An Evaluation Framework to Assess and Correct the Multimodal Behavior of a Humanoid Robot in Human-Robot Interaction

Duc Canh Nguyen, Gérard Bailly, Frédéric Elisei

An Evaluation Framework to Assess and Correct the Multimodal Behavior of a Humanoid Robot in Human-Robot Interaction

Duc-Canh Nguyen, Gérard Bailly & Frédéric Elisei

GIPSA-Lab, Grenoble-Alpes Univ. & CNRS, Grenoble, France

firstname.lastname@gipsa-lab.fr

Abstract

We discuss here the key features of a new methodology that enables professional caregivers to teach a socially assistive robot (SAR) how to perform the assistive tasks while giving verbal and coverbal instructions, demonstrations and feedbacks. We describe here how socio-communicative gesture controllers – which actually control the speech, the facial displays and hand gestures of our iCub robot – are driven by multimodal events captured on a professional human demonstrator performing a neuropsychological interview. The paper focuses on the results of two crowd-sourced experiments where we asked raters to evaluate the multimodal interactive behaviors of our SAR. We demonstrate that this framework allows decreasing the behavioral errors of our robot. We also show that human expectations of functional capabilities increase with the quality of its performative behaviors.

1. Introduction

Socially assistive robots

SAR are typically facing two situations with quite different timescales and related challenges: long-term vs. short-term interactions. Long-term interactions often target one single user with the challenge of engaging into open-domain conversations, establishing affective relation, such as performed by) (see Robinson et al., 2014 for a review). In contrast, short-term interactions are typically task-oriented (e.g. welcoming a client, giving directions, serving cocktails (Foster et al., 2014), conducting interviews (Bethel et al., 2016), repetitive and should cope with a large variety of user profiles.

Our work focuses on the development of socio-communicative abilities for short-term interactions. The target scenario is a neuropsychological interview with an elderly person.

2. The SOMBRERO Framework

The multimodal interactive behavioral model learning is performed by three main steps illustrated in Figure 23. Firstly, we collect representative interactive behaviors from human tutors especially by professional coaches. Secondly, the comprehensive models are trained from the collected data with considering a priori knowledge of users' models and task decomposition. Finally, the gesture controllers are built in order to execute the desired behaviors driven by the interactive model.

The interactive models of HRI systems are mostly inspired by Human-Human interaction (HHI). Therefore, they face several issues: (1) adapting the human model to the robot's interactive capabilities; (2) the drastic changes of human partner behaviors in front of robots or virtual agents; (3) the modeling of joint interactive behaviors; (4) the validation of the robotic behaviors by human partners until they are perceived as adequate and meaningful. The two first issues are solved by the framework used in SOMBRERO (Gomez et al., 2015) which allows coaches to involve and demonstrate an expected HRI behavior through immersive teleoperation technique: the so-called beaming method driving gaze, head movements and mouth of the robot in real-time during the interaction. The third issue has been addressed by (Mihoub et al., 2015; Mihoub et al., 2016). They

proposed to train statistical behavioral models that encapsulate discrete multimodal events performed by the interlocutors into a single dynamical system that could be further used to monitor behaviors of one interlocutor and generate behaviors of the other.
In this paper, we propose a method to address the fourth issue: the replay of interactive behaviors by the robot and its assessment by human raters.

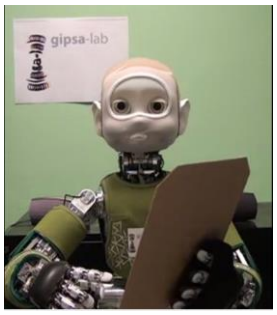

Figure 1. The iCub humanoid (named Nina) robot from the subject's perspective.

Figure 2. Capturing the multimodal behavior of the human tutor during HHI. Movements of the upper limbs (head, arms and hands) are captured by tracking 22 markers glued on these segments with a Qualysis ® mocap system. Gaze was tracked using Pertech ® head-mounted eye tracker.

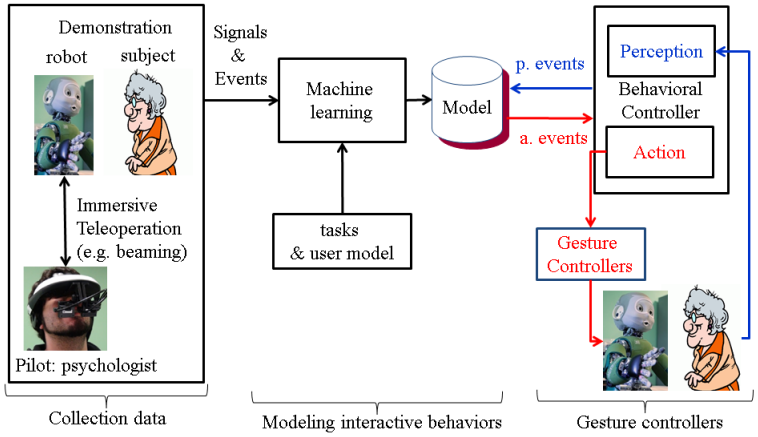

Figure 3. The three main steps of learning interaction by demonstration: collecting HRI data, learning a behavioral model and building appropriate sensorimotor controllers.

From HHI to HRI

The short-term interactive scenario involved here is a French adaptation of the Selective Reminding Test, so-called RL/RI 16 (Dion et al., 2015). It is often used to diagnose early loss of episodic memory. The test includes four phases: (1) words memorization (aka learning), (2) testing the words recall capability, (3) recognition of the words and (4) distractive task, which were described detail in (Nguyen et al., 2016).
In order to avoid complex gestures usually performed by human interviewers using scoring sheets and paper-based notes, the SAR uses two tablets as physical medium: one facing the robot to fake the note taking activity and the other facing the subject to display word items.
HHI demonstrations were performed by a female professional psychologist. We collected her multimodal behavior (speech, head movement, arm gestures and gaze, see Figure 2) when interviewing five different elderly patients together with the speech of the interviewees. These continuous signals were then semi-automatically converted into time-stamped events using Elan (Wittenburg et al., 2006) and Praat (Boersma and Weenink 1996) editors. With Elan, we basically determined hand strokes triggered by the interviewer to grasp and act on resources (workbook, notebook, chronometer) and regions of interest for fixations. With Praat, we hand-checked the phonetic alignment performed by an automatic speech recognition system and added prosodic annotations as well as special phonetic events related to backchannels and breath noises. The HHI multimodal score consists thus in time-stamped speech, head/arm/hands gestures and gaze events.

We then developed modality-specific gesture controllers to map these events to robotic behaviors. HHI to HRI retargeting is thus performed using multimodal events as pivots. This HHI multimodal score is available for download (see section 6).

Speech and Gesture controllers

We built four gesture controllers: arm, gaze, eyelids and speech, which will cooperate together to enable SAR to replicate the RL/RI scenario. The arm gesture controller is based on the iCub Cartesian Interface (Pattacini et al., 2010) and handles three basic gestures: resting, preparing to click and clicking to trigger display/hide items. The gaze gesture controller triggers fixations towards three regions of interest: subject's face, subject's tablet and robot's tablet. The gaze gesture controller synchronizes with the gesture controller for ensuring sensory-motor control, e.g. locking the gaze at finger target when initiating arm gesture. Conversely, when no such sensory-motor control is required, the gaze is driven by the other socio-communicative events. The eyelid gesture controller was added to cope with gaze direction, speech and blinking. Despite blinking rate is known to correlate with emotional state and cognitive state – notably thinking, speaking vs. listening (see Bailly et al., 2010) – blinks are generated according to a Gaussian distribution at 0.5 Hz +/- 0.1 Hz. Finally, the speech gesture controller was handled by our in-house audiovisual text-to-speech system (Bailly et al., 2009). The corpus-based of AV synthesis is fed with articulatory movements from a female adult that have been scaled to NINA's degrees of freedom and time-aligned with voice segments from a female teenager.

3. Evaluation

We want to evaluate if the coordinated behaviors are perceived and interpreted as expected by subjects. Since subjects can not both live and rate the interaction on-line, we thus asked third parties (observers or raters) to rate the final rendering of a multimodal score recorded during HHI – and replayed by our robotic embodiment.

State of the art

Most subjective evaluations of HRI behavior have been performed using questionnaires, where subjects or third parties are asked to score specific dimensions of the interaction on a Likert scale. (Fasola and Mataric, 2013) rated several aspects such as pleasure, interest, satisfaction, entertainment and excitation. (Huang and Mutlu, 2014) assessed a narration of a humanoid robot along several dimensions such as immediacy, naturalness, effectiveness, likability and credibility. (Zheng et al., 2015) compared control strategies for robot arm gestures along dimensions such as intelligibility, likeability, anthropomorphism and safety. Although delivering very useful information notably for sorting between competing control policies or settings, these questionnaire-based evaluations provide developers with poor information about how to correct faulty behaviors since the evaluation is performed offline and questions address global properties of the entire interaction.

Designing and performing on-line vs off-line evaluation

Following the procedure proposed by (Kok and Heylen, 2011), we opted for a method that enable raters to signal faulty events, since the HRI behaviors are essentially controlled by events. We thus designed an on-line evaluation technique that consists in asking raters to immediately signal faulty behaviors by pressing on the ENTER key of their computer when they just experience them. Following Kok & Heylen, we will use the term yuck responses to name these calls for rejection.

In order to gather a significant amount of yuck responses for a set of identical stimuli, we here ask our raters to evaluate the replay by the robot of the multimodal behavior, originally performed by our psychologist in front of one unique subject. We in fact filmed the robot's performance as fed by the multimodal score of the original situated interaction (arm gestures, head movement, gaze...). For now, the only original play-backed behavior is the subject speech. The camera remains fixed at the mean position of the location of the eyes of the subject. The raters can see the robot facing them, but not the patient that they replace. They can hear the robot, as well as what the subject says: they are spectators, but occupy the seat of the subject. For our first experimental assessment (Nguyen et al., 2016), we created a website (see section 6) where we ask people to look at a first-

person video and to press the ENTER key anytime they feel the robot behavior is incorrect. This provides a time-varying normalized histogram of incorrect behaviors. The maxima of the density function are cueing time-intervals for which a majority of raters estimate the behavior is inappropriate or hinders the interaction. Further diagnostic of what cues cause these faulty behaviors are later performed by roboticists and system designers. This on-line evaluation task is preceded by a quick screening of subjects (age, sex and mother tongue) and a familiarization exercise, and followed by a questionnaire that asks the subjects judgments (five-level Likert) on nine points: "Did the robot adapt to the subject?"; "Did the subject adapt to the robot?"; "Did you feel relaxed?"; "Did you feel secure?"; "Was the rhythm of the robots behavior well adapted?"; "Was the interaction pleasant?"; " Was the multimodal behavior appropriate?"; "Did the robot pay attention while speaking?"; "Did the robot pay attention while listening?"

The present paper builds up new results upon a previous experiment (Nguyen et al., 2016) performed by 50 French native subjects (26 males, 24 females, 32±12 years). The 25 most signaled events were related to the following problems: (1) Ungrounded pointing gestures; Underrepresented gaze contacts; (2) Inactivity during reverse counting or covert thinking; (3) Inadequate speech articulation; (4) Lack of facial expressions.

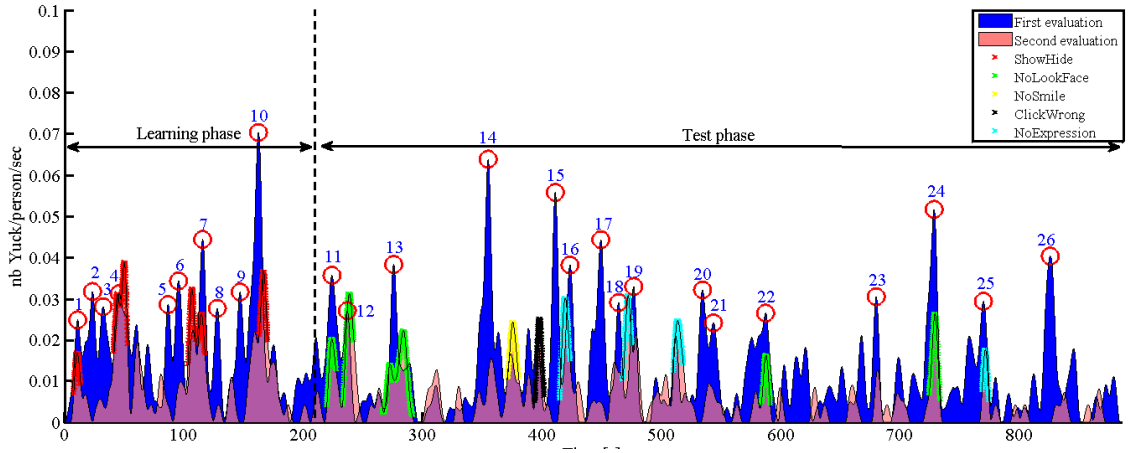

Figure 4. Comparing the yucking probability as a function of time for first vs. second assessment by the subjects (blue area: first evaluation, pink area: second evaluation, purple area: overlap between the first and second evaluations, dot-lines: annotated yuck).

For the current study, after correcting these faulty behaviors, we performed a new experimental assessment using the same experimental protocol. The second experimental assessment was performed by 46 French native subjects (16 males, 30 females, 36±16 years), 38 of whom already participated in the first assessment.

Results

Yuck responses

We remedied to these faulty behaviors by adding extra-rules to our gesture controllers. For example, in order to avoid immobility due the periods of poor external stimulation, the gaze controller automatically randomly loops on the two last regions of interest when the delay from the last fixation exceeds 3 sec.

With this rule, the number of yucks at timestamps 10, 11, 12, 13, 16, 18, 19 and 23 are significantly reduced as shown in Figure 4. However, this randomization should not be equally distributed and should favor the subject's face, since the participants still complain about its lack of engagement with the human subject (e.g. around peak 11, 12, 13 in counting task). This problem will be suppressed by systematically adding the subject's face to the current attention stack and favoring this region of interest in the gaze distribution.

In the first evaluation, yucks occuring at timestamps 2, 3, 4 were due to the wait-motion-done setting. In the redesign, theses faulty behaviors have been removed by disabling the wait-motion-done option that discards any new command while the current gesture has not reached its target

given a given precision. This policy is efficient: the yuck responses at landmarks 2, 3, 4 are significantly reduced in the second evaluation. The yucks at landmarks 14 and 15 were repaired by forcing the closing gesture at the end of phonation. Although many of the faulty behaviors are suppressed, several faulty detections still remain while some new yucks emerge from the background, notably the absence of expressiveness, e.g. smile responses to subject's embarrassment or head nodding normally associated with incentives, respectively cued by yellow vs. cyan extrema.

Subjective ratings and comments

We also compared subjective ratings from the first vs. second assessment (see Figure 5). While the new behavioural score results in an effective decrease of the yuck responses and descent behavior effectively improves – most other off-line subjective ratings degrade. Likelihood ratio tests comparing the combined multinomial model RATINGS ~ SEX+SESSION+EXPOSURE with the individual models RATINGS ~ SEX+SESSION, RATINGS ~ SEX+EXPOSURE and RATINGS ~ EXPOSURE+SESSION show that sex significantly contributes to the ratings of questions robot adaptation. In addition, the version ($p = 0.049$) and the number of evaluations ($p = 0.041$) has significant contributions on feels_relaxed. This means that people feel more relaxed and the robot was rated as more friendly in the second evaluation.

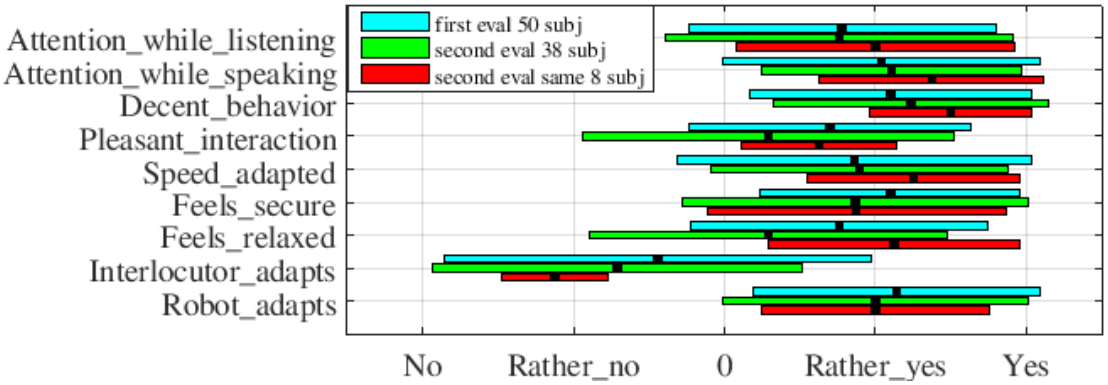

Figure 5. Comparing subjective ratings according to conditions

In the free comments, some raters of the first evaluation campaign mentioned the rather directive style of our female interviewer and the absence of emotional vocal and facial displays on our SAR e.g. laughs and smiles. While most raters of the second evaluation campaign underly the quality of gaze behavior, the majority criticize the poorness of emotional displays: "robot without human warmth!", "why robots never smile?", etc. It seems that the increased behavioral quality and appropriateness also increased the participant's expectations. When they have the impression that the robot is reactive, aware of the situation and monitors the interaction task in an appropriate way, they can allocate more attentional resources to the social and emotional aspects of the interactive behavior.

4. Conclusions

We have put forward an original framework for the on-line evaluation of HRI behavior that offers subsequent glass-box assessment: On-line evaluation provides developers with when something goes wrong. Post-hoc reverse engineering should be then performed by the socially assistive robot (SAR) designers to remedy for the potential causes of the most salient yuck responses, i.e. what went wrong. Off-line assessment provides developers with what is missing. These local vs global assessment procedures should be combined to maintain SAR at the top of the uncanny cliff.

We should augment the socio-communicative skills of our SAR with more expressive dimensions. While Nina is missing facial displays (notably articulated eyebrows), its available degrees-of-freedom (notably head, arm and body gestures) together with speech should be recruited to encode more linguistic and paralinguistic functions.

5. Acknowledgements

This research is supported by SOMBRERO *ANR*-14-CE27-0014 , PERSYVAL *ANR*-11-61 and ROBOTEX *ANR*-10-EQPX-44-01.

6. Appendix

The test page is available at http://www.gipsa-lab.fr/~duccanh.nguyen/assessment
Multimodal data/labels are freely available at: http://www.gipsa-lab.fr/projet/SOMBRERO/data

References

Bailly G. & Gouvernayre, C. (2012). *Pauses and respiratory markers of the structure of book reading*. In Interspeech. Florence, Italy, pp. 2218–2221.

Bailly G., Raidt, S. & Elisei, F. (2010). Gaze, conversational agents and face-to-face communication. *Speech Communication - special issue on Speech and Face-to-Face Communication*, 52(3), pp. 598–612.

Bailly G., Govokhina, O., Elisei, F. & Breton, G. (2009). Lip-synching using speaker-specific articulation, shape and appearance models. *Journal of Acoustics, Speech and Music Processing. Special issue on "Animating Virtual Speakers or Singers from Audio: Lip-Synching Facial Animation"*. Retrieved from https://asmp-eurasipjournals.springeropen.com/articles/10.1155/2009/769494

Bethel C., Henkel Z., Stives, K., May, D. C., Eakin D. K., Pilkinton, M., Jones, A., Stubbs-Richardson, M., (2016). *Using robots to interview children about bullying: Lessons learned from an exploratory study*. Preceeding of 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN),712–717.

Boersma P., & Weenink D., (2013*): Praat: doing phonetics by computer, version 5.3.51*. Retrieved from http://www.praat.org/

Dion M1, Potvin O, Belleville S, Ferland G, Renaud M, Bherer L, Joubert S, Vallet GT, Simard M, Rouleau I, Lecomte S, Macoir J, Hudon C. (2015). *Normative Data for the Rappel libre/Rappel indicé à 16 items (16-item Free and Cued Recall) in the Elderly Quebec-French Population*. The Clinical Neuropsychologist, 28(1), pp. 1–19. https://www.ncbi.nlm.nih.gov/pubmed/24815338

Fasola J. & Mataric, M. (2013). A socially assistive robot exercise coach for the elderly. *Journal of Human-Robot Interaction*, 2, pp. 3–32.

Foster M.E., Keizer, S. & Lemon, O. (2014). *Towards action selection under uncertainty for a socially aware robot bartender*. In Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction, pp. 158–159.

Gomez G., Plasson, C., Elisei, F., Noël, F., & Bailly, G. (2015). *Qualitative assesment of a beaming environment for collaborative professional activities*. European conference for Virtual Reality and Augmented Reality (EuroVR). Retrieved from https://hal.archives-ouvertes.fr/hal-01228890

Huang C.-M. & Mutlu, B. (2014). *Learning-based modeling of multimodal behaviors for humanlike robots*. In Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction, pp. 57–64.

Kok I. & Heylen, D. (2011). *Observations on listener responses from multiple perspectives*. Proceedings of the 3rd Nordic Symposium on Multimodal Communication, Northern European Association for Language Technology, pp. 48–55.

Mihoub A., Bailly, G. Wolf, C. & Elisei, F. (2016). Graphical models for social behavior modeling in face-to face interaction. *Pattern Recognition Letters* 74, pp. 82–89.

Mihoub A., Bailly, G. Wolf, C., & Elisei, F. (2015). Learning multimodal behavioral models for face-to-face social interaction. *Journal of Multimodal User Interfaces*, volume 9, issue 3, pp. 195–210

Nguyen D. C., Bailly, G. & Elisei, F. (2016). *Conducting neuropsychological tests with a humanoid robot: Design and evaluation*. In Cognitive Infocommunications (CogInfoCom), Warsaw, Poland, 337–342.

Pattacini U., Nori, F., Natale, L., Metta, G., & Sandini, G. (2010). *An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots*. International Conference on Intelligent Robots and Systems (IROS), pp. 1668–1674.

Robinson H., MacDonald, B. & Broadbent, E. (2014). The role of healthcare robots for older people at home: A review. *International Journal of Social Robotics* 6(4), pp. 575–591.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A.,& Sloetjes, H. (2006). *Elan: A professional framework for multimodality research*. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 1556-1559

Zheng M., J. Wang & Meng, M.Q.-H. (2015). *Comparing two gesture design methods for a humanoid robot: Human motion mapping by an RGB-D sensor and hand-puppeteering*. Proceesding of 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), IEEE, pp. 609–614.