On the Benefit of Independent Control of Head and Eye Movements of a Social Robot for Multiparty Human-Robot Interaction

Léa Haefflinger, Frédéric Elisei, Silvain Gerber, Béatrice Bouchot, Jean-Philippe Vigne, Gérard Bailly

On the Benefit of Independent Control of Head and Eye movements of a Social Robot for Multiparty Human-Robot Interaction

Léa Haefflinger[1,2], Frédéric Elisei[1][0000−0002−1295−3445], Silvain Gerber[1], Béatrice Bouchot[2], Jean-Philippe Vigne[2], and Gérard Bailly[1][0000−0002−6053−0818]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
{lea.haefflinger,frederic.elisei,silvain.gerber,gerard.bailly}@gipsa-lab.fr
[2] Innolab, ATOS, 38130 Echirolles, France
{beatrice.bouchot,jean-philippe.vigne}@atos.net

Abstract. The human gaze direction is the sum of the head and eye movements. The coordination of these two segments has been studied and models of the contribution of head movement to the gaze of virtual agents or robots have been proposed. However, these coordination models are mostly not trained nor evaluated in an interaction context, and may underestimate the social functions of gaze. Indeed, after analyzing human behavior in a three-party conversation dataset, we show that the contribution of the head to the gaze varies depending on whether the speaker is addressing two interlocutors or one of them: the conversational regime actually impacts the head/eyes coordination. We therefore propose an evaluation of different coordination policies in a social interaction context, using a Furhat robot to replay the human multimodal behavior from our data record. The verbal content and gaze targets are the same, but the robot uses four different head and eye coordination policies. (1) Furhat's default gaze control, whose eyes move faster and start before the head, but finally aligns both segments. (2) the robot head is fixed and only the eyes move. (3) the eyes are fixed and only the head moves. (4) Human-like control where the robot mimics the head movements of the human dataset, which naturally exploits independent eye and head control. Using an online crowdsourced test, we show that the human-like policy, which uses decoupled head and eye movements, is perceived significantly more natural than the others.

Keywords: Human-Robot Interaction · Head Orientation · Gaze · Multiparty Interaction · Multimodal Attention.

Introduction

Non-verbal cues are an essential part of human conversation, and in particular gaze cues. The gaze has many functions in human face-to-face interaction, such as giving feedback, complementing speech with emotional information, as well as

regulating the conversation and turn-taking in particular [1, 2]. In a multi-party conversation, the gaze is even more important to regulate the flow: it strongly contributes to addressee identification [3] or next speaker identification [4]. Conversely, social robots and virtual agents must also be able to generate gaze cues to interact smoothly with humans.

Gaze generation can be decomposed into two parts, the identification of where the robot should focus its attention, and the control policy that determines how body segments (from feet, trunk, and head to eyes [5]) direct and signal that attention. The gaze is essentially a combination of eye and head movements, and these two vectors have their own kinematics [6]. A realistic way to manage the attention of robots would be to control independently the eyes and the head of the robot. However, in Human-Robot Interaction (HRI) gaze is not always supported by these two vectors since some robots do not have the ability to reproduce human movements and cannot move their eyes freely – Nao robot [7] for example – or their neck – many telepresence robots for example [8]. In this case, gaze models only use the head or the eyes as attention vectors. In contrast, some robots are able to produce realistic eyes movements like the iCub [9] or Romeo [10] robots. More realistic kinematic models can be implemented on these robots.

Several bio-inspired control models that decouple eyes and head movements have been proposed, such as the model proposed by Itti [11] used by Zaraki et al for their robot [12] and Peters et al for the virtual agent Greta [13]. However, these control models are often tuned for the exploration of natural scenes and do not take into account the context of the interaction: the head orientation is determined by eye movements. However, the head should be considered as a vector of attention like the eyes [14], and each of them conveys redundant as well as complementary communicative information. A fixed coordination model that does not take into account the context seems to neglect the social functions of these two vectors. We therefore hypothesized that for a robot's gaze to be natural, it must use both the head and the eyes and the coordination of these two should take into account the context of the interaction.

The goal of our study is (a) to evaluate if gaze patterns with decoupled head and eye control are actually perceived as more natural than basic models with eye-only, head-only or eye-dependent head movements and (b) if certain eyes-head coordination strategies that depend on the interaction context could be identified.

To do so, we first analyzed an original dataset where the eye and head orientations of a human pilot were recorded in a multiparty conversation using immersive teleoperation of a robot. We evidence that independent head and eyes control makes a difference in such conversational situations. We then produced videos in which a virtual Furhat robot [15] mimics the recorded human attention behavior, with different eyes-head coordination strategies. A crowd-sourced comparative evaluation was then conducted to rank these strategies.

1 Background

Gaze is a widely studied nonverbal cue in both Human-Human Interaction (HHI) and HRI [16]. Two aspects of the gaze have been studied, mostly independently: its social functions during interaction and its kinematics. In multi-party conversation, gaze is an essential cue of the floor regulation between participants. Multu et al [17] showed that a robot can regulate the roles of participants in a conversation through appropriate gaze behavior, while Skantze et al [18] used gaze to impact turn-taking behavior. But in these studies, information about the chosen coordination strategy is not given. However, an appropriate eyes-head coordination can also impact speech distribution. Based on a talk time ratio, Gillet et al [19] adapted the robot's gaze behavior to balance participation between subjects, by controlling the distribution of head orientations of the robot. The dialogue role-based robot gaze control proposed by Shintani et al [20] is perceived as more natural when it combines head and eye movements especially for gaze aversions.

Beyond knowing where the robot should look, eyes-head coordination also plays a role in managing multi-party conversations. The coordination of the head and eyes in humans has been studied for a long time, but without really taking into account the interaction contexts. A lot of information about the kinematics of the human gaze is already known. The gaze is the addition of head and eye movements, whose kinematics are different. In particular, the eyes react before the head and move faster [6]. The head contribution in the gaze movement is roughly a linear function of the amplitude of the gaze [21]; but for small gaze shifts, only the eyes move [22]. Stiefelhagen and Zhu [23] show that in a multi-party conversation with four people, the contribution of the head varies among humans, but performs about 70% of the total gaze movement. Inspired by neurobiological observations, several models of gaze control have been proposed. The most cited model was proposed by Itti [11]. It is widely used for virtual agents and robots [12]. Note that this model was developed for screening natural scenes, and therefore doesn't take into account any interaction with human agents. Several studies try to evaluate the naturalness of this kind of neurobiological models depending on head contribution (from 0% to 100%) [13, 24]. To evaluate the different coordination strategies, subjects watch videos of a virtual agent looking at objects on a table. They found that a head contribution of 0% is perceived as the least natural, while a contribution of 75% is perceived the most natural [24]. Unfortunately, no evaluation was performed in an interaction context such as a multi-party conversation. The social functions of the gaze are neglected and it is not known whether or not the context impacts the control and perception of head movements. The data that we will present and use in the next sections show that we should care about that.

2 Collection of naturalistic HRI

2.1 Data collection

The dataset used in this paper is the RoboTrio corpus [25]. The interaction takes part through a collaborative game (see Fig. 1) called Unanimo® and involves two human players and a robot teleoperated by a human (animator). The purpose of the game is to find the most popular words associated to a seed word. The animator spells the played theme out, e.g. "shrimp" and players deliberate on proposing their answers, maybe "pink", "seafood", ... The role of the animator is to motivate the dyad and to provide the rank of the scoring of the possible answers, displayed on a screen only available to him. Each session is composed of 9 seed words, with one for warming-up. The immersive teleoperation platform used in this study is the one proposed by Cambuzat et al [26]. The robot is an iCub robot [9] with extended communicative capabilities [27], such as speech generation with synchronized articulated lips and jaw. Thanks to the immersive teleoperation platform, the iCub reproduces the three degrees-of-freedom of head (pitch, roll, yaw), and eye (azimuth, elevation, vergence), lips and jaw movements of a human pilot while diffusing his voice through the mouth. The pilot perceives the subjects through the robot's sensors: the human pilot wears a virtual reality headset to hear and see the conversation captured by the robot's ears and eyes, as a pair of live video streams. Moreover, thanks to augmented reality, the pilot can see in front of him a tablet containing the various information of the game in progress (seed word, best answers and their scores). With this setup, the players usually think they interact with an autonomous robot endowed with a human-like behaviour. In return, the monitored animator behavior already takes into account HRI limitations. All the signals (input and output) are logged and can be replayed.

The corpus is composed of 22 sequences, each sequence lasts approximately 20 minutes. Between all the sequences, the pilot remains the same but the players are different and are composed of same-sex pairs.

2.2 Data annotation

For this study only five sequences of men pairs were fully annotated and used for the experiment. The multimodal behavior of the robot pilot and the players has been annotated :

- **Verbal cues:** Verbal content as well as intention of players and animator have been manually annotated for each sequence using the ELAN [28] tool. The intentions of the pilot and the players are not similar due to their asymmetric roles in the interaction. In total, we have defined 24 intentions for the robot pilot ("Theme announce", "Ask for a proposition", "Ask for validation", "Give positive scoring", "Give null score"...) and 9 intentions for the players ("Proposition", "Positive Feedback", "Negative Feedback"...).

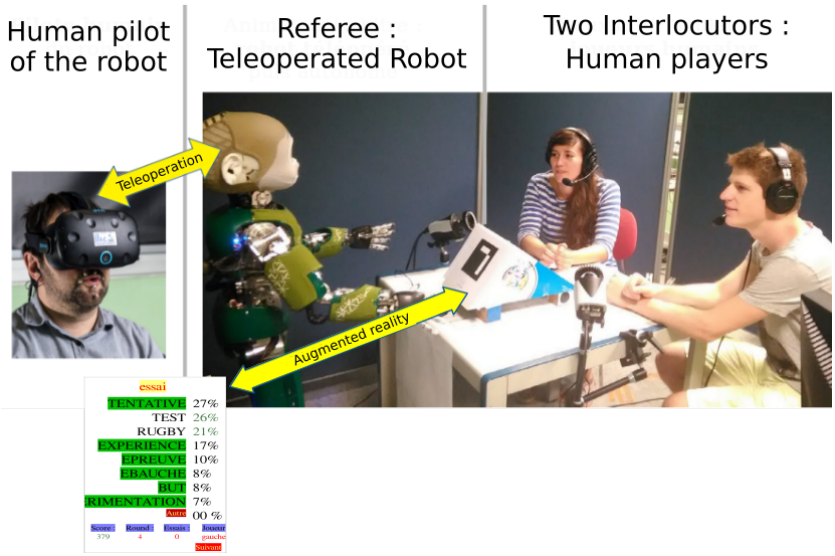

Fig. 1. Experimental setup for data collection of the RoboTrio corpus.

- **Pilot's gaze:** The recorded movement data of the pilot allowed us to compute the gaze focal points of the robot pilot in a virtual cylinder positioned at 1.2m from the robot (distance between the robot and the players). All the points corresponding to a fixation and not to a saccade have been automatically annotated with Gaussian Model Mixture (GMM). For the different classes, we have defined three regions of interest (RoI); "Left player", "Right player", and "Tablet", plus an "Elsewhere" class for gaze aversion. In addition to gaze focal points, we have also computed head focal points, corresponding to where the head is pointing on the same virtual cylinder. Head focal points have been classified in four classes with GMM; "Left", "Right", "Center", "Down".
- **Players' gaze:** We used Openface [29] to detect head and eyes orientations at each frame from the reference camera pointing to each player, from which we computed the corresponding focal point. If the focal point corresponds to a fixation, it is classified using GMM with one of the three labels: "Robot", "OtherPlayer", "Elsewhere".

3 Analysis of the human pilot's eyes-head coordination

In this study, we are interested in the contributions of head and eyes orientations to the gaze. Before comparing different coordination strategies, we have analyzed the human behavior of the robot pilot in RoboTrio. We first compared the position of the gaze focal points and those of the head computed during the annotation phase of the corpus. Fig. 2 shows the comparison of the distributions

of the focal points of the fixations depending on the head orientation for one sequence. Several observations can be made showing that **the orientations of the gaze and the head are not always identical.** First of all, the distribution of head focal points is much more restricted than that of the gaze. The contribution of head movements in the gaze seems to be closer to 30-40% than 100%, as implemented on several robots. Another observation is that a gaze point corresponding to a rightward head orientation (yellow) is not necessarily positioned in the rightmost RoI, and conversely for a gaze point whose head orientation is classified as leftward (blue). In an even more pronounced way, when the head is positioned in the center (red), between the two players, the gaze is mostly positioned on one of the two players and not on the middle.

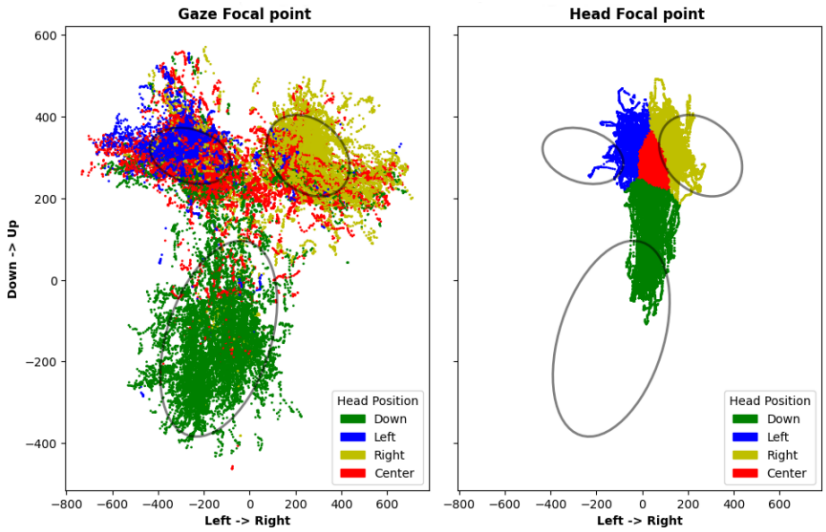

Fig. 2. Distribution of gaze and head focal points of the human pilot for one RoboTrio sequence. The colored point are obtained by a GMM classification of head focal points. The three ellipsoids show the Gaussians of the three ROIs (left player, right player, tablet) obtained with the GMM gaze classification.

We therefore try to better understand why at certain moments the pilot decides to position his head between the two players. **The hypothesis we made is that the head being a vector of attention, the pilot centers his head when he directs his attention on the two players, and directs it towards one player when he focuses his attention on only one.** To test this hypothesis, we decided to analyze and compare the orientation of the human pilot's head when addressing one or both players. To define the addressee, we detected in the utterances the use of the French pronouns "Vous" and "Tu". The "Vous" pronoun indicates that the pilot is addressing both players, while the "Tu" pronoun is used to address only one player. Of course, not all the

utterances of the pilot contain one of these two pronouns, so we had to limit our analysis to those containing them. For all utterances where a "Tu" was detected, we then manually annotated whether the "Tu" was directed at the left or right player. Then, we computed the median yaw angles performed by the head and the gaze (sum of head and eyes) during each utterance of 5 RoboTrio sequences annotated with "Tu" and "Vous". Fig. 3 shows the distribution of the median yaw angles of the pilot according to whether he pronounced a "Tu" or a "Vous" in the utterance. The left y-axis presents the bar plot distribution of the median angles, and the right y-axis shows the probability of the Gaussian distribution fitted on the median angles. We found that the distribution of head angles when the utterance contains a "Vous" is significantly more centered than when it contains a "Tu", which is not the case when we focus on the distribution of the gaze (combination of head and eye angles).

The verbal content of the interaction has thus an impact on the human pilot's behavior. **The contribution of the head in the gaze is weaker when the pilot addresses the two players than when he addresses only one of them. This result validates our hypothesis on the need to consider different eye-head coordination strategies in interaction, and in particular for multiparty conversations.**

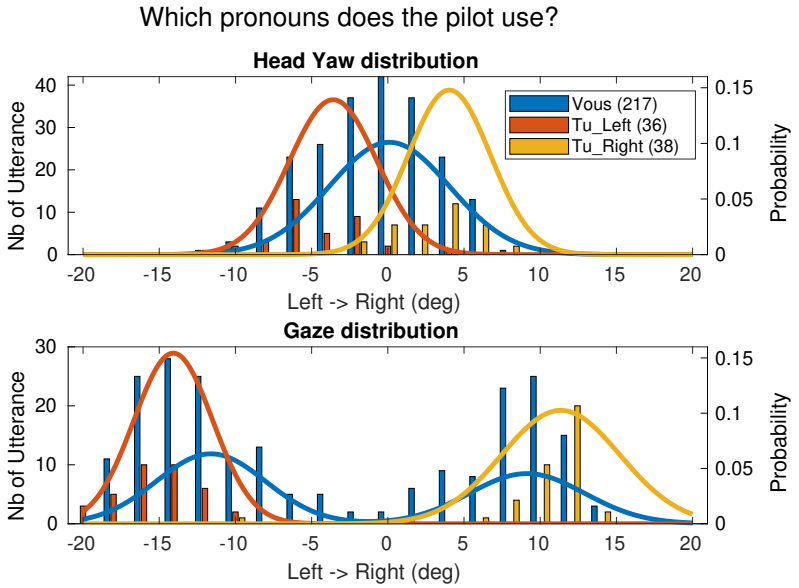

Fig. 3. Distribution of head and gaze median yaw angles according to who is/are addressed by the human pilot. The addressees are guessed thanks to the french pronouns, "Tu" for one player (left or right), "Vous" for both players. The left y-axis corresponds to the bar plot distribution, and the right y-axis to the probalility distribution after fitting a Gaussian distribution on the median yaw angles.

4 Subjective evaluation

In this study we exploit the data collected with the iCub robot to control a Furhat robot. While we always impose the gaze target of the original data, we could keep or modify the head part. To do so, we used the annotated gaze fixations of the pilot and the original verbal content so that Furhat could synthesize speech and **always attend the original gaze targets**. To perform these gaze fixations, different controls can be used for Furhat's head by imposing – or not – head movements matching the groundtruth data. This process allows **to compare different head/eyes contributions strategies while imposing the same gaze targets.**

4.1 Policies and hypotheses

In order to test the necessity to use and decouple head and eyes movements in attention management of Furhat, we decided to compare four policies.

- **EyesOnly policy:** Only the eyes of the robot move. The robot head is pointing to the center and is fixed. The trajectory of the eyes is computed by Furhat, according to the target that the robot must look at.
- **HeadOnly policy:** The head performs all the movement of the gaze. The eyes are enslaved to the head, as if they had no possibility to move freely. The pitch and yaw head angles are computed by adding eyes and head angles performed by the human pilot in our corpus. To smooth the trajectory and to avoid too fast movements for the head, a low-pass filter was applied on the eyes angles before the addition. The head roll of the head is kept as performed by the pilot.
- **Default Furhat policy (Baseline):** We used the default policy of Furhat. It computes eye and head movements from gaze targets. The eyes move first and faster than the head, but at the end of the movement both are aligned in the same direction.
- **EyesHead policy (Proposed policy):** This is the proposed policy, which is the closest to the pilot's behavior using the head and eyes. The three degrees of freedom of the head are kept identical to those performed by the pilot during the RoboTrio sequences. The eye trajectories to attend the imposed gaze targets are generated by Furhat.

From these four policies, we made two hypotheses:
(H1) The robot using EyesHead policy for attention management will be perceived as more natural than the others.
(H2) The preference between the other three policies will depend on the context of the interaction.

4.2 HEMVIP Evaluation

To evaluate these policies, we have decided to perform an on-line evaluation by third parties. We recorded video clips of Furhat (the way clips have been selected is explained in section 4.3) replaying interaction passages from RoboTrio

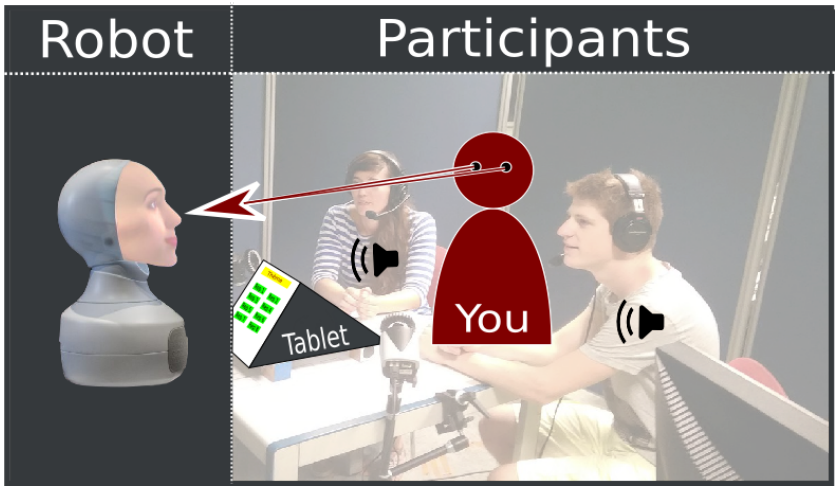

Fig. 4. Diagram presented in the introduction of the subjective evaluation to explain the perception context in the scoring interface: subject faces the robot, and can hear each player accordingly, i.e. on the right or on the left.

corpus, with the four control policies. For each interaction extract, four videos corresponding to the four policies were recorded. In these four videos, the verbal content and the robot's attention target are the same (left subject, right subject, tablet). Only the attention management are different according to the policy used. In order to keep environmental conditions identical, we recorded animations with the Furhat simulator. For all the clips, only the virtual robot is visible, the soundtracks of the RoboTrio's participants are broadcasted in stereo. In addition to explanations given in the experiment introduction, the context (see Fig. 4) is also shown to the subjects.

We used the HEMVIP[3] method [30] for the evaluation: evaluation is performed via several web pages; each page displays a panel with 4 sliders and "Play" buttons to score the 4 different video clips corresponding to the 4 different control policies for the same interaction extract. Subjects have to play each video clip at least once and give a score between 0 and 100 based on how natural they perceive the robot's behavior. The order of the control conditions is random. Each web page corresponds to a different extract of interaction, the order of the pages is also random: all subjects see and rate the same videos but not in a defined order. An example of the webpage is shown Fig. 5. At the end of the evaluation, subjects have to fill a general questionnaire about their familiarity with robots and can comment about the seen video clips.

[3] https://github.com/jonepatr/hemvip

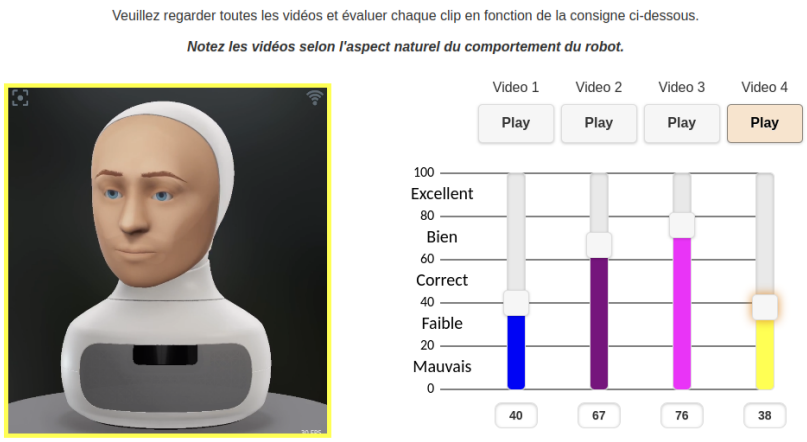

Fig. 5. Example of one of the fifteen web pages of the HEMVIP evaluation. The given instruction at the beginning means: "Rate the videos based on how natural the robot's behavior looks". Left area is used to display the video. Right area collects the evaluations with the four scales for the four videos, each having a dedicated play button.

4.3 Clips' selection

Short clips of interaction had to be selected for the evaluation. In order not to make the experiment too long, a total of 15 interaction clips were chosen, 3 clips for each of the 5 annotated sequences of the corpus. The average duration of these clips is 10.5 s, no clip is shorter than 9 s nor longer than 12 s. In addition to these 15 clips, 1 more clip was taken as a training clip for the subjects. The selection of the clips is achieved in three steps. No control policy has been favorized:

1. A first selection is performed automatically by focusing on the interaction moments where the head movements in the 4 control policies are the most different. For this, the sum of the absolute differences between the head Yaw angles (left/right) of the HeadOnly, the EyesOnly and the EyesHead policies are computed. We further add a constant according to whether there is verbal activity or not. The time course of these behavioral differences is smoothed by a median filter with a 10 s window. Then, peaks of maximum difference were detected. From these peaks, a first group of potential interaction extracts was obtained (red points on Fig. 6).
2. Once the peaks are detected, it is necessary to check that they correspond to relevant moments of the interaction (e.g. not corresponding to the explanation of the rules for example) and that the on-going context is clear, without the previous seconds. As a result, some ambiguous passages are rejected.
3. To finalize the selection, the main purpose was to diversify the extracts. This was arbitrated by the content of the interaction, to not only have passages where the animator gives the score, but also to have moments where the players debate between them, or make a proposal, etc. Moreover, it was also

done on the context of the interaction, so that the themes of the game are varied, and that passages are selected at the beginning, middle and end of the game sessions.

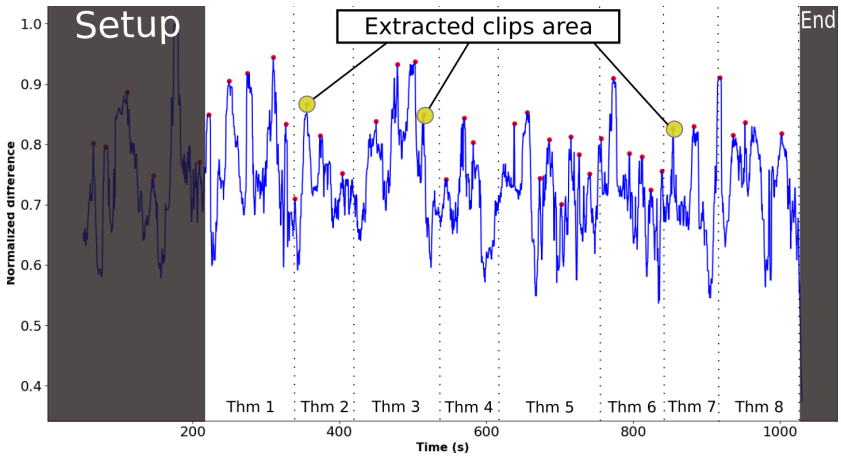

Fig. 6. Time course of the normalized difference between the head yaw trajectories of the different policies for one sequence. The position of the peaks of maximum differences are marked in red. The peaks corresponding to the three video clips finally selected for this sequence are marked with yellow dots.

4.4 Participants

Participants were recruited via the Prolific[4] platform. Access to this experience was restricted to French speakers residing in France, Belgium or Switzerland, so that participants can fully understand the verbal context of the videos. A total of 51 people, aged between 18 and 60, completed the evaluation. One submission was rejected, the completion time being too short. In the end, 50 submissions are considered, with a balanced number of Female and Male.

5 Results

The result of the subjective experiment are shown Fig. 7 and Fig. 8. The statistical significance of the distributions of subjective ratings has been studied by a beta regression with *clips_Id* and *users_Id* as random variables using the glmmTMB package [31] of R software [32]. Using a likelihood ratio test, we found that the policy significantly impacts the rated score (chisq(3)=744.53,

[4] https://www.prolific.co/

p<0.0001). We then conducted multiple pair-wise comparisons between the policies using the multcomp package [33] of R software; the Fig. 7 shows the adjusted p-value obtained. The EyesHead policy is significantly higher rated than the other policies. The closest coordination strategy from the human behavior is clearly perceived as more natural. The HeadOnly policy is the second highest rated policy, but strongly worst than the former. Ratings of Furhat and EyesOnly are not statistically different but significantly lower than the two preceding ones. With the beta regression, we found that familiarity significantly impacts the scores too (chisq(12)=197.84, p<0.0001). Nevertheless, for all the familiarity values, the EyesHead policy is the best rated. **The hypothesis (H1) is verified.** Moreover, we found that *clips_Id* significantly impacts the rated score too (chisq(9)=64.5, p<0.0001). But even if the rated score is not the same between the video clips, the EyesHead policy is always the highest rated (see Fig. 8). For the other policies, most of the time the HeadOnly policy is the second highest rated but it's not always the case. For example, for the "C" video clip, the EyesOnly policy scored higher than the two other policies, which probably means that smaller head movements are preferred for this interaction extract. For the "F" video clip, the Furhat policy obtained the second best score. **The (H2) hypothesis is also verified: depending on the extract of interaction the preference between the three other policies is different.**

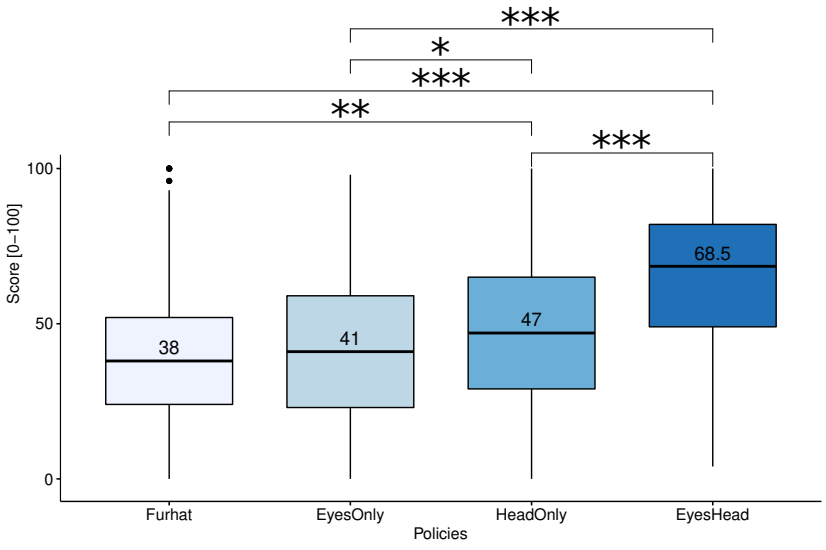

Fig. 7. Results of the subjective experience. Comparison of the reported naturalness-score, according to the policy. Each boxplot contains distributions of 50x15 points (number of subjects x number of clips). Significant p-values are indicated by * (<0.05), ** (<0.01) and *** (<0.001).

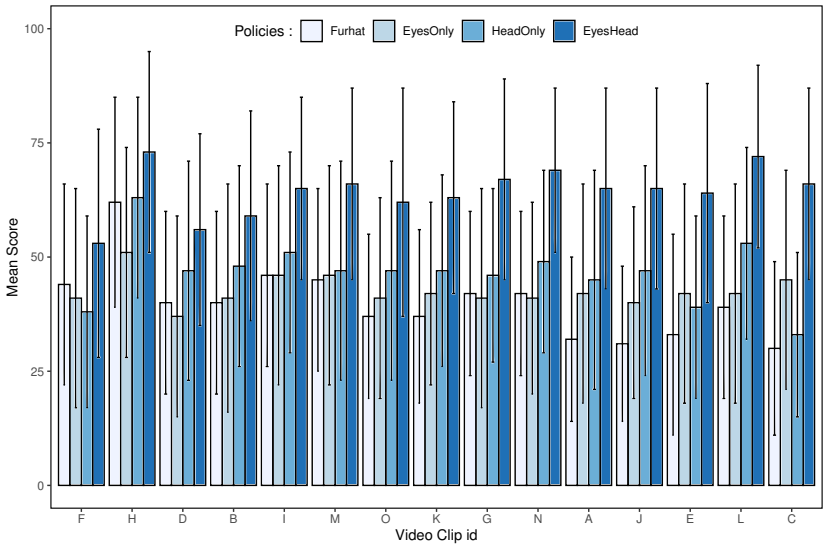

Fig. 8. Results of the subjective experience. Comparison of the average naturalness scores given for each policy according to the video *clip_Id*. Video clips are ordered by increasing difference between mean scores obtained by EyesHead and the mean of the other policies.

6 Discussion

In this study, we replayed multimodal human behavior recorded in multiparty conversation with different eyes-head coordination strategies for the robot's gaze. We show with a subjective online evaluation that driving only eyes or only head is less natural than the simulation of the original data involving both channels. Surprisingly, the Furhat policy using both eyes and head movements that align at the end doesn't obtained very good scores compared to the other policies. Using both vectors doesn't seem to be enough to manage gaze behavior in a natural way, even if the control takes into account biological aspects of human gaze, such as the vestibulo-ocular reflex. Multiple reasons could explain this low score. First, the default Furhat policy is not meant to cope with multiparty conversation, in particular with such context as collaborative games. Moreover, the head movement of the Furhat policy is quite slow, with a large amplitude. So when the pilot quickly shifts his gaze between two targets, the robot's behavior is not very natural. The better results obtained by the HeadOnly policy could be explained by the speed of the head which is much faster and so more natural than the default Furhat policy. Globally, the three policies (Default Furhat, EyesOnly and HeadOnly) have fixed coordination between head and eyes, and this lacks the variability that the multiparty context deserves. However, the results of Fig. 8 show different rating scores between the polices depending on the extract of the interaction. The preferred coordination between head and eyes depends

on the context of the interaction. This result is consistent with the previous VOUS/TU analysis of the human pilot's behavior that we have conducted about the difference between head movements according to the pilot's addressee.

Nevertheless, a first limitation of our study is that, except the EyesHead policy, the three others have a head movement propensity fixed at 0% (EyesOnly) or 100% (Default Furhat, HeadOnly). There is no intermediate head contribution as is proposed in [13, 24]. It would be interesting to compare EyesHead policy with policies with non extreme head movement contributions. Moreover, the different strategies were compared only on the naturalness of the behavior. Other questions could have been asked to the subjects concerning for example the personality of the robot, or the understanding of its intentions. Similarly, with the human behavior replay method, the subjective evaluation was performed with a third-person perspective. We therefore didn't have access to the feelings of a person who experienced the physical interaction with the robot. Note however that everything was done to allow the subject to understand the context of the interaction as well as possible. Pereira et al [34] showed that a third-person evaluation of gaze patterns in HRI provided similar results to a first-person evaluation.

Other limitations of this study have been identified for potential future improvements. For example, natural blinks were not transfered on the robot. We chose to use the model already implemented on the Furhat robot. So the blinks don't systematically occur at the same time between policies, but they are generated by the same model that is unaware of the cognitive activity of the robot nor it's communicative intentions. Similarly, the possible gaze targets for the robot are limited to the 3 RoI (left player, right player, tablet): there is no gaze aversion in the gaze replay nor gaze paths over the subjects' faces [35]. However, we hypothesize that for this evaluation in this interaction context, aversions are not paramount. The pilot never fixes at a player for very long time, and he can use the tablet to drop out of the ongoing conversation.

Another topic of discussion could be the use of Mixed Reality to collect groundtruth data on human behavior. Indeed, wearing a virtual reality headset could impact the behavior of the pilot. Pfeil et al [36] compared eyes-head coordination in physical and virtual environments and showed that subjects in virtual environments seemed to use their heads more, but in the study by Sidenmark et al [5] no significant difference was found. It is therefore difficult to conclude on some impact in our case, especially as we display real video streams rather than synthetic content.

Anyway, despite these possible limitations, our results show a strong preference for the coordination strategy combining head and eye movements in a very realistic way.

7 Conclusions and perspectives

This study argues for the independent control of the head and eyes movements for the generation of the robot gaze. Both body segments provide redundant and

complementary information about the conversational regime and communicative intentions throughout the interaction. Indeed, the orientation of the head seems to be a key element in the regulation of multiparty conversations, for example to communicate to whom we are addressing. We notably show that the head orientation may contribute to the identification of a message, delivering somehow the median direction of an "attention cone" in the assembly of attendees.

In the future, we will try to exploit the Robotrio corpus to train a multimodal gaze control model for our robot that takes into account its communicative intentions and the overt verbal and non verbal responses of the interlocutors.

The coordination between eyes, head and possibly other segments of the body depends on numerous factors such as the actual physical disposition of the interlocutors, their social roles and status (see the Multidimensional Dimensional Scaling analysis performed on gaze models in [37]) as well as the context of the interaction. We will see how a multimodal gaze control model can be biased by these physical and social settings.

Acknowledgements.

This work is supported by the ANR 19-P3IA-0003 MIAI. The first author is financed by a CIFRE PhD granted by ANRT. The authors thank Juliette Rengot for the automatic gaze classification tool and Nathan Loudjani, who was invaluable in the RoboTrio corpus recording (CNRS SI2H PEPS funding).

References

1. Kendon, A.: Some functions of gaze-direction in social interaction. Acta psychologica 26 1, 22–63 (1967)
2. Sacks, H., Schegloff, E., Jefferson, G.: A simple systematic for the organisation of turn taking in conversation. Language 50, 696–735 (12 1974)
3. Vertegaal, R., Slagter, R., van der Veer, G., Nijholt, A.: Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. p. 301–308. Association for Computing Machinery, New York, NY, USA (2001)
4. Ishii, R., Otsuka, K., Kumano, S., Yamato, J.: Predicting who will be the next speaker and when in multi-party meetings. NTT Technical Review 13 (07 2015)
5. Sidenmark, L., Gellersen, H.: Eye, head and torso coordination during gaze shifts in virtual reality. ACM Trans. Comput.-Hum. Interact. 27(1) (dec 2019)
6. Freedman, E., Sparks, D.: Coordination of the eyes and head: Movement kinematics. Experimental brain research 131, 22–32 (04 2000)
7. Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J., Maisonnier, B.: Mechatronic design of nao humanoid. In: 2009 IEEE International Conference on Robotics and Automation. pp. 769–774 (2009)
8. Kristoffersson, A., Coradeschi, S., Loutfi, A.: A review of mobile robotic telepresence. Advances in Human-Computer Interaction 2013, 3–3 (2013)

9. Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., von Hofsten, C., Rosander, K., Lopes, M., Santos-Victor, J., Bernardino, A., Montesano, L.: The iCub humanoid robot: An open-systems platform for research in cognitive development. Neural Networks 23(8), 1125–1134 (Oct 2010)
10. Pateromichelakis, N., Mazel, A., Hache, M.A., Koumpogiannis, T., Gelin, R., Maisonnier, B., Berthoz, A.: Head-eyes system and gaze analysis of the humanoid robot romeo. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1374–1379 (2014)
11. Itti, L., Dhavale, N., Pighin, F.: Photorealistic attention-based gaze animation. In: 2006 IEEE International Conference on Multimedia and Expo. pp. 521–524 (2006)
12. Zaraki, A., Mazzei, D., Giuliani, M., de rossi, D.: Designing and evaluating a social gaze-control system for a humanoid robot. IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans 44, 157–168 (04 2014)
13. Peters, C., Qureshi, A.: Graphics for serious games: A head movement propensity model for animating gaze shifts and blinks of virtual characters. Computers & Graphics 34, 677–687 (12 2010)
14. Hietanen, J.K.: Does your gaze direction and head orientation shift my visual attention? Neuroreport 10 16, 3443–7 (1999)
15. Al Moubayed, S., Beskow, J., Skantze, G., Granström, B.: Furhat: A back-projected human-like robot head for multiparty human-machine interaction. International Journal of Humanoid Robotics (01 2013)
16. Admoni, H., Scassellati, B.: Social eye gaze in human-robot interaction: A review. Journal of Human-Robot Interaction 6, 25 (03 2017)
17. Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., Hagita, N.: Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction. p. 61–68. Association for Computing Machinery, New York, NY, USA (2009)
18. Skantze, G., Johansson, M., Beskow, J.: Exploring turn-taking cues in multi-party human-robot discussions about objects. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. p. 67–74. Association for Computing Machinery, New York, NY, USA (2015)
19. Gillet, S., Cumbal, R., Pereira, A., Lopes, J., Engwall, O., Leite, I.: Robot gaze can mediate participation imbalance in groups with different skill levels. In: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. p. 303–311. Association for Computing Machinery, New York, NY, USA (2021)
20. Shintani, T., Ishi, C.T., Ishiguro, H.: 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. p. 332–336. Association for Computing Machinery, New York, NY, USA (2021)
21. Zangemeister, W., Stark, L.: Types of gaze movement: Variable interactions of eye and head movements. Experimental Neurology 77 3, 563–577 (1982)
22. Fuller, J.H.: Comparison of Head Movement Strategies among Mammals. In: The Head-Neck Sensory Motor System. Oxford University Press (1992)
23. Stiefelhagen, R., Zhu, J.: Head orientation and gaze direction in meetings. In: CHI'02 Extended Abstracts on Human Factors in Computing Systems. p. 858–859. Association for Computing Machinery, New York, NY, USA (2002)
24. Pejsa, T., Andrist, S., Gleicher, M., Mutlu, B.: Gaze and attention management for embodied conversational agents. ACM Transactions on Interactive Intelligent Systems 5, 1–34 (03 2015)

25. Prévot, L., Elisei, F., Bailly, G.: Robotrio (2020), https://hdl.handle.net/11403/robotrio/v1, ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr
26. Cambuzat, R., Elisei, F., Bailly, G., Simonin, O., Spalanzani, A.: 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. pp. 232–239. VDE, Munich, Germany (2018)
27. Parmiggiani, A., Randazzo, M., Maggiali, M., Metta, G., Elisei, F., Bailly, G.: Design and validation of a talking face for the icub. International Journal of Humanoid Robotics 12 (09 2015)
28. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: ELAN: a professional framework for multimodality research. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). European Language Resources Association (ELRA), Genoa, Italy (May 2006)
29. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 59–66 (2018)
30. Jonell, P., Yoon, Y., Wolfert, P., Kucherenko, T., Henter, G.E.: Hemvip: Human evaluation of multiple videos in parallel. In: Proceedings of the 2021 International Conference on Multimodal Interaction. p. 707–711. Association for Computing Machinery, New York, NY, USA (2021)
31. Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Maechler, M., Bolker, B.M.: glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R Journal 9(2), 378–400 (2017)
32. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2022), https://www.R-project.org/
33. Hothorn, T., Bretz, F., Westfall, P.: Simultaneous inference in general parametric models. Biometrical journal. Biometrische Zeitschrift 50, 346–63 (06 2008)
34. Pereira, A., Oertel, C., Fermoselle, L., Mendelson, J., Gustafson, J.: Effects of different interaction contexts when evaluating gaze models in hri. In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. p. 131–139. Association for Computing Machinery, New York, NY, USA (2020)
35. Bailly, G., Raidt, S., Elisei, F.: Gaze, conversational agents and face-to-face communication. Speech Communication 52(6), 598–612 (2010)
36. Pfeil, K., Taranta, E.M., Kulshreshth, A., Wisniewski, P., LaViola, J.J.: A comparison of eye-head coordination between virtual and physical realities. In: Proceedings of the 15th ACM Symposium on Applied Perception. Association for Computing Machinery, New York, NY, USA (2018)
37. Mihoub, A., Bailly, G., Wolf, C.: Social behavior modeling based on incremental discrete hidden markov models. In: Human Behavior Understanding: 4th International Workshop, HBU 2013, Barcelona, Spain, October 22, 2013. Proceedings 4. pp. 172–183. Springer (2013)