Towards metrics of Evaluation of Pepper robot as a Social Companion for Elderly People

Mélanie Garcia, Lucile Béchade, Guillaume Dubuisson-Duplessis, Gabrielle Pittaro, Laurence Devillers

Abstract

For the design of socially acceptable robots, field studies in Human-Robot Interaction are necessary. Constructing dialogue benchmarks can have a meaning only if researchers take into account the evaluation of robot, human, and their interaction. This paper describes a study aiming at finding an objective evaluation procedure of the dialogue with a social robot. The goal is to build an empathic robot (JOKER project) and it focuses on elderly people, the end-users expected by ROMEO2 project. The authors carried out three experimental sessions. The first time, the robot was NAO, and it was with a Wizard of Oz (emotions were entered manually by experimenters as inputs to the program). The other times, the robot was Pepper, and it was totally autonomous (automatic detection of emotions and decision according to). Each interaction involved various scenarios dealing with emotion recognition, humor, negotiation and cultural quiz. The paper details the system functioning, the scenarios and the evaluation of the experiments.

Mélanie Garcia
LIMSI-CNRS, Université Paris-Saclay, Orsay, France, e-mail: garcia@limsi.fr

Lucile Béchade
LIMSI-CNRS, Université Paris-Saclay, Orsay, France e-mail: lucile.bechade@limsi.fr

Guillaume Dubuisson-Duplessis
LIMSI-CNRS, Université Paris-Saclay, Orsay, France e-mail: gdubuisson@limsi.fr

Gabrielle Pittaro
LIMSI-CNRS, Université Paris-Saclay, Orsay, France e-mail: gabrielle.pittaro@univ-sorbonne.fr

Laurence Devillers
LIMSI-CNRS, Université Paris-Sorbonne, Orsay, France e-mail: devil@limsi.fr

1 Introduction

Currently, National and International teams work on projects for the elderly self-sufficiency [10] [11] [7], particularly on conversational agents design [14] [1]. To build a coherent and engaging conversational agent, social dialogue is essential.
An autonomous robot with clever perceptual analysis is more engaging for the user. Furthermore, it may lead to personalize relationship with the user [3]. Empathy may help a lot in the analysis and decision of answer tasks. The purpose of JOKER (JOKe and Emphathy of a Robot) project is to give a robot such a capability, as well as humor.
Besides, regarding human-robot dialogue, neither a clear common framework on social dialogue nor a procedure of evaluation exist. [2] tried to find metrics so as to better evaluate Human-Robot Interaction (HRI).
In this paper, the authors introduce a study as part of the ROMEO2 project. They explain the context, the system description, the proceedings of three HRI experimental sessions carried out with elderly people. They aim at evaluating objectively their multi-modal system in order to build a real personal robot assistant for elderly people.

2 Related Work

In Human-Robot Interaction, researches have focused on elderly users, and on the robot's ability to help the participant: to stand [17], to catch something [8] or to walk [16].
To maintain user engagement and increase acceptability of robot, social dialogue is crucial. Indeed, regarding media interaction, Reeves and Nass [13] emphasize that people react to media as if they were social actors. Several works address the issue of evaluating Human-Robot spoken interactions in a social context by considering the engagement of the human participant [6]. In assistive and social robotics, experiments with potential end-users provide a valuable feedback about researchers' expectations, and reliable data for the design of socially acceptable robots. Moreover, user feedback may help to improve the evaluation and the development of dialogue system. In that regard, [15] worked on an evaluation plan based on incremental stages corresponding to the improvement of a dialogue system according to user feedback.

3 Context

3.1 JOKER project

Social interactions require social intelligence and understanding: anticipating the mental state of another person may help to deal with new circumstances. JOKER researchers investigate humor in human-machine interaction. Humor can trigger surprise, amusement, or irritation if it does not match user's expectations. They also explore two social behaviors: expressing empathy and exchanging chat with the interlocutor as a way to build a deeper relationship.
The project gathers 6 international partner laboratories. LIMSI involvement is on affective and social dimensions in spoken interaction, emotion and affect bursts detection, user models, Human-Robot Interaction, dialogue, generation.

3.2 ROMEO2 project

Aldebaran launched ROMEO2 project with the objective to build and develop a personal robot companion for the elderly [9]. On the one hand, ROMEO will be capable of moving, take items and give simple information to the user. On the other hand, it will be able to think, to reason, in order to detect extraordinary situations, to decide, to adapt its answers. The multi-modality goal with multi-sensory perception and cognitive interaction (reasoning, planning, learning mechanisms) makes this project unique. Indeed, the robot ROMEO will be able to assist old people and to answer the best it can to their requests using both a predefined database and a learned database through its experiences. ROMEO will learn from its everyday life with the user and will be able to adapt and personalize its behavior. Furthermore, its capacity to detect emotions from the user will make its process decide the best behavior to adopt during a dialog.
LIMSI involvement is on emotion recognition, multimodal speaker identification, speech comprehension and social interaction.

3.3 Broca Hospital

LUSAGE Living Lab [12] at the Broca hospital, Paris, welcomed the experience, under the supervision of the gerontology service. Regularly, the Living Lab organizes workshops in the "Café Multimdia" project. Most of elderly people are badly aware of digital technologies, becoming more and more socially isolated. The goal of the project is to bridge that divide. The participants can discover the Information and communication Technologies, discuss them, and meet designers and researchers.
The authors of this study took part in the activities of the Lab: workshop on social

robotics for health-care and everyday life. On these occasions, they offered the participants to interact with a robot. After each experiment, the researchers discussed (individually and in group) with them, about the experiment, but also about their opinions on social and assistive robotics in general.

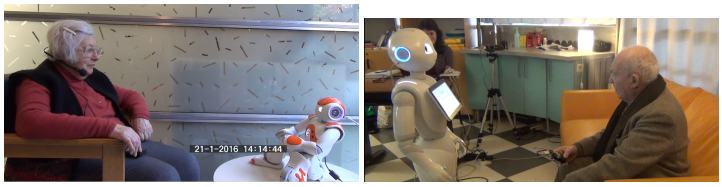

Fig. 1 Pictures taken during interactions between: an elderly and Nao (on the left), an elderly and Pepper (on the right).

4 System description

The system during the two last experimental sessions was totally autonomous. Each module of analysis of linguistic and paralinguistic could communicate through a multi-modal platform. The multi-modal platform enables to make the link between paralinguistic module and decision process while running the scenarios. Several levels of abstraction constitute the system, that allow to build a dialogue system on top of the crossbar architecture. From the lowest level to the highest level, the system uses:

- **The Event**: technical messages exchanged by crossbar components, can be asynchronous or synchronous
- **The Contribution**: a dialogic contribution within a dialogue turn which aggregate data from the input modules (e.g., linguistics, paralinguistics)
- **The Expectations**: expectations defined at the scenario level (e.g., an emotionally positive contribution, a given word, a silence, etc.)

After most Pepper intervention, the multi-modal platform requests a contribution from the user. It takes into account : a minimal time to wait for a contribution (if something is said, Pepper will stop listening after that time) and a maximal time to wait for a contribution (if nothing is said at all, Pepper will stop listening at that point). The multi-modal platform works according to the architecture shown in 2. The paralinguistic system features an emotion detection module based on audio [5]. The audio signal is cut into segments between 200 and 1600 milliseconds. Each segment contains or not a detected emotion. The robot takes into account the majority emotion during a speech turn. The emotion recognition module works with a linear Support Vector Machines (SVM) with data normalization and acoustic descriptors such as acoustic parameters in the frequency domain (e.g. fundamental frequency

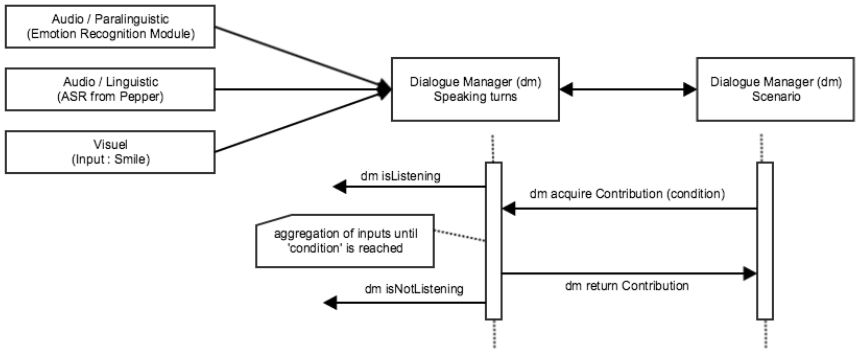

Fig. 2 Multi-modal platform system for social dialogue with Pepper

F0), in the amplitude domain (e.g. energy), in the time domain (e.g. rhythm) and in the spectral domain (e.g. spectral envelop or energy per spectral bands).

5 Experiment

5.1 Proceedings

First of all, the researchers gave to participants a general and collective explanation of the experiment: the aims of the researchers, the type of interaction and the nature of the robot they were going to meet. They carried out three experimental sessions within the same context. In the first session, the Humor system was a Wizard of Oz (totally operated by a human experimenter) with NAO. The experimenter provided a part of the inputs manually (emotion detection). In the second and third ones, all the system was autonomous. The authors used the robot Pepper. Each participant of the first and second session interacted with the robot only once while those of the third between one and four times. The comparison between the experiments is shown in Figure 3.

	1st Experiment	2nd Experiment	3rd Experiment
Number of participants	12	8	22
whose hearing elderly participants	12	4	16
Number of interaction per person	1	1	1 to 4
Robot	NAO	PEPPER	PEPPER
Type of system	Wizard of Oz	Autonomous	Autonomous
Scenarios	Humor, Negotiation	Emotion game, Humor, Negotiation, Quiz	Emotion game, Humor, Negotiation, Quiz

Fig. 3 Table comparing the experiments

5.2 Scenarios

The scenarios tested relate to possible daily life interactions between the robot companion and an elderly person.

- **Emotions**: A first scenario consists in asking the user to mimic chosen emotions while speaking. Pepper (the robot) asked the user to imitate the four emotions it can detect: Joy, Sadness, Anger, Neutral. If the speech from the user contains the requested emotion, the robot goes on following the rest of the scenario. If it does not, the robot says which majority emotion it detects and asks again the user to mimic (over 3 times, the robot stop asking and proceeds to the next question). This scenario enables to evaluate the emotion recognition system performance.
- **Jokes/Riddles**: Pepper can make several riddles and puns. For instance:
  - Question: "What is a cow making while closing its eyes?"
    Answer: "Concentrated milk!"
  - Question: "How do we call a dog without legs?"
    Answer: "We do not call it, we pick it up."

  During the experiment, according to the emotion detection on the user, Pepper adapted its humor to the user emotion profile. Pepper may also stimulate the memory of the user asking to repeat one joke he made. The paralinguistic module takes an important part in the humor process: it detects emotions, laugh. The behavior of the system depends on the receptiveness of the human to the humorous contributions of the robot. Positive reactions (e.g. laughter, positive comments or positive emotions) lead to more humorous contributions, whereas repeated negative reactions (e.g. sarcastic laughter, negative comments and negative emotions) drive the dialogue to a rapid end. If there is no reaction, the robot tries to change its kind of humor so as to make the user react.
- **Persuasion/Negotiation**: The robot as a companion has to take care of the user showing initiative. In this experiment, Pepper tried to convince old people to drink a glass of water. Before the simulation, the user was said to always refuse the proposition from Pepper. According to the user global reaction valence detected by the system, the robot chose a negotiation strategy: Humor (the robot makes derisive comments about itself so as to make the user accept its offer), Reason (the robot argues reasonably), Calming (the robot ensures it does not want to force the user after detecting anger). The robot calculates the global valence of reactions (positive or negative) from the user. Then, it can adapt its strategy of negotiation.
- **Cultural Quiz**: The robot makes the user listen to extract of music or movies, and asks the user to recognize the singer or actor, or the name of the song or the movie.

6 Results

6.1 Evaluation of system performance

The emotion recognition algorithm was built with the annotated audio corpus JEMO [4]. The performance of the emotion algorithm calculated by cross-validation is described by a F-score equal to 62,4.
Old people assessed the robot had difficulties to detect their emotions correctly. To evaluate the performance of our system on old people, the authors annotated with emotion labels (anger/sadness/joy/neutral) all the segments where the robot detected an emotion. The best detection performance was one correct detection over two. The worse was one over eight. The annotators noticed that sometimes segments were too short to recognize an emotion correctly. Figure 4 shows that good detections occur mostly for segment lengths higher than one second.
The corpus JEMO contains voices from people aged between 20 and 50 years old. The experiments participants were between 70 and 85. A higher speech rate (for set segment length) has been observed in the corpus JEMO than in the experimental corpus. This raises the problem on speech velocity variability. Therefore, it is necessary to take into account this feature while learning emotion detection algorithm, so to take into account the feature ”age” implicitly. Thanks to the data, the researchers will build a new emotion detection algorithm adapted to the elderly in future work.

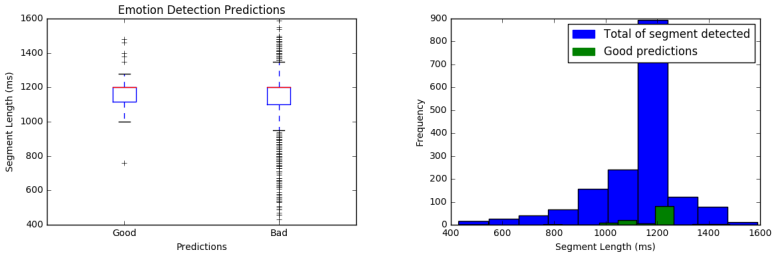

Fig. 4 Performance of the Emotion Detection on the Elderly: type of prediction boxplot and histogram according to Segment Lengths (in milliseconds). Annotation errors may put a bias.

6.2 Engagement metrics

User engagement is a precious piece of information. Being able to detect it, the robot could adapt its social behavior. Thus, engagement would be an essential feature to build a dialog adaptive algorithm. Indeed, the authors wondered if the more a user is involved in the interaction, the more it talks. Moreover, if a user reacts relatively

quickly, does it mean it is sensitive (positively or negatively) to what the robot is saying?
After each interaction, the user had to fill a questionnaire. This was about user's feelings during the interaction and global view on the interaction and on robots. In this section, answers related to engagement are studied: did you feel involved in the interaction? Did you feel self-confident?
The authors assumption is that engagement can be seen in three metrics: reaction time, silence time and speaking time during one speech turn of the user. To start a validation on that hypothesis, the authors compare these metrics to reliable information on engagement. Figure 5 looks to highlight links between speaking time and

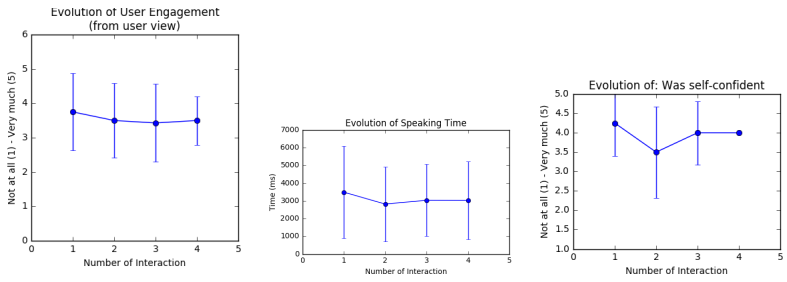

Fig. 5 Evolution of User Engagement (self-evaluation), Speaking Time (individual means of time speaking at each speech turn) and of User Self-Confidence according to the Number of Interaction. Graphs represent means and standard-deviation.Mean curves look to follow similar tendencies.

user engagement, and between speaking time and user self-confidence. Correlations between means of these features are respectively 0,866 and 0,881. Mean curves follow similar tendencies. The authors remind about the different sample sizes for the feature "Number of Interaction": 16 for the level one, 10 for the level two, 7 for the level three, 2 for the level four. Therefore, the hypothesis about speaking time as an engagement metric has to be validated with bigger samples in the future.
The engagement of the user may be explained by two variables: the understandability of the task requested (the user understands what the robot expects him to do so he can act spontaneously), and the attractiveness of the scenario (the scenario inspires the user who can answer faster than if it does not). Adding the metric success to the Emotion Challenge and to the Quizz may also help to distinguish those who played the game from those who did not. Furthermore in a next work, a metric about user engagement during the humor scenario will be studied.
Figure 6 may also show interesting link between silence time and robot dominance evaluation. If silence time represented a reliable metric of robot dominance, robot could adjust its behavior appearing more humble for the user.
Thanks to this short longitudinal session and other experiments planned next, the authors expect to build new strategies on engagement of the user during Human-Machine Interaction adapted to the elderly.

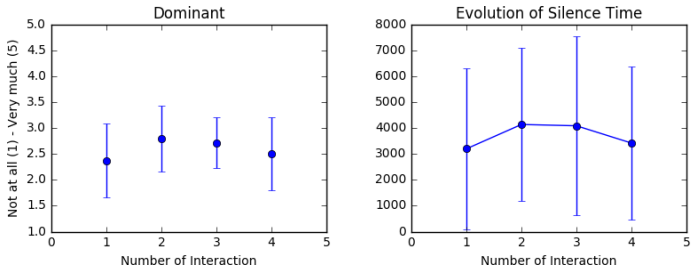

Fig. 6 Evolution of Robot Dominance (user evaluation) and Silence Time according to the Number of Interaction (individual means of silence time at each speech turn). Graphs represent means and standard-deviation. Mean curves look to follow similar tendencies.

6.3 Interaction Appraisal

The researchers took recommendations from the users through questionnaire about their opinion on the robot operating, their feelings during the interaction and their thoughts about such a technology.
In the first experimental session, the experimenters entered manually each user emotional expressions. The users liked interacting with the robot. Thus, the researchers could study the acceptability of robot as a vector of communication for the elderly. However, users were different between each experimental session, and sample size was the smallest in the first one. It is not possible to draw conclusions about the impact of the autonomy of robot.
Regarding the second and third experimental sessions, the robot detected correctly most of the elderly moods (positive or negative attitude in reactions), thanks to the summarization of emotions detected in speech turns. Thus, the robot could adapt its answers correctly. Sometimes, it failed at recognize the speech from the user and then gave incoherent answer. Figure 7 shows an example of an interaction which functioned correctly with a user while Figure 8 shows one of an interaction which worked badly.

What the researchers drew from these experiments is positive. Participants reported on the socio-affective states they felt during the experiment in self-report questionnaire. Participants feelings were globally positive: participants assessed to have more positive affective states than negative ones during the interaction. Figure 9 shows the evolution of enjoyment interacting with the robot according to the number of interactions (third experimental session). The slight decrease between the first and the second time may be explained by a feeling of lassitude and habituation: users do not have the same surprise than in their first interaction. The size of the sample does not allow statistically significant result but it gives an interesting overview on the interest of the adaptive behavior of the robot.
As for the preference between each scenario, the elderly preferred the emotion game and the quizz to a great extent. Those correspond to scenarios in which the robot clearly asks the user to say something. The users talked much more during the emo-

P: ”It’s been a long time since you drank for the last time.
Take some water please.”
U: ”Not at all.”
P: ”You should drink a bit.”
U: ”I’m not thirsty, I only drink when I’m thirsty.”
P: ”So I’m going to make jokes until you accept.
Take your glass and I’ll show you how I do a headstand!”
U: ”It does not interest me, do your headstand alone !” *(ANGER)*
P: ”Calm down, I don’t do this to command you.
I only suggest you to take a glass of water, it’s not that much.”

Fig. 7 Example of an interaction which functioned correctly with a user: Pepper adapted its behavior correctly in relation to the user’s reactions (scenario of negotiation, ”P” for Pepper, ”U” for User). Consequently, the user got more involved into the conversation.

P: ”Where does this extract come from?”
(MUSIC Title: ”Ya d’la joie”, Singer: ”Charles Trenet”)
”Do you know who is it ?”
U: ”Mmm I think it is Trenet but I’m not sure...” *(LOW VOICE)*
P: ”Can you repeat please ? I didn’t understand well.”
U: ”Maybe Charles Trenet, I don’t know...” *(LOW VOICE and HESITATION)*
P: ”You look wrong, but you look like a good sport!
It was an extract from ”Ya d’la joie”, sung by Charles Trenet.”

Fig. 8 Example of an interaction which worked badly: Pepper did not detect correctly the good answer from the user (scenario of quiz, ”P” for Pepper, ”U” for User). Consequently, the user doubted about the quality of the conversation and turned unwilling to dialog with the robot.

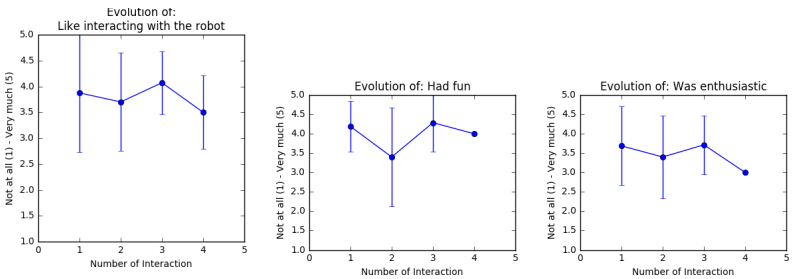

Fig. 9 Third Experimental Session – Answers to Enjoyment Questions (”Did you like interacting with the robot?”, ”Did you have fun?”, ”Did you feel enthusiastic?”) according to the Number of Interaction. The slight decrease between the first and second time interacting with the robot may be explained by user habituation.

tion game. Moreover, the quizz made them react unequally. Some of them talked a lot during the quizz, others did not dare to (maybe scared of being wrong).
Moreover, this longitudinal study will help to provide more data to find new benchmarks specific to interactions between robot and elderly people. These will be taken into account to evaluate the system. Nevertheless, a robot which perfectly fits everyone is an hard to reach ideal. Indeed, figure 10 shows the difficulty to have a robot whose all the elderly finds normal in its behaviors. That shows the diversity between

user evaluation of robot behavior during the interaction. That result may also be due to the adaptive behavior and different strategies it used according to the user reactions detected. Thus, there is a double variability: that of the user personalities, that of the robot choices on behavior strategies.

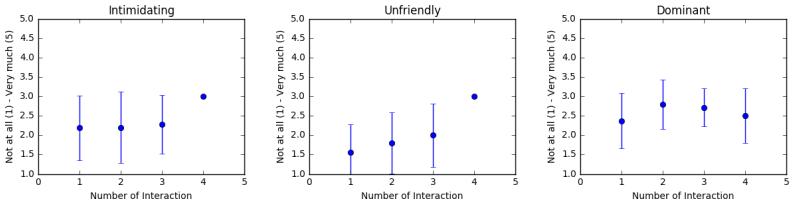

Fig. 10 Third Experimental Session – Answers about Robot Behavior (The robot was comforting/intimidating, friendly/unfriendly, humble/dominant) according to the Number of Interaction. This shows the participant view variability about the concept of "normal" robot.

7 Conclusion and Discussion

To build functional and socially acceptable conversational agents, having benchmarks is primordial. What the authors are trying to do is to find metrics and benchmarks by collecting data from the end-users, in order to evaluate objectively their system, and to go on improving it. In the case of ROMEO2 project, these metrics may be proper to the elderly. The authors are wondering how to make it ethically adapted, efficient and understandable, useful and easy-to-use (from the elderly people perspective) to make them more easily use the technology.
The authors started collecting data first with a WoZ system with Nao, next with an autonomous system and the Pepper robot. The last experimental session allowed them to study the evolution of user reactions according to the individual number of interaction with the robot. The experiments at Broca hospital emphasize main issues: the user concerns about data safety and the adaptation of the robot to the user. In this second point, it is necessary to take into account the age of the user. The robot has to adapt its vocabulary and its behavior according to the user's age. It also has to change its speech velocity and its way to detect emotions (according to the speed of the user speaking).
The system will be improved using data collected at the experiments described and more tests will be done at the Broca Hospital in collaboration with healthcare workers about the acceptability of such a technology.

Acknowledgements The authors wish to thank again Pr. Anne-Sophie Rigaud, head of the gerontology department at the Broca hospital, and her team, for providing access to LUSAGE Living

Lab facilities, hence allowing the authors to carry out this study. The authors would also like to show their gratitude to all the participants of the experiment.

References

1. R. Agrigoroaie, F. Ferland, and A. Tapus. The enrichme project: Lessons learnt from a first interaction with the elderly. In *ICSR*, 2016.
2. A. Aly, S.S. Griffiths, and F. Stramandinoli. Metrics and benchmarks in human-robot interaction: Recent advances in cognitive robotics. *Cognitive Systems Research*, July 2016.
3. A. Aly and A. Tapus. A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. In *HRI*, 2013.
4. L. Devillers, M. Tahon, M. Sehili, and A. Delaborde. Détection des états affectifs lors d'interactions parlées: robustesse des indices non verbaux. *TAL*, 55-2, July 2014.
5. 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*, pages 1–13, 2015.
6. Guillaume Dubuisson Duplessis and Laurence Devillers. Towards the consideration od dialogue activities in engagement measures for human-robot social interaction. In *International Conference on Intelligent Robots and Systems, Designing and Evaluating Social Robots for Public Settings Workshop,*, pages 19–24, 2015.
7. S. Avallone et al. he enrichme robotics and aal project: new tools for the elderly independent living and for behavioral and physiological monitoring. In *TeleMediCare*, 2016.
8. S. Kumar, P. Rajasekar, T. Mandharasalam, and S. Vignesh. Handicapped assisting robot. In *International Conference on Current Trends in Engeneering and Technology (ICCTET)*, 2013.
9. A. Kumar Pandey, R. Gelin, R. Alami, R. Vitry, A. Buendia, R. Meertens, M. Chetouani, L. Devillers, M. Tahon, D. Filliat, Y. Grenier, M. Maazaoui, A. Kheddar, F. Lerasle, and L. Fitte Duval. Ethical considerations and feedback from social human-robot interaction with elderly people. In *AIC*, 2014.
10. M. Panou, E. Bekiaris, K. Touliou, and M.F. Cabrera. Use cases for optimising services promoting autonomous mobility of elderly with cognitive impairments. In *TRANSED*, 2015.
11. M. Panou, M.F. Cabrera, E. Bekiaris, and K. Touliou. Ict services for prolonging independent living ofelderly with cognitive impairments. In *AAATE*, 2015.
12. M. Pino, S. Benveniste, R. Picard, and A.S. Rigaud. User driven innovation for dementia care in france: the lusage living lab case study. *Interdisciplinary Studies Journal*, 3:1–18, February 2017.
13. B. Reeves and C. Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* Cambridge University Press., 1996.
14. S. Tobis, M. Cylkowska-Nowak, K. Wieczorowska-Tobis, M. Pawlaczyk, and A. Suwalska. Occupational therapy students'perceptions of the role of robots in the care for older people living in the community. *Occupational Therapy International*, February 2017.
15. V. J.M. Van der Zwaan Van der Zwaan, J.M. Dignum, V. Dignum, and C.M. Jonker. A bdi dialogue agent for social support: Specification and evaluation method. In *International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS)*, 2012.
16. X. Wei, X. Zhang, and Y. Wang. Research on a detection and recognition method of tactile-slip sensation used to control the elderly-assistant amp; walking-assistant robot. In IEE, editor, *INterational Conference on Automation Science and Engineering (CASE)*, 2012.
17. G. Xiong, J. Gong, T. Zhuang, T. Zhao, D. Liu, and X. Chen. Development of assistant robot with standing-up devices for paraplegic patients and elderly people. In IEE, editor, *International Conference on Complex Medical Engineering (CME)*, 2007.