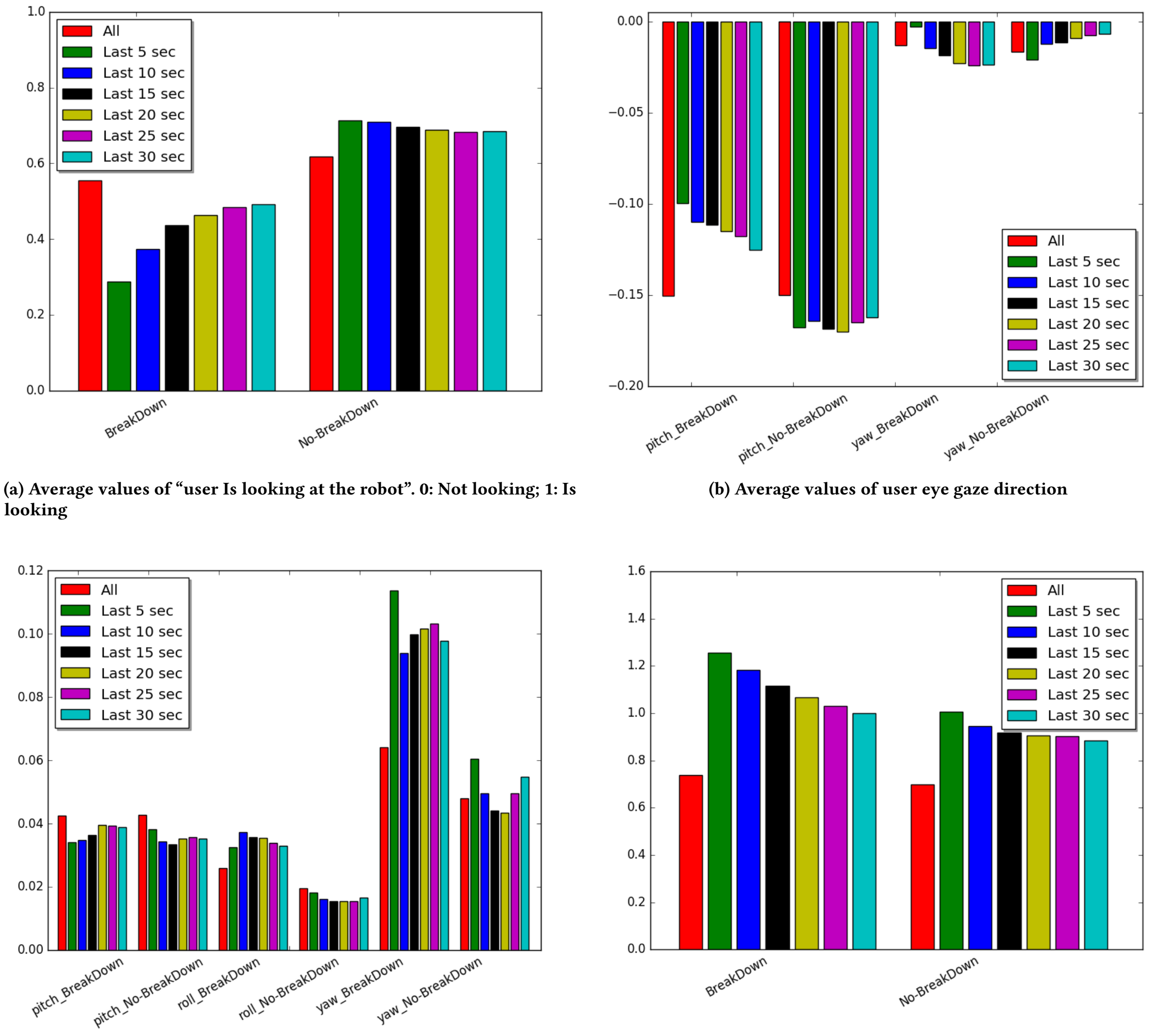


Figure 5: Number of departures by phase.

4.1 Data Analysis

Figure 6 presents an objective study based on a comparison between the behaviour of users who left the interaction before the end (denoted by BreakDown) and those who stayed until the end (denoted by No-BreakDown). The comparison is made on whole interactions as well as the last seconds before departure in both cases (from last 30 seconds before leaving to the last 5 seconds). Both mono-user and multiparty interactions are used. The analysis is on the data of the focused user. Regarding eye gaze, participants who stayed until the end look more at the robot than those who left before the end (Figure 6a). This could be confirmed with vertical gaze direction around pitch axis in Figure 6b. Figure 6c displays head motion variation over users. It illustrates a shaking (*i.e.* yaw axis) and tilting (*i.e.* roll axis) tendency for users who left before the end of the interaction than who stayed until the end. Figure 6d shows the average distance of users toward the robot computed

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



(a) Average values of "user Is looking at the robot". 0: Not looking; 1: Is looking

(b) Average values of user eye gaze direction

(c) Average user head angles variances

(d) Average user distance computed using front sonar sensor (in meters)

Figure 6: Comparison between users' behaviours. BreakDown: Leaving before the end; No-BreakDown: Staying till the end.

using the front sonar (see Figure 4). The participants who stayed until the end were closer to the robot than those who left before the end.

A subjective study based on a questionnaire was presented at the end of the interaction (*i.e.* the final phase). On a 5-level Likert scale (from disagree "1" to agree "5"), the participant was asked to indicate

(1) satisfaction with the interaction,
(2) involvement in the interaction,
(3) desire to leave the interaction,
(4) desire to continue the interaction during the welcome phase,
(5) desire to continue the interaction during the dialog phase,
(6) desire to continue the interaction during the cucumber phase,
(7) desire to continue the interaction during the survey phase,
(8) desire to stay during the interaction,
(9) belief that the robot wanted to continue the conversation,
(10) desire to continue the conversation,
(11) feeling about involvement in the interaction,
(12) was the interaction boring or fun,
(13) was the information interesting,
(14) did they like the interaction.

Figure 7 shows the average participants' response to the survey over 53 participants. The participants wanted to continue the interaction more clearly during the welcome phase, where they watched the robot introduce itself and give instructions, compared to the other phases, where they had to communicate with the robot. They thought that the robot wanted to stay and to continue the conversation with them more than themselves wanted to do so.

Participants who stayed until the end were satisfied and liked the interaction. However, it does not seem to be the case for those who left the interaction. They are divided into two groups: 1) those who left because of constraints (*e.g.* time constraint: they have to leave to attend a course or meeting) and 2) those who were frustrated and did not enjoy the proposed scenario.

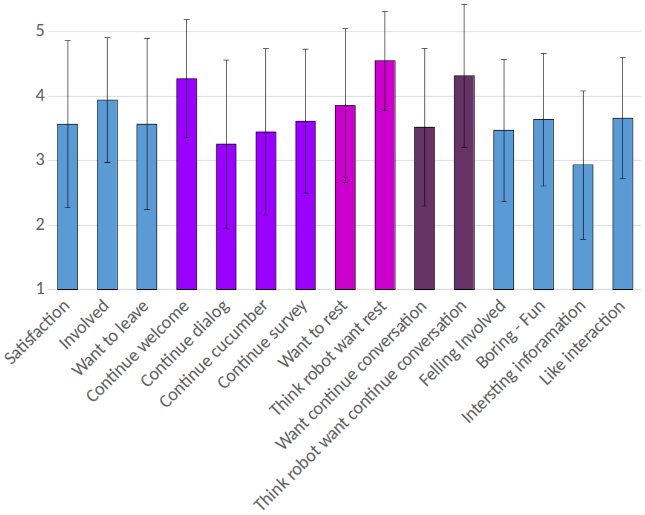


Figure 7: Average survey response over 53 participants. Error bars represent the standard deviation.

5 USE CASES

The UE-HRI dataset can be exploited in various research fields. In the following, we refer to some interesting applications that could take advantage of our dataset.

Early detection of engagement breakdown: Given the high number of users leaving before the end of the interaction, it would be interesting to analyse this phenomenon of *engagement breakdown* and try to forecast it as soon as possible, so as to consider re-engagement strategies that would prevent it. Higashinaka [12] *et al.* define a dialogue breakdown as *"a situation in a dialogue where users cannot proceed with the conversation"*. We define engagement breakdown as a failure to complete successfully the interaction with the robot and leaving before finishing it. *Early detection of engagement breakdown* is the prediction of a failure to finish an interaction with success and leaving before the end.

Affect burst recognition: Affect bursts could be an indicator of the emotional and affective state of the user. They were defined by Scherer [21] as "*very brief, discrete, nonverbal expressions of affect in both face and voice as triggered by clearly identifiable events*". Schröder [22] defined them as "*short, emotional non-speech expressions, comprising both clear non-speech sounds (e.g., laughter) and interjections with a phonemic structure (e.g., "Wow!"), but excluding "verbal" interjections that can occur as a different part of speech (like "Heaven!," "No!," etc.)*".

These affect burst are present in our dataset, because participants are in realistic situation, and showing them when enjoying the lively animation of the robot or when hating the curiosity of the robot. In order to construct a social relationship with humans as well as a long-term adaptation, our dataset could be used to detect user's affect burst.

Affect detection: It is challenging to detect automatically the user's affect. Affect detection would be helpful to design more natural interactions. Detection of frustration as defined by Barker [3] as "*any situation in which an obstacle - physical, social, conceptual or environmental - prevents the satisfaction of a desire*" was studied in natural human-computer dialog [1, 5].

Different affects observed in our corpus such as frustration, boredom, nervousness, disappointment. Our spontaneous dataset can be used for studying affect variation as well as developing and evaluating an affect detector.

6 CONCLUSION

In this paper, we have presented a new multimodal dataset of spontaneous interaction between humans and a robot, in which 54 interactions are freely available to download and use. The dataset provides a rich set of data streams that could be used by research and development groups in a variety of areas.

We believe that the provided dataset can be used to analyse and improve Human-Robot interaction. It could be used to develop and test a variety of modules in the robot architecture. For instance, this dataset is currently used by our research teams to develop a real-time early engagement breakdown detector based on user multimodal input data. Prediction of engagement breakdown ahead of time is an important step towards a smoother and more engaging human-robot interaction.

ACKNOWLEDGMENTS

The work is supported by the PSPC ROMEO2.

REFERENCES

[1] Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. 2002. Prosody-Based Automatic Detection Of Annoyance And Frustration In Human-Computer Dialog. In *in Proc. ICSLP 2002*. 2037–2040. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.4027

[2] Salvatore M. Anzalone, Sofiane Boucenna, Serena Ivaldi, and Mohamed Chetouani. 2015. Evaluating the Engagement with Social Robots. *International Journal of Social Robotics* 7, 4 (aug 2015), 465–478. https://doi.org/10.1007/s12369-015-0298-7

[3] ROGER G. BARKER. 1938. The Effect of Frustration upon the Cognitive Ability. *Journal of Personality* 7, 2 (dec 1938), 145–150. https://doi.org/10.1111/j.1467-6494.1938.tb02284.x

[4] Dan Bohus and Eric Horvitz. 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, New York, USA, 2–9. https://doi.org/10.1145/2663204.2663241

[5] Hynek Boril, Seyed Omid Sadjadi, Tristan Kleinschmidt, and John H L Hansen. 2010. Analysis and detection of cognitive load and frustration in drivers' speech. In *Proceedings of {INTERSPEECH}*. Makuhari Messe International Convention Complex, Chiba, Makuhari, Japan, 502–505.

[6] Nigel Bosch and Sidney D'Mello. 2015. The Affective Experience of Novice Computer Programmers. *International Journal of Artificial Intelligence in Education* (oct 2015), 1–26. https://doi.org/10.1007/s40593-015-0069-5

[7] P T Costa, R R McCrae, and Inc Psychological Assessment Resources. 1992. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Psychological Assessment Resources. https://books.google.co.in/books?id=mp3zNwAACAAJ

[8] John W Creswell. 2013. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

[9] Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2 (2012), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

[10] P. Dybala, M. Ptaszynski, R. Rzepka, and K. Araki. 2009. Activating humans with humor - A dialogue system that users want to interact with. *IEICE Transactions on Information and Systems* E92-D (2009), 2394–2401. Issue 12.

[11] Lynne Hall, Sarah Woods, Ruth Aylett, Lynne Newall, and Ana Paiva. 2005. Achieving empathic engagement through affective interaction with synthetic characters. In *Affective computing and intelligent interaction*. Springer, 731–738.

[12] Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

[13] Serena Ivaldi, Sebastien Lefort, Jan Peters, Mohamed Chetouani, Joelle Provasi, and Elisabetta Zibetti. 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 HumanâĂŞRobot Assembly Task. *International Journal of Social Robotics* 9, 1 (jan 2017), 63–86. https://doi.org/10.1007/s12369-016-0357-8

[14] Iolanda Leite, Marissa McCoy, Daniel Ullman, Nicole Salomons, and Brian Scassellati. 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, New York, USA, 99–105. https://doi.org/10.1145/2696454.2696466

[15] Marwa Mahmoud, Tadas Baltrušaitis, Peter Robinson, and Laurel Riek. 2011. 3D corpus of spontaneous complex mental states. In *Conference on Affective Computing and Intelligent Interaction*.

[16] Gary McKeown, Michel F. Valstar, Roderick Cowie, and Maja Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo*. IEEE, 1079–1084. https://doi.org/10.1109/ICME.2010.5583006

[17] Tatsuya Nomura, Takayuki Kanda, and Tomohiro Suzuki. Experimental investigation into influence of negative attitudes toward robots on humanâĂŞrobot interaction. (????). https://doi.org/10.1007/s00146-005-0012-7

[18] Christopher Peters, Ginevra Castellano, and Sara de Freitas. 2009. An exploration of user engagement in HCI. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*. ACM, 9.

[19] I Poggi. 2007. *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*. Weidler Buchverlag Berlin.

[20] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L. Sidner. 2010. Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 375–382. https://doi.org/10.1109/HRI.2010.5453163

[21] KR Scherer. 1994. Affect bursts. *Emotions: Essays on emotion theory* (1994). https://books.google.fr/books?hl=fr

[22] Marc Schröder. 2003. Experimental study of affect bursts. *Speech Communication* 40, 1 (2003), 99–116. https://doi.org/10.1016/S0167-6393(02)00078-X

[23] Candace L. Sidner, Christopher Lee, Cory D. Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166, 1-2 (aug 2005), 140–164. https://doi.org/10.1016/j.artint.2005.03.005

[24] Dominique Vaufreydaz, Wafa Johal, and Claudine Combe. 2016. Starting engagement detection towards a companion robot using multimodal features. *Robotics and Autonomous Systems* 75 (jan 2016), 4–16. https://doi.org/10.1016/j.robot.2015.01.004

[25] Johannes Wienke, David Klotz, and Sebastian Wrede. 2012. A Framework for the Acquisition of Multimodal Human-Robot Interaction Data Sets with a Whole-System Perspective. *LREC 2012 Workshop on Multimodal Corpora for Machine Learning* (2012). https://pub.uni-bielefeld.de/publication/2487085