RoboTrio2: Annotated Interactions of a Teleoperated Robot and Human Dyads for Data-Driven Behavioral Models

Frédéric Elisei[1,*], Léa Haefflinger[1,2] and Gérard Bailly[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, F-38000 Grenoble, France
[2]Atos, Échirolles, France

Abstract

We present here RoboTrio2, an annotated multimodal corpus of interactions between an autonomous-looking social robot and two humans, and the original way to record it: an immersive tele-operation of the robot, which makes it behave naturally and efficiently, and captures many signals (gaze including vergence, head and neck movements, the exact subjective stereo views that motivate the decisions in the interaction, binaural audio...). With this high level of embodiment, the pilot provides the robot with demonstrations of conversational skills to conduct a natural interaction with humans and successfully perform the intended task (social interactions in a gaming scenario, with gaze and speech turnovers). The behaviors of its two human partners are also recorded through static HD cameras and headset microphones to ease annotation. Training autonomous behavioral models for our social robot is the main goal of this 8-hour corpus, but the study of elicited human behaviors is also possible with the corpus and annotations we released.

Keywords

human-robot interaction, social robotics, multi-party, cooperative game, head and gaze orientation, immersive teleoperation

1. Introduction

Nowadays machine learning needs adequate training data. What do social robots need to train to social interaction with naive humans? Would it be successful to imitate human signals, just like children do? The generation of verbal and non-verbal behaviors for robots is frequently based on human-human interaction dataset [1, 2, 3]. But robots – even humanoids ones – have different bodies and capabilities, will not grow up in a human body and must be prepared to be alternatively considered by users as agents or objects, or even ignored when their behavior is no more appropriate...

Some studies have already highlighted the differences between Human-Human Interaction (HHI) and Human-Robot Interaction (HRI): Children appear to be more expressive when playing with another child than with a robot [4], the position of a human's head during turn-taking varies depending on whether the change is occurring between two humans or between a human and

HHAI-WS 2024: Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 10—14, 2024, Malmö, Sweden
*Corresponding author.
frederic.elisei@gipsa-lab.grenoble-inp.fr (F. Elisei)
0000-0002-1295-3445 (F. Elisei); 0009-0009-6592-040X (L. Haefflinger); 0000-0000-0002-6053 (G. Bailly)

a robot [5], the duration of gaze fixations of a robot face is longer than for a human face [6], humans modify their prosody when addressing a machine [7]. To tackle this problem and study human-robot interactions, the wizard of Oz method is often used [8], where the robot is remotely controlled by a human using buttons and predefined actions [9]. A second possible method is a robot controlled by rules [10, 11]. However both options restrict the possible actions to those that are predefined, Both introduce a lack of naturalness and fluidity in the interaction, also modifying the transistions. In addition, they often limit the experiment to the study of a unique aspect of the interaction.

If learning by imitation, robots should do **from other robots that have the same sensors/actuators (body)** and already engage successfully in fluid, natural, ecological interactions with humans ... while humans interact with this yet-to-be autonomous robot!

We describe here **an original method for collecting such corpora with fluid humans/robot interactions: immersive teleoperation of a humanoid robot** can collect multimodal signals intrinsically adapted to the specific robot sensors/actuators and its specific reaction times, while bringing the social know-how, language understanding, and decision-taking of a human tutor into the sensory-motor coupling.

The **here delivered 8-hour RoboTrio2 corpus** [12] is such an example, with many annotations. It consists of a task-oriented interaction of a robot with 23 humans pairs, collected in French by a single pilot. This dataset was successfully used to train a machine learning model for robot gaze control [13]. It could also be used to study conversational modes, gaze and head behavior, or to compare with behavior from HH data.

2. Immersive teleoperation

Figure 1 shows our setup for RoboTrio2. The human pilot that immersively drives the robot is in a remote room and wears an HTC Vive VR headset, that embeds two SMI eyetrackers. He also uses stereo earbuds and a microphone.

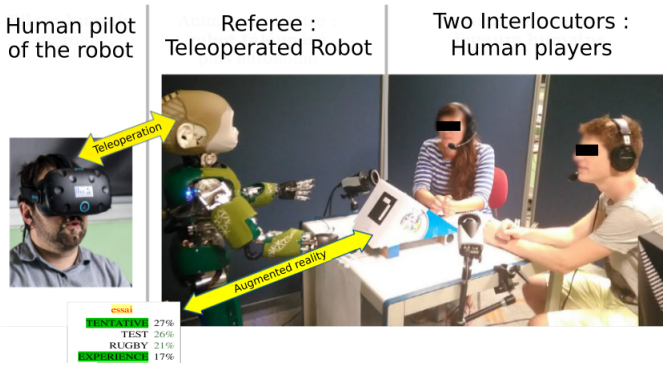

Figure 1: Immersive teleoperation of a robot, to collect natural interaction data of 2 human users in front of an autonomous-looking social robot (and its tablet, using mixed-reality to show in-game help). The table sides also support two HD cameras, directed towards the humans to ease the automatic/manual annotation process.

His chin and lip corners have motion capture markers to drive the robot face in real time with his articulations: The robot is a modified iCub [14] that has an articulated mouth (hiding a speaker), mobile eyes (each embedding a VGA camera) that move like human ones (with 3 degrees of freedom, including vergence), and a microphone in each ear (with human-shaped ear pinnas). The robot neck allows human-like orientations of the head (with 3 degrees of freedom).

Logged streams: With our original immersive teleoperation system [15], all these real time streams are synchronized and recorded, and control the iCub robot in real-time. Pilot's gaze and pilot head movements are 60 Hz streams. We also record what the robot does from these motor instructions to drive the equivalent 6 degrees of freedom (3 for the neck, 3 for the eyes including vergence). The pilot generates his interactive behaviors (where to look, what to say, head movements...) from what he hears and what he sees through the robot sensors, captured by the robot ears and through the pair of cameras embedded in its mobile eyes. No joystick or button press for the pilot, he just directs his own face and gaze, that the VR headset and its dual eye-tracker track. Pilot's jaw and lip movements are tracked by a Qualisys motion capture system and drive the robot face to shape its mouth, that relays the pilot speech through a speaker.

3. The RoboTrio2 corpus

The corpus involves **a cooperative game played by the two humans.** They sit in front of a social robot that acts as a game animator and referee. This robot is teleoperated as previously described, resulting in a high level of embodiment. What is demonstrated by the pilot is **a viable solution with the specific robot sensors/actuators** to conduct a natural real time interaction with humans (decoding and generating meaningful gaze and aversion, speech turnovers...) to successfully perform the intended social interaction needed by the gaming scenario. What is experienced by the human players is **an autonomous-looking robot that utilizes natural language and ecological head and gaze patterns to perform the joint task.**

This 8-hour corpus logs data streams and events linking perception and action, making it **ideal for building autonomous behavior models** for our social robot, Nina, a modified iCub. But with all the provided annotations, it can also be used to study all the humans that interacted with this robot: we recorded 23 interactions with different human pairs (either male or female) while the robot is always teleoperated by the same human pilot (to help build a coherent one-to-many robot behavior model).

The game: It is played by a team of two humans, trying to find the words most commonly associated with a given theme (previously played online by other human players). E.g. for the "eat" theme, the words that would score the most are "drink", "food", "lunch", "diner", "swallow" and "feed". The same 9 theme words are played for all the games, and 5 answers are collected per theme. During the game, our players collaborate to find the best answers and look/question the robot at will. This scenario generates a lot of interaction and social cues; thinking about the theme, brainstorming and debating potential answers, etc. The robot guides them as its human

pilot would, and frequently takes part in the conversation. **The corpus is both complex and rich in verbal and non-verbal content** for the players and the robot (mutual gaze, gaze aversion, speech overlap, backchannel . . .). *Videos, annotations and extra details can be seen online* [16].

To ease the post-recording annotations, the two human players are also recorded by **two fixed HD cameras** (synchronized with the Qualisys capture system). These are not used by the robot nor the pilot, but were helpful in annotating the interaction signals and meaningful events: gaze directed to robot/other player, prephonatory gestures, thinking attitudes... while the first person cameras are low resolution and will exhibit motion blur.

We also ran OpenFace [17] to **extract head rotations and eye movements** as well as FACS Action Units for every player (seen by the HD cameras), giving access to higher-level events (e.g. lip opening for prephonatory gestures).

3.1. What has been annotated

We use Elan from MPI [18] to concentrate all the multi-channels audio and video streams in parallel tracks, plus the annotations of the robot streams/motion capture that form the corpus. Figure 2 shows some of the hierarchical annotations of the verbal content.

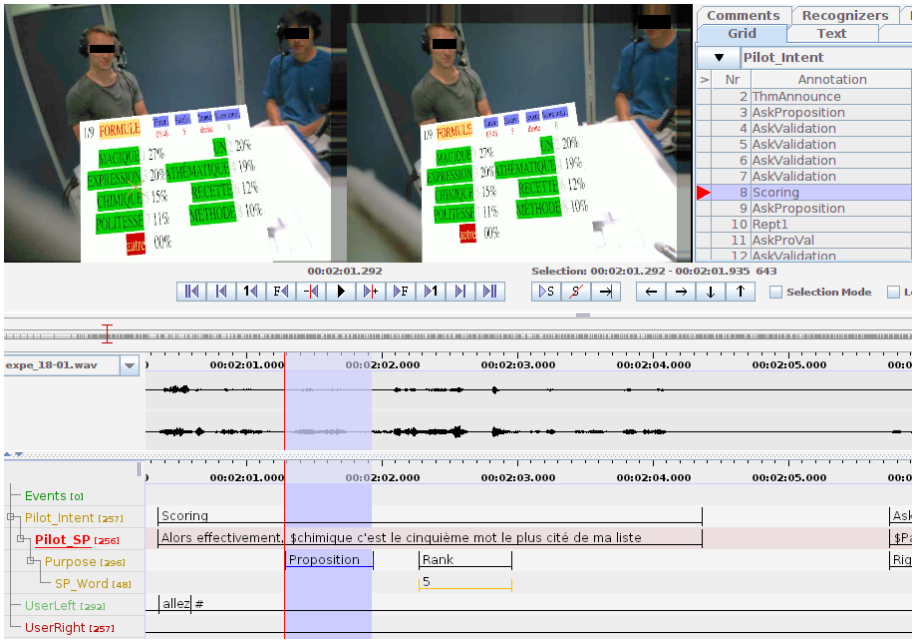

Figure 2: Hierarchical annotation of speech intents in Elan. Highlighted instant is scoring the proposition “chimique" (chemical), ranked fifth. Top image pair corresponds to the first-person view of the pilot through the two eyes/cameras of the robot, and shows the virtual tablet with the theme “formule" (formula) being played currently, and the eight best answers (magic, one – as in formula one –, expression, mathematical, chemical, recipe, polite, method). Grid on top right lists the previous/next intents of the robot/pilot in this dialog.

Speech transcriptions: Speech has been transcribed for the pilot as well as the human players. In our specific scenario, some spot-words play specific roles and have been annotated specifically: **the themes** that the referee gives to the players (known beforehand), and all the words that the players may give as a **proposition**, or discuss together before making a formal proposition.

Speech acts of the robot/pilot: We listed 23 different classes for the pilot speech intents, including: ask for a proposition (or a validation), repeat a proposition or the theme again, give the score/the theme/an explanation or feedback, wait for players after each round.

Gazes of the robot/pilot: The pilot gaze focal point is computed from the 60 Hz recordings of the pilot's head and eyes movements. After detecting ocular saccades, these points were classified using Gaussian Mixture Models (GMM) into 4 different targets: LeftUser (leftmost player), RightUser (rightmost player), Tablet (live game info), and Elsewhere.

Gaze of the users: By combining the players' head and eye positions provided by OpenFace, their gaze was classified using GMM (after detection of ocular saccades) according to their three targets: Robot, other User, Elsewhere.

4. Statistics on the corpus

Of the 23 recorded sequences, 11 (nearly 4 hours) are fully annotated both verbally and non-verbally. To illustrate the richness and interest of this corpus, this section presents some statistics on the behavior of the pilot and the users, as well as some findings.

Verbal statistics: As the roles in this corpus are asymmetrical, verbal behavior of the pilot/robot and players differ. As seen in Table 1, the **number of utterances** is equivalent between the participants, but the average duration of these and therefore the total speaking time differ significantly. Indeed, users produce a lot of **backchanneling** (few hundred) or positive/negative **feedbacks** (almost 2,000) to share their reactions to the proposals or scores given, resulting in very short utterances, unlike the pilot, who animates the game and may have to use longer sentences when announcing the theme or scores.

Table 1
Verbal statistics of the participants

Stats	Pilot	LeftUser	RightUser
#Utterrances	2812	2714	2663
Mean duration	1.84s	0.91s	0.99s
Speaking Time	86min	41min	44min

Concerning **the pilot's intentions,** 10 of the 23 occurred for at least 90 utterances. The 3 most frequent are *"ask for the validation of a proposal", "give score"* and *"ask for a proposal"*, with 574, 367 and 363 utterances respectively.

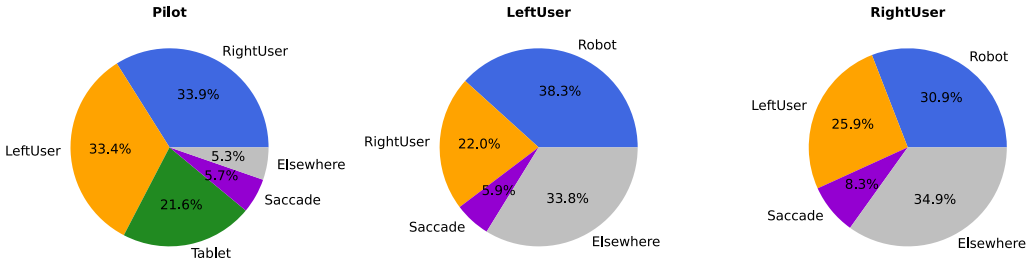

Figure 3: Distributions of the pilot & players gazes

Furthermore, as this corpus is multi-party, the pilot can address one player or both at the same time. This allows the study of the different participant roles in a conversation (speaker, listener, side participant,...), and their regulation, called "Footing" in [19]. To obtain an indication of who **the pilot's addressee(s)** are, we detected the use of French pronouns in his sentences: *"Tu"* (You) for one addressee vs. *"Vous"* (You) for both. As a result, 172 utterances contain the *"Tu"* pronoun and 632 the *"Vous"* pronoun. Players' first names were also used in 139 utterances. This corpus has already been used successfully to compare the different behaviors of the teleoperated robot's head depending on whether he was addressing one or both players [20].

Gaze statistics: As gaze is one of the most important non-verbal cues in human and HRI conversations [21], this section presents some general statistics on it. First, figure 3 shows the gaze distributions of the pilot/robot and the users for the 11 sequences. Both users are looked at equally by the pilot (there could be some variations inside sequences but not globally). We can also see the significant use of the tablet, which is looked at by the robot/pilot to retrieve needed information. The players have looked at the robot a lot, indicating that they haven't put the robot aside and have included it in the conversation. The *Elsewhere* class is also well represented, due to the many moments when players need to think.

As the pilot's behavior can be affected by his role in the conversation, speaker or listener, table 2 shows in more detail the distribution of his gaze according to his activity. When he's speaking, he looks more at the tablet that displays information about the game in progress. When one of the players is speaking, the pilot will often look at him/her (green cells), however **the proportion of his gaze directed at the other player is not low** (yellow cells). The role of

Table 2
Distribution of the pilot's gaze depending on his activity as speaker or as listener.

Pilot's target	Who speaks? Self	LeftUser	RightUser
LeftUser	23.6%	50.8%	31.6%
RightUser	24.0%	31.4%	51.6%
Tablet	39.0%	8.6%	8.8%
Elsewhere	6.3%	4.0%	3.9%
Saccade	7.2%	5.0%	5.1%

game facilitator requires the pilot **to observe the reactions of the players and motivate their collaboration,** making his behavior a complex one (best generated with machine learning and ad hoc data).

5. Conclusion

We have shown here how the immersive teleoperation of a robot can produce a valuable corpus, especially for training social robotics models. In contrast to human-human corpora [22, 23], **our data may include the change of expectations and behavior in front of robotic bodies.** In addition, this setup provides data from a fluid interaction in HRI, where the robot's behaviors are less constrained than when using usual wizard of Oz or rule-based methods [24, 25, 26]. The recorded behaviours are also suitable for studying conversational modes, gaze and head behavior in natural HRI. As an example, we provide RoboTrio2, an 8-hour annotated multimodal corpus of a multi-party game (available online [12]), and outline here both some of its contents and first results from its analysis.

Acknowledgments

This data collection was funded by a CNRS S2IH PEPS project, involving GIPSA-lab, LPL and INT. We are grateful to our subjects and to the people who contributed to the immersive teleoperation platform (M. Sauze, R. Cambuzat, and C. Plasson) and to the RoboTrio2 recording/annotation (N. Loudjani, O. Granier, and J. Rengot). Part of this work is funded by ANR 19-P3IA-0003 MIAI and a PhD granted by ANRT (2021/0836).

References

[1] T. Shintani, C. T. Ishi, H. Ishiguro, Analysis of role-based gaze behaviors and gaze aversions, and implementation of robot's gaze control for multi-party dialogue, in: Proceedings of the 9th International Conference on Human-Agent Interaction, HAI '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 332–336.

[2] C. Oertel, P. Jonell, D. Kontogiorgos, K. F. Mora, J.-M. Odobez, J. Gustafson, Towards an engagement-aware attentive artificial listener for multi-party interactions, Frontiers in Robotics and AI 8 (2021) 555913.

[3] J. Domingo, J. Gómez-García-Bermejo, E. Zalama, Optimization and improvement of a robotics gaze control system using lstm networks, Multimedia Tools and Applications 81 (2022) 1–18.

[4] S. Shahid, E. Krahmer, M. Swerts, Child–robot interaction across cultures: How does playing a game with a social robot compare to playing a game alone or with a friend?, Computers in Human Behavior 40 (2014) 86–100.

[5] M. Johansson, G. Skantze, J. Gustafson, Head pose patterns in multiparty human-robot team-building interactions, in: Social Robotics: 5th International Conference, ICSR 2013, Bristol, UK, October 27-29, 2013, Proceedings 5, Springer, 2013, pp. 351–360.

[6] C. Yu, P. Schermerhorn, M. Scheutz, Adaptive eye gaze patterns in interactions with human and artificial agents, ACM Transactions on Interactive Intelligent Systems (TiiS) 1 (2012) 1–25.
[7] E. Shriberg, A. Stolcke, D. Hakkani-Tür, L. P. Heck, Learning when to listen: Detecting system-addressed speech in human-human-computer dialog., in: INTERSPEECH, 2012, pp. 334–337.
[8] L. D. Riek, Wizard of oz studies in hri: A systematic review and new reporting guidelines, J. Hum.-Robot Interact. 1 (2012) 119–136.
[9] N. Dahlbäck, A. Jönsson, L. Ahrenberg, Wizard of oz studies: Why and how, in: Proceedings of the 1st International Conference on Intelligent User Interfaces, IUI '93, Association for Computing Machinery, New York, NY, USA, 1993, p. 193–200.
[10] G. Skantze, M. Johansson, J. Beskow, Exploring turn-taking cues in multi-party human-robot discussions about objects, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 67–74.
[11] M. Moujahid, H. Hastie, O. Lemon, Multi-party interaction with a robot receptionist, in: Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction, HRI '22, IEEE Press, 2022, p. 927–931.
[12] F. Elisei, L. Haefflinger, L. Prévot, G. Bailly, The robotrio2 corpus, 2023. URL: https://hdl.handle.net/11403/robotrio/v2, ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr.
[13] L. Haefflinger, F. Elisei, B. Bouchot, B. Varini, G. Bailly, Data-driven generation of eyes and head movements of a social robot in multiparty conversation, in: International Conference on Social Robotics, Springer, 2023, pp. 191–203.
[14] A. Parmiggiani, M. Randazzo, M. Maggiali, F. Elisei, G. Bailly, G. Metta, An articulated talking face for the icub, in: 2014 IEEE-RAS International Conference on Humanoid Robots, 2014, pp. 1–6. doi:10.1109/HUMANOIDS.2014.7041309.
[15] R. Cambuzat, F. Elisei, G. Bailly, O. Simonin, A. Spalanzani, Immersive Teleoperation of the Eye Gaze of Social Robots Assessing Gaze-Contingent Control of Vergence, Yaw and Pitch of Robotic Eyes, in: ISR 2018 - 50th International Symposium on Robotics, VDE, Munich, Germany, 2018, pp. 232–239.
[16] F. Elisei, Presentation of the robotrio corpus, https://www.gipsa-lab.grenoble-inp.fr/~frederic.elisei/RoboTrio, 2024. [Online; accessed 1-April-2024].
[17] B. Amos, B. Ludwiczuk, M. Satyanarayanan, OpenFace: A general-purpose face recognition library with mobile applications, Technical Report, CMU-CS-16-118, CMU School of Computer Science, 2016.
[18] H. Brugman, A. Russel, X. Nijmegen, Annotating multi-media/multi-modal resources with elan., in: LREC, 2004, pp. 2065–2068.
[19] E. Goffman, Footing, Semiotica 25 (1979) 1–30.
[20] L. Haefflinger, F. Elisei, S. Gerber, B. Bouchot, J.-P. Vigne, G. Bailly, On the benefit of independent control of head and eye movements of a social robot for multiparty human-robot interaction, in: M. Kurosu, A. Hashizume (Eds.), Human-Computer Interaction, Springer Nature Switzerland, Cham, 2023, pp. 450–466.
[21] H. Admoni, B. Scassellati, Social eye gaze in human-robot interaction: A review, J.

Hum.-Robot Interact. 6 (2017) 25–63.

[22] A. D. Marshall, P. L. Rosin, J. Vandeventer, A. Aubrey, 4d cardiff conversation database (4d ccdb): A 4d database of natural, dyadic conversations, Auditory-Visual Speech Processing,{AVSP} 2015 (2015) 157–162.

[23] O. Celiktutan, E. Skordos, H. Gunes, Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement, IEEE Transactions on Affective Computing 10 (2017) 484–497.

[24] D. B. Jayagopi, S. Sheiki, D. Klotz, J. Wienke, J.-M. Odobez, S. Wrede, V. Khalidov, L. Nyugen, B. Wrede, D. Gatica-Perez, The vernissage corpus: A conversational human-robot-interaction dataset, in: 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2013, pp. 149–150.

[25] A. Ben-Youssef, C. Clavel, S. Essid, M. Bilac, M. Chamoux, A. Lim, Ue-hri: A new dataset for the study of user engagement in spontaneous human-robot interactions, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 464–472.

[26] E. Kesim, T. Numanoglu, O. Bayramoglu, B. B. Turker, N. Hussain, M. Sezgin, Y. Yemez, E. Erzin, The ehri database: a multimodal database of engagement in human–robot interactions, Language Resources and Evaluation (2023) 1–25.