Data-Driven Generation of Eyes and Head Movements of a Social Robot in Multiparty Conversation

Léa Haefflinger, Frédéric Elisei, Béatrice Bouchot, Brice Varini, Gérard Bailly

▶ **To cite this version:**

Léa Haefflinger, Frédéric Elisei, Béatrice Bouchot, Brice Varini, Gérard Bailly. Data-Driven Generation of Eyes and Head Movements of a Social Robot in Multiparty Conversation. ICSR 2023 - 15th International Conference on Social Robotics (ICSR 2023), Dec 2023, Doha, Qatar. pp.191-203, 10.1007/978-981-99-8715-3_17 . hal-04335472

HAL Id: hal-04335472
https://hal.science/hal-04335472

Submitted on 11 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Data-driven Generation of Eyes and Head Movements of a Social Robot in Multiparty Conversation

Léa Haefflinger[1,2][0009−0009−6592−040X], Frédéric Elisei[1][0000−0002−1295−3445], Béatrice Bouchot[2], Brice Varini[2], and Gérard Bailly[1][0000−0002−6053−0818]

[1] GIPSA-Lab, Grenoble-Alps Univ., France
[2] Atos, France

Abstract. Given the importance of gaze in Human-Robot Interactions (HRI), many gaze control models have been developed. However, these models are mostly built for dyadic face-to-face interaction. Gaze control models for multiparty interaction are more scarce. We here propose and evaluate data-driven gaze control models for a robot game animator in a three-party interaction. More precisely, we used Long Short-Term Memory networks to predict gaze target and context-aware head movements given robot's communication intents and observed activities of its human partners. After comparing objective performance of our data-driven model with a baseline and ground truth data, an online audiovisual perception study was conducted to compare the acceptability of these control models in comparison with low-anchor incongruent speech and gaze sequences driving the Furhat robot. The results show that our data-driven prediction of gaze targets is viable, but that third-party raters are not so sensitive to controls with congruent head movements.

Keywords: Human-Robot Interaction · Gaze · AI · Head · Multiparty

1 Introduction

The importance of non-verbal cues in human conversations is no longer to be proven: authors of [4] consider that 60% of communication intents would pass through this channel. For a robot to interact with humans in the most natural way, it must be able to perceive, understand and generate such cues.

One of the most studied non-verbal cues for Human-Robot Interaction (HRI) is gaze [1]. And rightly so, it is a major social cue in face-to-face Human-Human interactions (HHI). Indeed, in addition to transmitting emotions, it is a powerful regulator of conversations [16, 27]. This function is particularly important in multiparty conversations, for turn-taking management and role detection, such as who will be the next speaker, or who is the current addressee [13, 32]. This impact of the gaze has also been emphasized in HRI. For example, the gaze control proposed by Multlu et al [22] allowed their robot to signal the roles of participants in the conversation (bystander, overhearer, ...). In the same way,

thanks to its gaze, a robot can influence turn-taking behaviors [30] and even regulate speaking times [10]. In addition to having an impact on the conversational regime, appropriate gaze behaviour increases participants' engagement in the conversation and positively impacts their perception of the robot [29, 18, 6].

All these studies confirm the importance of providing our robot with the most natural gaze control possible. This study introduces a unique method to control the robot gaze in a multi-party interaction by combining two Long Short-Term Memory (LSTM) models, one to predict the attention targets of the robot, and one to generate the corresponding head movements. We believe that the use of LSTM models will enable the generation of more subtle behaviours based on elements of the interaction context that the robot can perceive (who speaks, where the interlocutors look, ...) or related to its own intentions (who it is addressing, what it is talking about, ...). These models will first be evaluated objectively, then subjectively, through an online perception study, in order to compare them with a baseline model and ground-truth behaviours.

2 Related Works

Given the importance of gaze in HRI, a large number of gaze models have already been proposed and tested [1]. However, most of these models are developed for dyadic interactions and not for multiparty interactions as in this study. Two categories of models can be distinguished, models based on human interaction data, called data-driven, and those using rules extracted from human behavior, called heuristics. On the side of multiparty heuristic models, we can find the model proposed by Zaraki et al [34], where each participant gets a coefficient of attention computed from multimodal cues, or the model proposed by Mishra et al [20] for a robot playing a game with two humans. For data-driven models, Mutlu et al [21] proposed a control to monitor roles of participants, Nakano et al [23] built a model taking into account dominance in a conversation, and Shintani et al [28] focused on gaze behavior during turn-taking. Some models use machine learning algorithms, as proposed by Stefanov et al [31] who tested artificial neural networks using or not LSTM to model attention, or Huang et al [12] who used Support Vector Machine (SVM) for their gaze prediction model.

Furthermore, beyond the prediction of gaze targets, this study also focuses on the generation of head movements that allow subtle control of head-eye coordination. Head-eye coordination has been extensively studied in humans [33, 8, 7]. For HRI head is mostly considered as a passive contributor of eye movements but not a component per se of the robot's communicative intentions [14, 34, 3]. However, Gillet et al [10] were able to influence participants' speaking times by manipulating the head movements. Among the few studies that have implemented a context-aware control of head-eye coordination, are the model of Mishra et al [20] where the contribution of the robot's head depends on the duration of fixation of the attention target, and the models proposed by [31, 24] that predict both eye-gaze direction and head orientation.

Fig. 1. Setup for the RoboTrio data collection.

The two studies most similar to our method are those by Stefanov et al [31] and Huang et al [12] through the use of machine learning algorithms. However, Stefanov et al [31] did not propose a subjective evaluation of their models and our study differs from that of Huang et al [12] due to the roles beeing asymetric in our interaction (robot plays game as animator), and their robot could not move its head and its eyes. Another major difference is that we use the robot's addressee as an input of our model. As shown in [11], the contribution of the head in the gaze depends on whether one or two people are addressed at a time.

3 Creating the Models

3.1 Gaze/head data collection: immersive robot teleoperation

To train and evaluate our models, we use multimodal data from three-party interactions in a collaborative game context [26], Figure 1. The game is scored by finding the most quoted words for a given theme (previously played online by human players). E.g. for the "sea" theme, the words that would score the most are "ocean", "water", "beach", "mediterranean", "boat" and "fish".

The behavior we want to model is that of the game's animator. This animator is in fact an iCub robot [19] controlled by a human pilot through immersive teleoperation [5]. This setup allows to interact with two human players through the robot sensors and actuators. A tablet is placed in front of the robot so that the pilot can scan the information about the game in progress. The animator must report the themes, invites the players to propose words, and then reports the scores for the proposed words. **All head and eye movements of the pilot, including vergence, are reproduced in real time by the operated robot** (3+3 Degrees of Freedom, aka DoF). These gaze and head movements, as well as audio and video of the three-party interaction, are recorded as "the corpus". We use 11 recorded and annotated game sequences, with different pairs of players, but keeping the robot's pilot the same. This amounts to almost 4 hours of recording, where each sequence lasts about 20 minutes. A sequence consists of 9 rounds (new theme word), and 5 collected answers per theme. While playing, the players collaborate to find the best answers and look/ask the robot at will. So there's a lot of interaction and

social cues; thinking about the theme, sharing and gauging ideas on a potential answer, etc. The robot monitors them like its human pilot would do, and is included regularly in the conversation. **The corpus is therefore complex and rich in verbal and non-verbal content** for the players and the robot (mutual gaze, gaze aversion, speech overlap ...).

3.2 Models Implementation

In order to make the attention control of our robot as natural as possible, we propose to generate both attention targets and head movements. For this purpose, we decided to cascade two models.

Tasks definition Our first model predicts gaze targets. To train this model, the gaze of the robot pilot was classified with Gaussian Mixture Models (GMM). After detection of the ocular saccades, the gaze was divided into 4 classes of attention; one for the leftmost player in the game *UserL*, one for the rightmost player *UserR*, one for the $Tablet$ screened by the animator, and finally a class $Elsewhere$. To simplify the training of our model, the frames where the gaze is classified as $Saccade$ or $Elsewhere$ are grouped into an $Other$ class. In addition, to filter out errors due to classification, fixations with a duration of less than 150 ms were merged with the preceding fixations. The distribution of gaze classes in the dataset is not completely balanced, with *UserL* and *UserR* representing 32.4% and 32.1% respectively, while $Tablet$ represents 21.7% and $Other$ 13.8%.

Then, the second model predicts the three DoF of the head: pitch (up/down), roll (tilt), yaw (left/right).
The outputs of both models are generated continuously at 60 Hz.

Input multimodal features Multimodal features about the activity of the pilot and the players are given as input to both models. These features have been selected against others as they can be observed in real-time (targeting a future implementation with our Furhat robot [2]). Each feature, when composed of N classes, is decomposed into N channels, with only 0 or 1 values:

- 11 channels for **Robot pilot activity**:
  - *Speech:* whether pilot is speaking or not
  - *Speech Intent:* intent of the sentence, 7 different classes (ask for a proposition, give the score, the theme, an explanation or feedback,...)
  - *Addressee:* pilot's adressee(s) *UserL*, or *UserR* or *Both*, value is 0 for the three channels if the addressee is unknown
- 6 channels for **UserL and UserR activities**:
  - *SpeechL, SpeechR:* whether left (resp. right) user is speaking or not
  - *GazeL:* 2 classes, whether given user is looking at the other user, or at the robot. Value is 0 for both channels if the user is looking at elsewhere
  - *GazeR:* same as previous, but for the right user.

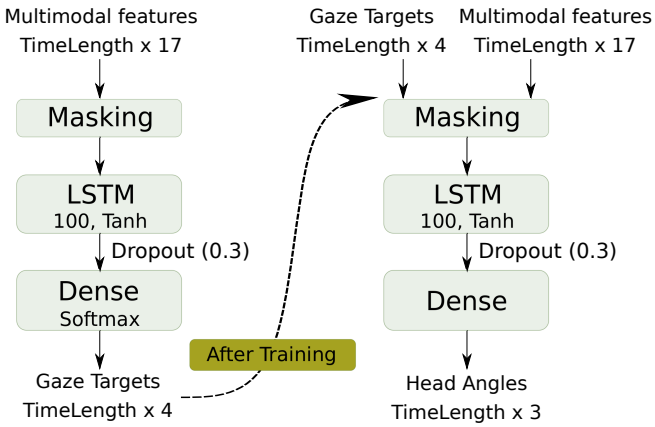

Fig. 2. Structure of the models, the one predicting gaze targets on the left, and the one generating head movements on the right.

In addition to these **17 input channels**, the model generating head angles also receives pilot gaze features, **adding 4 channels for a total of 21.**
Verbal features were annotated manually, while users gaze features were automatically annotated using GMM. The robot's addressee was annotated using the French pronouns "Vous" (*Both*) and "Tu" (*UserR/L*), see [11].

Models training To take into account possible temporal dependencies, the models use LSTM cells. Their input are temporal sequences, which have been cut to correspond to a whole game theme. The 11 interactions being composed of 9 themes, we obtain a total of 99 mini-batches. As the duration of the themes is variable and the input of the networks must be of fixed dimension, padding was applied to standardize the length of the sequences (TimeLength). The models have a many-to-many architecture; their output is a temporal sequence whose length matches the input. The structure and parameters of the two models are presented in Figure 2. The masking layer is used for padding detection. The main differences between these two models are the input and output dimensions, and the use of a "softmax" activation function for gaze target classification. The networks are trained with 200 epochs, a batch size of 10, and an Adam optimizer [17] with a learning rate of 10^{-4}. For the gaze target classification model, the loss function is the categorical crossentropy, and for the head angle regression model the loss function is the Mean Squared Error (MSE).

Model performances To best evaluate the performance of our models, we used the K-fold cross validation method. For each training, the test dataset is composed of the n-th theme of each sequence (11 temporal sequences), and the training dataset of the 8 others (88 temporal sequences). The two networks are thus each trained and evaluated 9 times with different datasets (9 folds). The

Table 1. F1-score of the gaze classification model according to the interaction context.

Gaze Class	Pilot is Speaking	Pilot is Listening	No One is Speaking	ALL
UserL	0.50	0.54	0.51	0.52
UserR	0.52	0.59	0.51	0.55
Tablet	0.77	0.41	0.45	0.67
Other	0.05	0.01	0.03	0.03
Weighted F1-score	0.54	0.48	0.44	0.49

average accuracy of the attention target classification model is 52.9±1.4%. The average MSE of the head angle generation model is 7.93±0.51.

In Table 1, the performance of the gaze classification model is analysed in detail, by calculating the F1-score for each class depending on the robot pilot's activity. First, we notice that the *Other* class is particularly badly predicted, which is not surprising as it does not correspond to a specific target and acts as a garbage collector. Moreover, contrary to the results presented in [31, 12], the proposed model is better when the pilot speaks than when he listens. This can be explained by the pilot's role as the game animator who, when speaking, will often look at his tablet to consult the game information. Moreover, the model knows the verbal intention of the pilot, which is not the case when a user speaks.

Ablation Study To study the influence of each input feature on model performance, we conducted an ablation study (Table 2). To do this, we trained separately our models under the same conditions as before (same parameters and 9-fold cross-validation) but removed selected input features (✗ in Table 2). Removing the *Intent* feature has the biggest impact on gaze prediction, which can be explained by the importance of intentions in determining whether the robot should look at the tablet (theme announcement, scores) or specific/both players (ask for proposal, validation) when speaking. When all *Robot* features are removed, performance drops drastically, and the same applies to *Users* features. It therefore seems interesting to take into account both endogenous and exogenous information from the robot.
For head generation, *Robot_Gaze* is clearly the feature that provides the most information. Nevertheless, when all *Robot* features are removed, performance drops even further, assuming that the other *Robot* features are also important.

4 Subjective Evaluation

4.1 Method

Goal The objective evaluation of model performance is not decisive. Indeed, it is not because the predicted target is different from the original target that this choice is less relevant or natural, and the same for the generation of head movements. A subjective evaluation is therefore necessary to validate the viability of our proposed attention control. To evaluate it, we predicted the gaze targets, for

Table 2. Results of the ablation study on input features.

Robot			Users		**Gaze prediction: Accuracy**
Speech	Intent	Addressee	Speech	Gaze	
-	-	-	-	-	$52.9 \pm 1.4\%$
-	✗	-	-	-	49.1 ± 0.8
-	-	-	-	✗	50.9 ± 1.3
-	-	-	✗	-	51 ± 1.2
-	-	✗	-	-	51.5 ± 1.8
✗	-	-	-	-	52.5 ± 1.1
✗	✗	✗	-	-	45.0 ± 1.7
-	-	-	✗	✗	47.9 ± 1.3

Robot				Users		**Head prediction: MSE**
Speech	Intent	Addressee	Gaze	Speech	Gaze	
-	-	-	-	-	-	7.93 ± 0.51
-	-	-	✗	-	-	13.48 ± 0.66
-	-	-	-	-	✗	8.26 ± 1.04
-	-	✗	-	-	-	7.97 ± 0.46
-	✗	-	-	-	-	7.96 ± 0.47
-	-	-	-	✗	-	7.92 ± 0.44
✗	-	-	-	-	-	7.91 ± 0.45
✗	✗	✗	✗	-	-	16.60 ± 0.57
-	-	-	-	✗	✗	7.88 ± 0.38

each frame (60 Hz), for all 11 sequences, and reused these predictions as input to the head generation model trained on ground-truth data (see Figure 2). For each theme, the models used for prediction were those that were not trained with that theme. Finally, the predictions were filtered, removing gaze fixations shorter than 150 ms, and smoothing head movements with a Blackman filter. These attention behaviors are evaluated in this section.

Compared attention controls We decided to compare our cascaded data-driven model with 3 other controls, to test both the prediction of attention targets, and the generation of head movements. To do so, we replay sequence of game interactions on a virtual Furhat robot [2] with different gaze and head behaviors. Between the different conditions, the verbal content is identical and is synthesized by Furhat, only the control of its eyes and head differs. The possible targets of attention are limited to *UserL*, *UserR* and *Tablet*. The different four models are listed below :

- **Lstm Model:** The proposed data-driven control that combines the two LSTM models described in this paper, that take into account the interaction context. All three angles of the head are controlled.
- **Heuristic Model:** This model focuses on the head movement generation, it is close to the observation model proposed by [14]. The robot looks at the same targets as the *LstmModel*, and the head movements are generated from these targets only, without taking into account the context (no pilot and

players activities). The model calculates the distance between two fixations, if it is lower than a threshold value, the head does not move, otherwise it performs a defined percentage of the path. The percentages were set to be as close as possible to what was done by the pilot, 30% for the yaw angle and 45% for the pitch angle. For the calculation of the threshold value, and the trajectories of the head, we used the equations proposed by Itti et al in [14]. Moreover an attraction-middline effect is also implemented for better realism [9]. This control uses only 2 DoF, pitch and yaw. The Table 3 shows the Root Mean Squared Error (RMSE) between the head angles of the control and those of *LstmModel*. Logically, the error is maximum on the Roll angle, this angle being equal to zero for this *HeuristicModel* control. The differences being not negligible, we hypothesized that they will be perceived during the perceptive study.

- **Simulated Ground Truth (GT):** *High_Anchor* This control corresponds to the original human behaviour of the pilot in the data collection, same attention targets, and same head movements (pitch, roll, yaw). In Table 3, RMSE between this control and the two previous ones are high, since targets of attention are not necessarily the same.
- **Shifted Ground Truth:** *Low_Anchor* This control is the same as the previous one, but uses data shifted in time. The robot will reproduce the same behavior as the pilot but 1 minute ahead. The head movement corresponds to the current target of attention, but this target is incongruous. As sustained conversational states last several seconds, we chose a 1 minute shift to get a close context without matching the original attention targets.

Online evaluation For the perception evaluation, 21 clips of interaction were selected and extracted, 2 per game sequence plus 1 for the initial training example. These selections correspond to extracts where the head movements between the conditions *HeuristicModel* and *LstmModel* differ the most. Each of these extracts of interaction result in 4 video clips of about 10 s, corresponding to the 4 controls to be compared. Only the virtual robot is visible on these videos, the players are perceived and differentiated using stereo audio, with the left (right) speaker using the left (right) audio channel. For the evaluation, we used the HEMVIP [15] method. On each page, the subjects compare 4 renderings of the same interaction segment. They must rate each video between 0 and 100. 20 web pages corresponding to the 20 extracts are presented in a random order, as well as the 4 videos of the different controls. The evaluation instruction given to the

Table 3. Root Mean Squarred Error between (RMSE in degree) the head angles of the different controls.

Comparison	Pitch	Roll	Yaw	ALL
$LstmModel$ vs $HeuristicModel$	1.73	2.56	3.61	2.63
$LstmModel$ vs $SimulatedGT$	4.72	4.28	2.66	3.89
$HeuristicModel$ vs $SimulatedGT$	4.85	5.00	4.49	4.78

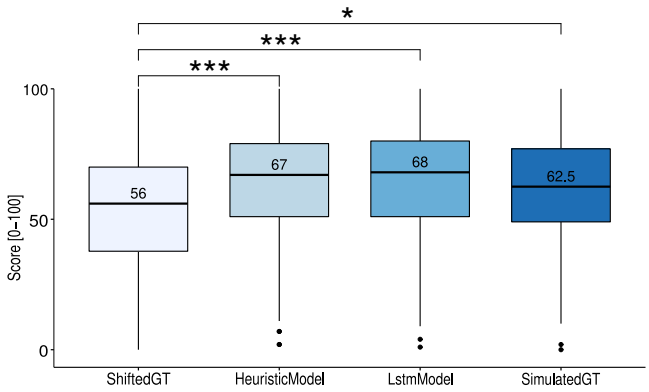

Fig. 3. Results of the online evaluation depending on the robot controls used. Each boxplot contains 600 points (number of subjects x number of clips). Significant p-values are indicated by * (<0.05), and *** (<0.001).

subjects was this: *"Rate the videos based on the relevance of the robot's behavior and gaze relative to the context"*. Before the experiment, the subjects are shown an explanation of the game and a picture of the scene being observed, so that they know what the targets of the robot's gaze are. At the end, they have to fill in a survey about their familiarity with the robots and their general feeling. We recruited 30 participants via the crowdsourcing platform Prolific[3], all native French speakers, with an equal representation of men and women.

4.2 Results

Figure 3 presents distributions of *rating_score* obtained by the 4 control policies. The significance of the results was tested by building a beta regression model, with *clips_Id* and *subject_Id* as random variables. A likelihood ratio test shows that the type of control significantly impacts the *rating_score* (chisq(3)=16.511, p=0.0008). Multiple pair-wise comparisons between the different controls resulted in the adjusted p-values presented in Figure 3. The only significant differences were found between *ShiftedGT* and the other three controls. This confirms that the gaze target management proposed by our model is viable. Despite a small supportive bias for *LstmModel*, no significant difference was found between the proposed control *LstmModel* and the control *HeuristicModel*. The difference between the head movements generated by a context-aware model and a non-context-aware model is not perceived. Despite the scores obtained by *SimulatedGT*, there is no significant difference between this hypothesized high anchor and the two controls, *HeuristicModel*, and *LstmModel*.

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

4.3 Discussion

The subjective evaluation revealed that despite a moderate objective performance, the proposed control was rated as appropriate as the original behavior in the data set. However, even though no significant differences were found, it is surprising that the trend in the score of the *SimulatedGT* control is lower than that of the two controls *HeuristicsModel* and *LstmModel*. Indeed, this control is supposed to reproduce a human behavior, more subtle than the two others. A first comment is that subjects do hear but not see the two players. Their estimation of the addressee(s) of human partners is degraded. This lack of context may create a misunderstanding of the robot's behavior, which mimics the behavior of a knowledgeable human who was participating in the interaction. Secondly, although third-party evaluations have been shown to find similar results to those of internal participants in the interaction [25], it is possible that this evaluation method has limitations. This limitation would be especially valid when finely comparing controls. Finally, the subjective evaluation only focused on the question of appropriateness of the behavior, but other characteristics could have been interesting to evaluate, such as naturalness or engagement.

5 Conclusion

In this study, we introduced a data-driven gaze control for HRI multi-party interaction, where the robot is an animator of a collaborative game. The control is based on two cascaded LSTM networks trained on multimodal data, one for gaze target prediction, one for head movement generation. Using an online perception study, we showed that this control is viable, with attention target prediction comparable to human behaviour, but no advantage of context awareness was revealed for head movement generation. These promising modeling results need to be further developed. Future work will aim to identify the reasons for this non-perception of differences. Two approaches are envisaged: improving the models by using embedding or CNN layers, for example, but also conducting a new perception evaluation with raters facing the physical robot for checking the engagement hypothesis.

Acknowledgements.

The RoboTrio corpus was supported by CNRS through a S2IH PEPS funding. This research is supported by the ANR 19-P3IA-0003 MIAI. The first author is financed by a CIFRE PhD granted by ANRT (2021/0836).

References

1. Admoni, H., Scassellati, B.: Social eye gaze in human-robot interaction: A review. Journal of Human-Robot Interaction Steering Committee **6**(1), 25–63 (2017)

2. 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 (2013)
3. Aliasghari, P., Taheri, A., Meghdari, A.F., Maghsoodi, E.: Implementing a gaze control system on a social robot in multi-person interactions. SN Applied Sciences **2**, 1–13 (2020)
4. Birdwhistell, R.L.: Background to kinesics. ETC: A Review of General Semantics **13**(1), 10–18 (1955)
5. 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)
6. Correia, F., Campos, J., Melo, F., Paiva, A.: Robotic gaze responsiveness in multiparty teamwork. International Journal of Social Robotics **15** (2022)
7. Freedman, E., Sparks, D.: Coordination of the eyes and head: Movement kinematics. Experimental brain research **131**, 22–32 (2000)
8. Fuller, J.H.: Comparison of Head Movement Strategies among Mammals. In: The Head-Neck Sensory Motor System. Oxford University Press (1992)
9. Fuller, J.H.: Head movement propensity. Experimental Brain Research **92**, 152–164 (2004)
10. 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, USA (2021)
11. Haefflinger, L., Elisei, F., Gerber, S., Bouchot, B., Vigne, J.P., Bailly, G.: On the benefit of independent control of head and eye movements of a social robot for multiparty human-robot interaction. In: Kurosu, M., Hashizume, A. (eds.) Human-Computer Interaction. pp. 450–466. Springer Nature Switzerland, Cham (2023)
12. Huang, H.H., Kimura, S., Kuwabara, K., Nishida, T.: Generation of head movements of a robot using multimodal features of peer participants in group discussion conversation. Multimodal Technologies and Interaction **4**(2), 15 (2020)
13. 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)
14. Itti, L., Dhavale, N., Pighin, F.: Photorealistic attention-based gaze animation. In: 2006 IEEE International Conference on Multimedia and Expo. pp. 521–524 (2006)
15. 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)
16. Kendon, A.: Some functions of gaze-direction in social interaction. Acta psychologica **26 1**, 22–63 (1967)
17. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2014)
18. Kuno, Y., Sadazuka, K., Kawashima, M., Yamazaki, K., Yamazaki, A., Kuzuoka, H.: Museum guide robot based on sociological interaction analysis. p. 1191–1194. CHI '07, Association for Computing Machinery, New York, NY, USA (2007)
19. 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 (2010)

20. Mishra, C., Skantze, G.: Knowing where to look: A planning-based architecture to automate the gaze behavior of social robots. In: 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). p. 1201–1208. IEEE Press (2022)
21. Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., Ishiguro, H.: Conversational gaze mechanisms for humanlike robots. ACM Trans. Interact. Intell. Syst. **1**(2) (2012)
22. 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, USA (2009)
23. Nakano, Y.I., Yoshino, T., Yatsushiro, M., Takase, Y.: Generating robot gaze on the basis of participation roles and dominance estimation in multiparty interaction. ACM Trans. Interact. Intell. Syst. **5**(4) (2015)
24. Nguyen, D.C., Bailly, G., Elisei, F.: Comparing cascaded LSTM architectures for generating head motion from speech in task-oriented dialogs. In: Kurosu, M. (ed.) Human-Computer Interaction. Interaction Technologies. pp. 164–175. Springer International Publishing, Cham (2018)
25. 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)
26. Prévot, L., Elisei, F., Bailly, G.: The robotrio corpus (2020), https://hdl.handle.net/11403/robotrio/v1, ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr
27. Sacks, H., Schegloff, E., Jefferson, G.: A simple systematic for the organisation of turn taking in conversation. Language **50**, 696–735 (1974)
28. 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. p. 332–336. HAI '21, Association for Computing Machinery, New York, USA (2021)
29. Sidner, C.L., Kidd, C.D., Lee, C., Lesh, N.: Where to look: A study of human-robot engagement. In: Proceedings of the 9th International Conference on Intelligent User Interfaces. p. 78–84. IUI '04, Association for Computing Machinery, New York, USA (2004)
30. 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)
31. Stefanov, K., Salvi, G., Kontogiorgos, D., Kjellström, H., Beskow, J.: Modeling of human visual attention in multiparty open-world dialogues. J. Hum.-Robot Interact. **8**(2) (2019)
32. 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)
33. Zangemeister, W., Stark, L.: Types of gaze movement: Variable interactions of eye and head movements. Experimental Neurology **77 3**, 563–577 (1982)
34. 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 (2014)