Multifaceted Engagement in Social Interaction with a Machine: the JOKER Project

Laurence Devillers, Sophie Rosset,
Guillaume Dubuisson Duplessis, Lucile Béchade
LIMSI, CNRS, Université Paris-Saclay
Orsay, France
devil@limsi.fr

Yücel Yemez, Bekir B. Türker,
Metin Sezgin, Engin Erzin
Koç University
Istanbul, Turkey
bturker13@ku.edu.tr

Kevin El Haddad, Stéphane Dupont
University of Mons
Mons, Belgium
kevin.elhaddad@umons.ac.be

Paul Deléglise, Yannick Estève, Carole Lailler
LIUM, Le Mans University
Le Mans, France
yannick.esteve@univ-lemans.fr

Emer Gilmartin, Nick Campbell
Trinity College Dublin
Dublin, Ireland
nick@tcd.ie

***Abstract*—This paper addresses the problem of evaluating engagement of the human participant by combining verbal and nonverbal behaviour along with contextual information. This study will be carried out through four different corpora. Four different systems designed to explore essential and complementary aspects of the JOKER system in terms of paralinguistic/linguistic inputs were used for the data collection. An annotation scheme dedicated to the labeling of verbal and non-verbal behavior have been designed. From our experiment, engagement in HRI should be multifaceted.**

***Keywords*-Human-Robot Interaction; Dataset; Engagement; Speech Recognition; Affective Computing**

I. Introduction:

The JOKER project aims to build a generic intelligent user interface providing a multimodal dialogue system with social communication skills including humour and other social behaviours [1]. One of our main objectives is to study and measure the user-robot relationship as well as user engagement by combining verbal and nonverbal behaviour and contextual information. In this direction, recent work in Human-Robot interaction (HRI) intends to recognise and quantify human engagement in dialogue in order to adapt the behaviour of the robot. Engagement can be defined as "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake" [2], [3]. Engagement process involves nonverbal and verbal behaviours, as well as low-level processes (such as behaviour synchrony, mimetics) and high-level cognitive processes (such as answering a riddle).

From our experiment, engagement in HRI should be multifaceted. To that purpose, we propose to evaluate human engagement by combining verbal and nonverbal behaviour along with contextual information. In this paper, after discussing work related to engagement (Section II) and data collection (Section III), we describe an annotation scheme (Section IV) and first analyses of the annotations in order to investigate measures of engagement (Section V). Results presented in this paper are exemplified on a corpus collected in a cafeteria at LIMSI and on a corpus collected at TCD setting as part of the JOKER project.

II. Multifacet Engagement

In the context of collaborative task-oriented interaction between a human and a robot, [3] have identified four types of connection event (directed gaze, mutual facial gaze, delay in adjacency pair, and backchannel) involved in the computation of statistics on the overall engagement process. Our focus is on social dialogue rather than task-oriented one. We consider the interpretation of cues in the context of a dialogue act. Interestingly, cues such as the four types of connection event could be integrated to our approach. We can also consider engagement 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" [4]. In order to favour the user's engagement level other previous research manipulated the agent's non-verbal behaviour including gestures, gaze and facial displays [5], as well as the agent's verbal behaviour including the form [6] and prosody [7].

III. JOKER Data Collection

Four different corpora were collected [1], [8] and annotated at different levels (see Section IV) at LIMSI. Four different systems were used for the data collection (two fully automatic system, one semi-automatic and one Wizard of Oz).

These systems were designed to explore essential and complementary aspects of the JOKER system in terms of paralinguistic/linguistic inputs. The first system, called the *paralinguistic system*, involves a social interaction dialog. Its objective is to tell riddles to the user and adapt this telling to some aspects of the user model. It is fully automatic and features an emotion detection module based on audio [9], a dynamic user model [10], and a finite-state based dialogue manager. The user model is based on user's attitude towards the robot (*interactional representation*) and user's affective tendencies in the course of the interaction (*emotional representation*).

The second system, called the *linguistic system*, offers a *discover my favorite dish* challenge to the user. By asking questions to the robot, the participant has to find out a recipe. It includes a question-answering system adapted to the culinary challenge similar to the open-domain dialogue system Ritel [11], a natural language generation system, and a database of recipes and ingredients automatically crawled from the web. For this system, speech recognition has been carried out manually by a human operator who typed the utterances produced by the participant.

The third system, called the *emotion challenge system*, offers an *emotion* challenge to the user. Nao asks the participants to play 4

978-1-5386-2335-0/18/$31.00

emotions : joy, anger, sadness and neutra state.

The fourth system is a *Wizard of Oz* that is based on a graphic user interface remotely controlled by a human operator. A predefined dialog tree specifies the text utterances, gestures and laughter that can be selected by the human operator to be executed by Nao. At each node, the operator chooses the next node of dialogue to visit according to the human dialogue participant's reaction.

Table I contains an example of the data collected with each of those systems.

Thirty-seven participants (62% male, 38% female) were recorded. Each of them interact with all the systems. The average age of the participants is 35.1 (standard deviation: 11.97; min: 21; max: 62). All were volunteers working at the LIMSI laboratory, and French speaking. The total duration of the data recorded with each of the systems is shown in Table III.

IV. Corpus Annotation Procedure

We define an annotation scheme dedicated to the labeling of verbal and non-verbal behavior produced by the human participant while interacting with the robot. The annotation scheme can be divided into the dimensions of: neutral, discourse, emotional and multimodale. The neutral dimension annotate a relative neutral "position" for a given participant, to have a capture of the human without emotions nor engagement. The discourse dimension contains the annotation of contextual reactions of participant to the task (EventAct) and the annotation of dialogue act adapted from [12](DAct). The emotional dimension is annotated with classic aspects coming from speech-based emotion detection systems [10]: activation behavior (EmotionAct); emotion labels (FeelingAct) and mental states (FeelingAct). Finally, the multimodal dimension contains annotation of laughter (LaughAct), backchannels (BackchannelsAct), mouth movements of smiling (MouthAct), head gestures (HeadAct) and the contextual reactions of participants to the robot's humor (HumorAct). All these labels can be verbal or non-verbal. Humor labels describe the contextual human response to a humorous intervention from the robot viewed as the second part of an adjencency pair of humorous act and humor response [8]. Laugh labels describe the intention disseminated by the participant laugh. Backchannels can be verbal or non-verbal expressions such as nodding, gazing or minimal responses, reactive tokens and continuers [13]. The ELAN annotation tool[1] was used as it allows for complex and multilayer annotation for video and audio data. The annotation schema is multi-layers. The different layers concern discourse and multimodality. In total thirteen layers are defined, all of them having a controlled vocabulary. Table III describes the annotation schema.

V. Engagement Analysis

A. Statistics

For each of the 13 layers, the annotation procedure gives the duration of each annotation type (for example, the duration of each *Move* for the layer *HeadAct*). Table IV presents these basic statistics.

On Figure 1 one can observe clear differences concerning the usage of acts (i.e. the annotation layer) given the kind of interaction.

[1]https://tla.mpi.nl/tools/tla-tools/elan/

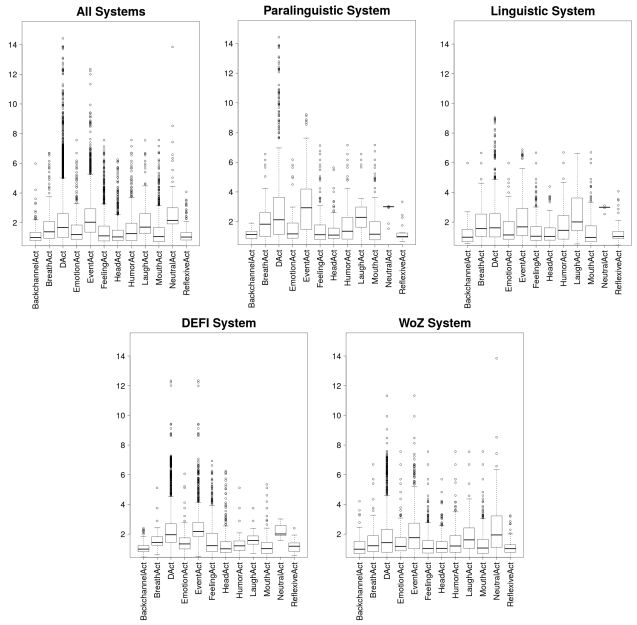

Figure 1. Durations of Annotation Layers

Also, the variation is important. Table V shows the amount of different nonverbal expressions, in this database, which were reported to occur frequently in human-human dyadic interactions: laughter, smiles, head and eyebrow movements can frequently be found in human-human interactions. These values are reported per system. We also study the occurrences of these expressions with respect to the agents utterances. In Table VI, we show the average amount of times each of the previously mentioned expressions occurred at least once during an entire recording session (and per system). And this while the agent is still speaking (upper value in each cell) and between one sentence from the agent and another, i.e. while the participant is reacting (lower value in each cell). For example this table shows that the participants laughed at least once on average 7.7% of the time while the agent was still speaking and 2.6% of the time while they were reacting in the *paralinguistic system*. Calculating these values relayed on the transcriptions (verbal) layer of the annotation scheme. For technical reasons, the transcriptions were not available at the writing time of this paper for the *Emotion Challenge system*. That is why we report values only related to the other systems. The results show that, except for the Nodding-Neg expressions and for the Linguistic system, all the expressions occurred while the agent was still speaking rather than after for the Paralinguistic and the Linguistic systems. This might be due to the nature of the scenarios in each system and the timing (automatic or not) between the speaking turns between the participant and the agent. We can also see that, for each utterance from the agents, the participants produced these expressions, in general and on average, more during the *paralinguistic system* than the other two systems and more during the *linguistic system* than the *WoZ* system. An interesting point we could notice is, the more automatic the system is, the more often the expressions tend to occur. This might of course be due to the nature of the interactions but it could also be due to the fact that the agent acts "less human" when "more free". Indeed the agent acting more freely might produce more

	Example
Paraling. syst.	S: Tu as l'air triste, je vais te remonter le moral avec une devinette. Comment appelles-t-on un chien sans pattes ? U: Bin je sais pas. Chien sans pattes ? S: J'ai détécté une réaction négative à ma devinette. La réponse était, on ne l'appelle pas, on va le chercher.
Ling. syst.	S: Nao:Bonjour ! Je suis Joker, et j'aime bien discuter. J'ai un petit jeu pour toi ! Devine mon plat préféré ! H: alors bonjour Joker euh peut-être que ça va être le plat de spaghetti S: Je n'ai jamais gouté ça. Propose moi autre chose ! H: euh du riz basmati
Emo. Challenge syst.	S: Je vais maintenant te demander de jouer des émotions afin de tester mes capacités. C'est parti ! Parle-moi en ayant l'air joyeux. U: Oh bah là c'est pas difficile hein ! Je suis, je suis plutôt joyeux hein ! S: J'ai détécté de la joie. Super, tu as réussi ! On continue.
WoZ syst.	S:Bon, je connais aussi des devinettes laisse-moi t'en raconter une. U: vas-y je t'écoute. S: Comment ramasse-t-on la papaye ? U: Euh avec les mains. S: La réponse était avec une foufourche! U: Elle est bonne. Tu en as d'autres ?

Table I
EXAMPLES COLLECTED WITH DIFFERENT SYSTEMS.

	Paraling. Syst.	Ling. Syst.	Emo. Challenge Syst.	WoZ Syst.	Total
Duration	1h 30min 5s	1h 25m 20s	1h 42m 27s	3h 20m 57s	7h 58m 50s
average	2min 26s	2m 18s	2min 46s	5m 25s	3m 14s
σ	14s	55s	32s	1m 00s	1m 28s

Table II
DURATION OF THE COLLECTED DATA PER SYSTEMS. σ IS THE STANDARD DEVIATION.

odd actions from it and thus more nonverbal expressions from the participants.

B. Engagement measures

The reaction to humor, or humor support, is important to show the understanding and appreciation of a joke. It is a way of showing involvement in a discussion. Hay and Bell [14], [15] pointed out that there are many different humor support strategies in verbal or non-verbal behaviour of participants. Using the annotation scheme of humor reactions and all other dimensions allows us to have cues for humor support in the non-verbal behavior. Theses cues are used to find objective metrics to predict the user satisfaction and engagement in the humorous topic.
Furthermore, we analyze the human-robot interaction experiments of JOKER dataset in terms of engagement measurement [16]. There are two distinct experiments in which autonomous and wizard-of-oz (WoZ) setups are used. Engagement measures are extracted for both setups. We take the engagement level of WoZ setup experiments as a gold standard in JOKER dataset. We evaluate the autonomous setup experiments by comparing with the gold standard engagement measurements as in [17].

C. Lexical analysis and automatic speech recognition

Speech collected from the human participants in the different system has been manually transcribed and these transcriptions are a part of the annotations. These transcriptions of human speech contain 10,784 word occurrences of 1,125 different words. In comparison, the robot uses a vocabulary of 516 words for 6,755 word occurrences. The intersection between the vocabulary of humans and the robot vocabulary is made of 378 words, showing that only 33.6% of words uttered by a human were also used by the robot. In addition to these manual transcriptions of human speech, automatic transcriptions will be also added to the JOKER data. The automatic speech recognition used to transcribe these data is based on the Kaldi toolkit, using chain TDNN acoustic models [18] trained on more than 500 hours of speech in French. It is a French variant of the LIUM ASR systems that ranked second during the Multi-Genre Broadcast Challenge 2016 in Arabic [19] and the Multi-Genre Broadcast Challenge 2015 in English [20]. The language model has been trained from data crawled on websites dedicated to cooking, since the *linguistic system* is based on a culinary challenge while the other systems are also strongly influenced by this topic. Two kinds of text corpora were collected: comments and recipes, which one was used to estimate a single language model respectively LM_c and LM_r. Since no specialized data were already available for spoken human/robot interactions with humor before the JOKER project, these crawled text data dedicated to cooking are one of the most close data that we could use. We split the entire French JOKER data into a development corpus (containing $\frac{1}{3}$ of the data) and a test one (the other $\frac{2}{3}$). By this way, we were able to optimize the value of the linear interpolation of LM_c and LM_r on the development data. The final language model contains around 70K words, and the out-of-vocabulary rate is 1.49%. On the test data, the perplexity value was 253, which is a high value that indicates the difficulty of the language modeling for this ASR task. The automatic transcriptions will be available very soon, with confidence measures, but also word-lattices and confusion networks, allowing researchers who

Layer	Definition	Vocabulary
NeutralAct	Annotate relative neutral "position" for a given participant.	-
DAct	Annotate all dialog acts related to past utterances.	SOM, AutoFeedback, AlloFeedback, ContactManagement, TimeManagement, Event, TurnManagement, DiscourseSM
EventAct	Reactions to the task.	CookQuestion, CookIgnore, CookAnswer, JokeNao, JokeAnswer, JokeIgnore, Behavior
HeadAct	Dedicated to the head's moves	EyeBrow, Nodding-Neg, Nodding-Pos, Move
MouthAct	Dedicated to the Mouth's moves	SmileOnPOS, SmileOffPOS, SmileOnNEG, SmileOffNEG
HumorAct	Annotate the human reaction "in context"	Humor, Like, Dislike, Sarcasm,
EmotionAct	Notify the engagement of the Human Being in front of Nao	Active, Passive
FeelingAct	Notify a strong feeling or emotion in front of Nao	Surprise-Para, Surprise-Ling, Surprise, Sadness-Para, Sadness-Ling, Sadness, Joy-Para, Joy-ling, Joy, Doubt-Para, Doubt-Ling, Doubt, Angry-para, Angry-Ling, Angry, Contemp-Para, Contemp-Ling, Contemp, Pride-Para Pride-Ling, Pride, Disap-Paragraph, Disap-Ling, Disap Awar-Para, Awar-Ling, Awar, POS-Ling, POS-Para, POS NEG-Ling, NEG-Para,NEG
LaughtAct	Annotate laughter from the Human	Embarasment, Amused, Sarcastic, Politeness, Relief Non-Understanding
BreathAct		BreathPOS, BreathNEG
ReflexiveAct	Annotate all feedbacks of the Human in a paralinguistic way.	Robot, Pers, Other, Situation
BackchannelAct	Notify how to maintain the channel of the attention/engagement.	-

Table III
ANNOTATION SCHEMA: A BRIEF DESCRIPTION.

Systems	Time.m	Time.s
Paraling.	2.309148	1.952638
Ling.	1.844974	1.278916
DEFI	2.132364	1.451157
WoZ	1.702166	1.268233

Table IV
DURATION GIVEN THE CORPUS OF EACH EVENT. TIME.M STANDS FOR THE MEAN DUREATION, TIME.S GIVES THE STANDARD DEVIATION.

Expressions	Emo Chal.	Woz.	Ling.	Paraling.
Laughs	115	228	75	98
SmileOnPos	13	13	11	22
Nodding-Pos	59	81	18	42
Nodding-Neg	15	27	11	22
Move	141	137	74	72
EyeBrow	99	63	37	38

Table V
NUMBER OF SOME NONVERBAL EXPRESSIONS PER SYSTEM. LAUGHS BEING THE CONCURRENCE OF SMILEONPOS AND BREATHONPOS.

Expressions	Woz.	Ling.	Paraling.
Laughs	2.4%	6.9%	7.7%
	4.2%	3.1%	2.6%
SmileOnPos (Smiles)	2.9%	5%	16.1%
	4%	3.5%	8.13%
Nodding-Pos	2%	4.6%	10.3%
	3.4%	1.5%	3.7%
Nodding-Neg	0.2%	0.3%	3.6%
	0.8%	1.9%	2.7%
Move	3.3%	9.15%	15.8%
	5.9%	8.4%	7.5%
EyeBrow	1.6%	4%	8.5%
	2.7%	2.4%	3.3%

Table VI
AVERAGE PERCENTAGE OF TIME, NONVERBAL EXPRESSIONS OCCURRED AT LEAST ONCE PER RECORDING SESSION. IN EACH CELL, UPPER VALUE SHOWS THE OCCURRENCES WHILE THE AGENT IS STILL TALKING AND THE LOWER ONE WHILE THE PARTICIPANT IS REACTING.

cannot develop their ASR system to deal with different kinds of ASR outputs.

VI. ACKNOWLEDGMENTS

The described work has been sponsored by ERA-Net CHIST-ERA (http://www.chistera.eu/), the "Agence Nationale pour la Recherche" (ANR, France), the "Fonds National de la Recherche Scientifique" (FNRS, Belgium), the "The Scientific and Technological Research Council of Turkey" (TUBITAK, Turkey, grant number 113E324) and the "Irish Research Council" (IRC, Ireland).

REFERENCES

[1] L. Devillers, S. Rosset, G. Dubuisson Duplessis, M. A. Sehili, L. Béchade, A. Delaborde, C. Gossart, V. Letard, F. Yang, Y. Yemez, B. B. Türker, M. Sezgin, K. El Haddad, S. Dupont, D. Luzzati, Y. Estève, E. Gilmartin, and C. Nick, "Multimodal data collection of human-robot humorous interactions in the joker project," in *ACII*, Xi'an, China, September 2015.

[2] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1, pp. 140–164, 2005.

[3] C. Rich and C. L. Sidner, "Collaborative discourse, engagement and always-on relational agents." in *AAAI Fall Symposium: Dialog with Robots*, 2010.

[4] I. Poggi, *Mind, hands, face and body. A goal and belief view of multimodal communication*. Weidler, 2007.

[5] D. Bohus and E. Horvitz, "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*. Association for Computational Linguistics, 2009, pp. 225–234.

[6] N. Glas and C. Pelachaud, "User engagement and preferences in information-giving chat with virtual agents," in *Workshop on Engagement in Social Intelligent Virtual Agents (ESIVA)*, 2015, pp. 33–40.

[7] G. Dubuisson Duplessis and L. Devillers, "Towards the Consideration of Dialogue Activities in Engagement Measures for Human-Robot Social Interaction," in *International Conference on Intelligent Robots and Systems*, ser. Designing & Evaluating Social Robots for Public Settings Workshop, Hambourg, Germany, Sep. 2015, pp. 19–24. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01206675

[8] L. Bechade, G. Dubuisson-Duplessis, and L. Devillers, "Empirical study of humor support in social human-robot interaction," in *Streitz N., Markopoulos P. (eds) Distributed, Ambient and Pervasive Interactions. DAPI 2016.*, 2016.

[9] L. Devillers, M. Tahon, M. A. Sehili, and A. Delaborde, "Inference of human beings' emotional states from speech in human-robot interactions," *International Journal of Social Robotics*, pp. 1–13, 2015.

[10] A. Delaborde and L. Devillers, "Use of nonverbal speech cues in social interaction between human and robot: Emotional and interactional markers," in *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments*, ser. AFFINE '10. New York, NY, USA: ACM, 2010, pp. 75–80.

[11] B. van Schooten, S. Rosset, O. Galibert, A. Max, R. op den Akker, and G. Illouz, "Handling speech input in the Ritel QA dialogue system," in *InterSpeech'07*, Antwerp, Belgium, 2007.

[12] H. Bunt, "The DIT++ taxonomy for functional dialogue markup," in *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, 2009, pp. 13–24.

[13] R. Gardner, *When Listeners Talk: Response Tokens and Listener Stance*. Amsterdam: J. Benjamins Publishing, 2001.

[14] N. D. Bell, "Responses to failed humor," *Journal of Pragmatics*, vol. 41, pp. 1825–1836, 2009.

[15] J. Hay, "The pragmatics of humor support," *Humor – International Journal of Humor Research*, vol. 14, no. 1, pp. 55–82, 2001.

[16] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human robot interaction," *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 375–382, MArch 2010.

[17] B. B. Türker, Z. Buçinca, E. Erzin, Y. Yemez, and M. T. Sezgin, "Analysis of engagement and user experience with a laughter responsive social robot," in *18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, 2017.

[18] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] N. Tomashenko, K. Vythelingum, A. Rousseau, and Y. Estève, "LIUM ASR systems for the 2016 Multi-Genre Broadcast Arabic challenge," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 285–291.

[20] V. Gupta, P. Deléglise, G. Boulianne, Y. Estève, S. Meignier, and A. Rousseau, "CRIM and LIUM approaches for multi-genre broadcast media transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 681–686.