Investigating Conversational Dynamics in Human-Robot Interaction with fMRI

Ekaterina Torubarova (ekator@kth.se)
Division of Speech, Music and Hearing, KTH Royal Institute of Technology
Stockholm, Sweden

Caroline Arvidsson (caroline.arvidsson@ling.su.se)
Department of Linguistics, Stockholm University

Julia Uddén (julia.udden@psychology.su.se)
Department of Psychology, Department of Linguistics, Stockholm University

André Pereira (atap@kth.se)
Division of Speech, Music and Hearing, KTH Royal Institute of Technology

Abstract

We investigated how verbal communication with a robot differs from talking to a human in terms of brain activity by analysing an open-source fMRI dataset. We focused on modeling conversational dynamics rather than conversation as a whole, by analysing fine-grained events, in particular turn initiation. The results indicate that turn initiation in a conversation with a human involves higher activation in auditory and visual cortex than turn initiation with a robot. Conversely, listening to the robot showed higher engagement of auditory cortex than listening to a human. We suggest that verbal and non-verbal turn-taking cues provided by the human agent engage more cognitive processing for picking up the turn. On the other hand, listening to a robot agent requires more processing than listening to a human. Both findings suggest that the accurate simulation of appropriate turn-taking cues and behaviors will help robots to establish more natural conversation dynamics and that the use of brain imaging can provide valuable objective measurements for assessing user states in human-robot interaction.

Keywords: Cognitive Neuroscience; Robotics; Intelligent agents; Language and thought; Pragmatics; fMRI

Introduction

Social robotics is a quickly developing field for creating embodied intelligent agents able to communicate with the user. Compared to typically disembodied smart voice assistants such as Siri or Alexa, embodied robots are expected to play more complex social roles to be used in education, healthcare, elderly care, etc., where they can engage the user for a longer time (Leite, Martinho, & Paiva, 2013). When perceiving the user or producing a response, they should do it in a human-like manner, exploiting multimodal behavioral cues such as gaze and head movements, facial expressions, gestures, prosody, etc. (Duric et al., 2002). As humans, we automatically use these cues in communication. They help us understand not only each other's literal verbal message but also the underlying intention, affect, level of engagement and other pragmatic features (Streeck, Goodwin, & LeBaron, 2011).

Empirical studies on human-robot interaction (HRI) often evaluate the robot's performance by using subjective user reports (e.g. (De Graaf & Allouch, 2013; Corrigan et al., 2015; Hall et al., 2014)). This method has certain limitations: importantly, it does not allow for studying the dynamics of the interaction, since the reports are usually done post-hoc, when participants evaluate their experience as a whole. Many HRI studies use behavioral data for evaluating user state. Communicative behavioral cues are associated with some inner cognitive processes (e.g. attention, enjoyment, etc.): count of head nods, patterns of eye movements, change in body orientation, among others. (e.g. (Ishihara, Nitta, Nagasawa, & Okada, 2018; Ma et al., 2019; Ben-Youssef, Varni, Essid, & Clavel, 2019)). Usually, the behavior is annotated by one or several annotators, or more recently, computer vision algorithms are used to extract behavioral features, which is widely used for building models of autonomous user state detection based on behavior (Leo, Medioni, Trivedi, Kanade, & Farinella, 2017). The limitation of this method is that the mapping between the behavioral cues and the presumable inner state experienced by the user can be arbitrary and may depend on the annotator's judgment. In addition, the personal inner state is not always expressed in observable behavior (Song & Nakayama, 2009), which can be due to many factors such as individual differences or variations in conversational settings. To account for the limitations of subjective reports and behavioral annotations there is a need for online non-interrupting profound evaluation of user experience in HRI. Applying cognitive neuroscience methods in HRI can help overcome these limitations and provide a deeper insight into the perception of the robots by users.

The first studies applying fMRI to communication with social robots date back to 2008 (Krach et al., 2008; Hegel, Krach, Kircher, Wrede, & Sagerer, 2008). With the advances in both social robotics and brain imaging methodology, more works appear in the recent years (for review see (Cross, Hortensius, & Wykowska, 2019)). The focus of many fMRI studies in HRI is on social cognition, i.e. whether a social robot can involve brain mechanisms associated with social cognition similarly to a human. Various social attributes of human-robot interaction were investigated using fMRI, such as empathy (Rosenthal-Von Der Pütten et al., 2014; Cross, Riddoch, et al., 2019), emotion perception (Chaminade et al., 2010), mentalizing (Chaminade et al., 2012; Özdem et al., 2017), action observation (Gazzola, Rizzolatti, Wicker, &

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Keysers, 2007; Miura et al., 2010). It was shown that interacting with or observing social robots can elicit brain activity similar to humans under certain conditions.

Less work has been done on the differences in verbal communication with a robot and a human. One factor restricting this research has been high susceptibility of brain imaging signal to noise caused by movement artifacts during speaking. A novel within-subject paradigm was presented in (Rauchbauer et al., 2019), studying free bidirectional conversation of participants with a human and a robot agent. Here, the robot was used as a control condition for studying social cognition: unlike the human agent, it provided minimum social behavioral cues such as facial expressions or head movements, while the speech content was similar. The authors found that talking to both agents yielded activation in speech perception and production areas. However, only interaction with a human agent involved activation in Theory of Mind areas such as the temporoparietal junction and hypothalamus. The authors published the open-source multimodal dataset used in this study (Rauchbauer et al., 2020). The analysis of this dataset was further deepened in (Spatola & Chaminade, 2021). Here, the authors analysed how brain activity changes over time across several interactions depending on the agent. The human condition but not the robot condition involved an increase in activation in posterior cingulate cortex. This area, among other functions, has been linked to social cognition. The authors assumed that social engagement is updated along the interaction according to the interlocutor's feedback. In another study (Hogenhuis & Hortensius, 2021), the authors approached the dataset from another perspective, i.e. which elements of the conversation contribute to the differences in agent perception. While the original work (Rauchbauer et al., 2019) contrastet the whole 1-minute conversations, here the authors divided the interactions into speech production and comprehension episodes. During production, only subtle differences in activity were found for different agents. Comprehension showed significantly different activation profiles depending on the agent. The authors found differences in activation in the language network as well as the visual person perception network, but not the theory-of-mind network. Specifically, listening to a robot contrasted with listening to a human revealed more activation in the bilateral Heschl's gyrus, lateral occipital cortex, insular cortex, and inferior frontal gyrus. Thus, the authors concluded that the differences in agent comprehension occur at lower perceptual level rather than higher-order cognitive level.

The previous studies demonstrate that focusing on different conversational events may reveal additional information about differences in processing human and robot agents. In this paper, we aimed to further deepen the analysis by focusing on the processes involved in turn initiation during the conversation. We focused on turn initiation as one of the key components of turn-taking in a dialogue. The turn-taking process is crucial for establishing common ground and understanding the interlocutor's intention, which provides a more engaging interaction. By focusing on turn initiation, we can investigate how the nature of the agent affects the way users plan their responses in a conversation. The action of initiating a turn during conversations likely involves numerous complex operations. Prior to taking the floor, the upcoming speaker is presumably required to pay attention to implicit turn-ending cues (e.g., prosodic, syntactic, gestural (Schaffer, 1983; Beňuš, Gravano, & Hirschberg, 2011; Stivers et al., 2009)), while simultaneously planning their own turn, and signal the intention to initiate speech in order to generate smooth turn transitions (Levinson & Torreira, 2015). Coordinating speaker change may thus involve pragmatic processes over and beyond core language, possibly also implicating areas related to cognitive control and social cognition.

In this paper, we are going to report an analysis performed on an openly available fMRI dataset of human-human and human-robot conversations provided by (Rauchbauer et al., 2020). The aim of our study was to deepen the analysis of this dataset by focusing on language processing in the brain during more fine-grained conversational events, in particular turn initiation. In this paper, we went for whole-brain analysis given that turn initiation has not been extensively studied before with fMRI and a large set of outcomes was possible. Similarly, for comprehension and production, the only study comparing these events in human-human vs human-robot interaction is (Hogenhuis & Hortensius, 2021). Since we used the same dataset, defining a set of regions of interest from this study would be so-called double dipping (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009), which would result in invalid statistics.

To our knowledge, the differences in brain activity when picking up the turn after a human or a robot agent have not been studied before. We assume that studying turn initiation may provide insights into the inner user experience during conversations with a robot compared to a human. We hypothesize that how a human agent and a robot deliver their turn may result in differences in brain activity during the participant's turn initiation.

Method

Dataset

The openly available dataset by (Rauchbauer et al., 2020) consists of the recordings of 25 native French participants. Each participant had four sessions of six one-minute conversations, alternating between a human agent and a robot agent. The robot used in the experiment was a humanoid robotic head Furhat (Al Moubayed, Beskow, Skantze, & Granström, 2012) whose utterances were pre-written and controlled by an operator using a Wizard-of-Oz procedure. The participants were supposed to discuss the images they saw on a screen with an agent who was outside of the scanning room and connected via online video stream and bidirectional audio. In total, 24 minutes of conversation for each participant were recorded: 12 with a human and 12 with a robot. During the conversations, the brain activity of the participants and the

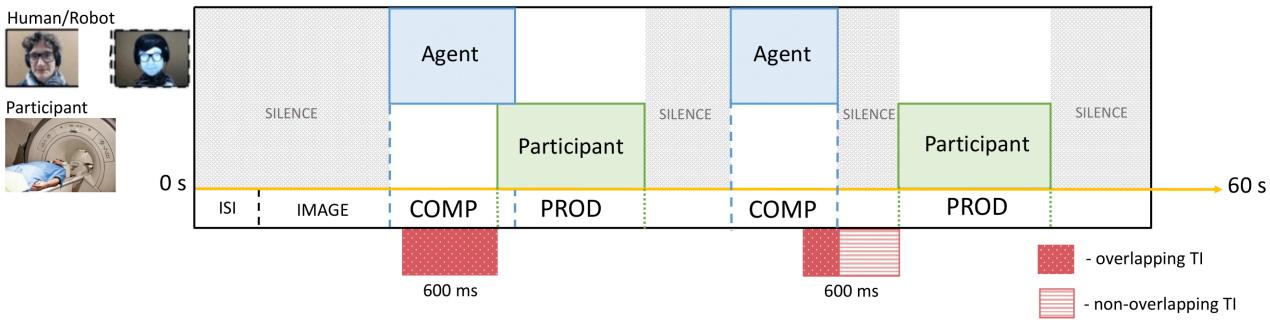


Figure 1: An example of a 60s conversation and the corresponding events in the fMRI model. Each conversation started with a fixation cross (event *ISI*) and the presentation of an image (event *Image*). Light blue boxes correspond to the utterance of the agent (human or robot), the event *Comprehension*; light green boxes – utterance of the participant, the event *Production*; grey areas – silence of both participant and the agent, the event *Silence*; red boxes - 600 ms turn initiation time window before the onset of production, divided into the events *Overlapping Turn Initiation* (dark red dotted segment) and *Non-Overlapping Turn Initiation* (light red segment) in respect to overlap with comprehension.

audio of the conversation were recorded, along with other parameters which are out of the scope of the current analysis. For the details about data acquisition, see (Rauchbauer et al., 2019).

Preprocessing

The preprocessing and the analysis of fMRI data were carried out using SPM12[1]. Out of 25 participants, two were excluded from the following analysis: one due to excessive head movement and another due to incomplete data. Preprocessing included rigid body transformation (realignment) with 6 parameters (translations and rotations); independent head movement check in x, y, z (1 participant excluded at this step); functional images coregistration to one of the anatomical images (T1) and normalization to a standard Montreal Neurological Institute (MNI) space with affine regularization and resampling of the voxels to 2x2x2 mm with a 4th degree B-spline interpolation. The normalization step also included white and grey matter segmentation and bias correction. Finally, functional images were spatially smoothed using a 3D isotropic 5 mm full-with-at-half-maximum Gaussian kernel. A temporal high-pass filter (cycle cut-off at 128 sec) was utilized to account for various low-frequency effects.

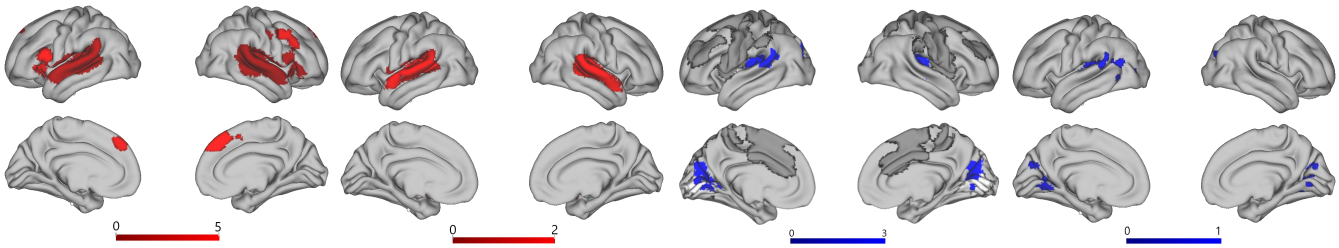
First Level Analysis

We created several first-level models by using the transcriptions of the participant's and the agent's speech[2]. The onsets and durations of the utterances were extracted from the transcriptions using a Python script. The hemodynamic response function (HRF) in each model was modeled for the events as regressors and six motion parameters using a general linear model. The regressors were convolved with a canonical HRF using a 2 mm within brain mask.

Model 1: Original Events Durations We created a model with 7 event classes: *Fixation Cross* at the beginning of the conversation, *Image Presentation, Comprehension, Silence, Turn Initiation, Production* (Fig. 1). These events, apart from fixation cross and image presentation, were modeled separately for each agent. With this model, we first aimed at replicating the results in (Hogenhuis & Hortensius, 2021) in comprehension and production. Then, we aimed at investigating turn initiation. Our model included silence and turn initiations, while the model in (Hogenhuis & Hortensius, 2021) included only audible speech and laughter segments. We defined turn initiation as a 600 ms time window before the onset of the participant's utterance, based on picture naming literature suggesting 600 ms to be a minimum picture name articulation latency (Indefrey & Levelt, 2004). While there is evidence that turn planning is also distributed across comprehension, (Bögels, Casillas, & Levinson, 2018) suggest a trade-off between comprehension and turn planning. The chosen time window in our study is unique as it contains processes related to initiating the turn based on the previous planning, while it is relatively absent of concurrent comprehension processes.

Model 2: Fixed Events Durations As mentioned above, Model 1 was used to replicate the results in (Hogenhuis & Hortensius, 2021) regarding comprehension and production. The analysis of the conversation transcriptions showed that on average, the utterances of the human agent (M=2.09s, SD=1.66s) were longer than the robot agent (M=1.27s, SD=0.74s), $t=22.9$, $p<0.05$. This difference in event duration could explain the previous results since this difference was not accounted for in the previous study. We thus created another first-level model with a fixed duration for comprehension and production events for both agents of 600 ms. This model was used to further investigate a potential caveat regarding the results in (Hogenhuis & Hortensius, 2021). We used this model only to compare its results with the results of the original model for comprehension and production.

[1] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/
[2] Available at https://www.ortolang.fr/market/corpora/convers/v2



(a) Comp_r > Comp_h (Original Model)
(b) Comp_r > Comp_h (Fixed Duration Model)
(c) TI_h > TI_r (Original Model)
(d) TI_h_n/o > TI_r_n/o (No overlap TI Model)

Figure 2: Increased activations during comprehension of the robot compared to the human (a,b) and turn initiation with the human compared to the robot (c,d) in different first-level models. In (c), turn initiation with a robot compared to baseline is shown in dark grey. Only significant clusters are shown (p FWE-corr < 0.05).

Model 3: Overlapping/Non-Overlapping Turn Initiation The analysis of the transcriptions showed that the turn initiation window defined in Model 1 was more likely to overlap with comprehension for the human agent comparing to the robot, because the segments of silence were on average shorter with the human agent. To account for this difference, we created another first-level model, in which turn initiation was divided into *Overlapping* and *Non-Overlapping* segments for both agents. Only events longer than 300 ms were included in the model, which resulted in a procedure where we classified the dominant type (either the overlapping or non-overlapping segment) of each turn initiation. This model was used to further investigate and confirm the results of turn initiations with Model 1, circumventing the issue of overlaps with comprehension.

Second Level Analysis

For the second level analysis, the cluster-forming threshold of *p-uncorr.* was set to .005 (no extent-level threshold, $k = 0$). Family-wise error (FWE), as implemented in SPM12, was utilized as multiple comparison correction method (at the cluster and peak-level). The significant clusters were projected on mid-thickness surface using Connectome Workbench [3]. Anatomical labelling was done using the ROI MNI V7 anatomical parcellation database in AAL3 toolbox in SPM12 (Rolls, Huang, Lin, Feng, & Joliot, 2020), with region inclusion margin set as minimum 10% label.

Results

Model 1: Original Events Durations

In an attempt to replicate the results of (Hogenhuis & Hortensius, 2021), we compared comprehension and production events of the original model depending on the agent. Directly contrasting production segments did not reveal significant differences depending on the agent. No significant activation was found higher in comprehension of the human compared to the robot. Comprehension of the robot elicited higher activation than comprehension of the human agent in bilateral Heschl's gyrus, bilateral superior temporal gyrus, bilateral middle temporal gyrus, and bilateral inferior frontal gyrus, among other areas (Figure 2a). See all significant clusters in Table 1.

In order to analyse turn initiation, we first contrasted it with the baseline (fixation cross) to see which regions in general are involved during turn initiation. During turn initiation, widely distributed activity was found for both agents. Directly contrasting turn initiation segments between a robot and a human conversation did not reveal any higher activation for the robot. The reverse contrast, turn initiation with a human compared to the robot, showed higher activation in bilateral temporal superior gyrus, bilateral Heschl's gyrus, bilateral calcarine sulcus, bilateral cuneus, left superior occipital gyrus, left lingual gyrus, left Rolandic operculum (Figure 2c, Table 1).

Model 2: Fixed Events Durations

Given that comprehension and production events differed in durations in the original model, we repeated the same comprehension and production contrasts in the fixed durations model. Similarly to the original model, in the fixed 600 ms window model no significant differences were found neither for production events, nor for comprehending the human agent compared to the robot agent. Comprehension of the robot agent, however, showed higher activation in bilateral Heschl's gyrus, bilateral superior temporal gyrus, among other areas (Figure 2b, Table 1).

Model 3: Overlapping/Non-overlapping Turn Initiation

In the original model, turn initiation segments were more likely to overlap with comprehension in the human-human conversation. To account for this potential caveat, in Model 3 we contrasted overlapping and non-overlapping turn initiation segments separately depending on the agent. During the overlapping period of turn initiation, no clusters were found more active neither for the human nor for the robot agent. For the non-overlapping segments of turn initiation, a significant increase in activation was found only for the human agent. In

[3] https://www.humanconnectome.org/software/connectome-workbench

particular, compared to the robot agent, there was higher activation in the left temporal superior gyrus, left Heschl's gyrus, left lingual gyrus, as well as right calcarine sulcus and right cuneus (Figure 2d, Table 1).

Discussion

We investigated the differences in brain activity during turn initiation in a free conversation with a human and a robot agent. We used an open source fMRI dataset (Rauchbauer et al., 2020), which was previously analysed at a whole-conversation level and at the level of speaking and listening events. We aimed at further deepening the analysis of this dataset by further defining conversational events: in our models, we included production, comprehension, silence and turn initiation. We created several first-level models that account for the differences between the human and the robot condition in the dataset, such as the length of the agents' utterances and transitional gaps duration.

The results of our models showed that the involvement of the auditory cortex in human vs robot conversation differed radically during comprehension compared to turn initiation. As the participants were listening to what the agent was saying, auditory areas were increasingly involved when listening to the robot compared to the human. On contrary, when participants were preparing to speak, more activation was found for the human vs the robot condition in the auditory and visual cortex. These results indicate different involvement of low-level perceptual processes (for auditory and visual input) depending on the agent. More relative processing is needed for auditory comprehension of the robot, whereas more relative processing is required for taking in the human agent, both audibly and visually, specifically during the critical moment when the participant is initiating a turn. Importantly, we did not find significant differences in areas associated with high-level processes such as Theory of Mind. Thus, we assume that the differences in the processing of the agent were caused by the perceptual features of the agents' behavior rather than processing of their perceived intrinsic nature. It should be kept in mind that in this dataset, the two agents had behavioral differences. The robot was used as a control condition to study social interaction and was lacking facial expressions, prosodic cues and its utterances were pre-written and produced using a Wizard-of-Oz procedure that created larger unnatural gaps between turns. Therefore, the processing of the robot's utterance could, as we have sketched, involve more extensive neural activity, given that the participants could not rely on natural behavioral cues that indicate turn-taking. On contrary, during turn initiation, even during silence of both the participant and the agent, more processing was required for the human agent. We assume that the process of turn initiation involves monitoring the interlocutor's behavior and speech in order to plan the production of one's own utterance, which was less expressed in the robot.

By analysing comprehension and production, we aimed at replicating the results of the previous study (Hogenhuis & Hortensius, 2021). Silence and turn-initiations were included in our models alongside with the speech events. Unlike in (Hogenhuis & Hortensius, 2021), we did not find a significant increase in activation for comprehending a human over a robot. For comprehension of the robot versus human, we primarily replicate the results showing increased activity in bilateral Heschl's gyrus. Our analysis, however, revealed other involved areas not found in (Hogenhuis & Hortensius, 2021). Given that the utterances of the human and robot agent differed in average duration, we deepened our analysis using a fixed durations model, which showed similar results: the only significant difference was found in comprehension of robot versus human agent. Here, the activation was again found in bilateral Heschl's gyrus. Thus, even after redefining the conversational events and considering the utterance duration differences, we partially replicated the results of the previous study. Hence, we find increased activation in bilateral Heschl's gyrus to be the most robust finding comparing comprehension between robot and human.

Our result can be associated with the novelty effect of listening to the robot. Given that the gaps for taking over the turn after the participant were longer for the robot, which slows down the natural conversational dynamics, higher activation in auditory processing of the robot can be associated with the difference in expectancy of the agent's response. More analysis is needed on how the activity in the auditory cortex depending on the agent changes during comprehension over the course of interaction, similarly to the analysis in (Spatola & Chaminade, 2021) which focused on changes in social cognition over time.

The neural correlates of turn initiation in human-robot interaction have not been studied before. The focus of the current study was on the differences during these events between the human and robot agent. We saw that in general, turn initiation, no matter the agent, involves a distributed activity including areas of auditory, visual, somatosensory and prefrontal cortex. Our results suggest that turn initiation with a human agent, presumably, requires more visual and auditory processing. Interestingly, after taking into account a potential caveat - the overlap with comprehension - we found significant differences only in the non-overlapping segments of turn initiation. Thus, this difference arises mainly when then the agent has already finished the turn. This can indicate that before producing their own turn, the participants kept processing the human agent's previous utterance and current non-verbal behavior.

Similarly to (Hogenhuis & Hortensius, 2021), our models did not show strongly significant differences during production depending on the agent. Based on that finding, we assume that aligning one's turn with the agent's turn was more affected by the agent's behavior than the process of delivering the utterance. However, it is interesting to investigate how the conversational setting, e.g. the topic of discussion, might affect this result. For instance, the task of the current dataset required discussing an image without the need to come up

Anatomical region	MNI Local maxima			Cluster		Voxel	
	x	y	z	Size	p-FWE	t value	p-FWE
1) comp_r >comp_h (Orginal Model)							
Rheschl/RSTG/RROL/RINS/RMTG/RTPsup/RIFGop	50	-22	8	7199	<.001	7.98	.001
Lheschl/LSTG/LROL/LTPsup/LSMG	-58	-20	8	5390	<.001	7.84	.002
LIFGop/LIFGorb/LIFGtr/LINS	-48	16	12	1034	.024	n.s	
RMFG/LMFG	4	42	42	1063	.021	n.s	
RIFGop	50	6	40	1065	.021	n.s	
2) comp_r >comp_h (Fixed Duration Model)							
Rheschl/RSTG/RTPsup/RINS/RINFop/RMTG	46	-30	10	4019	<.001	10.37	<.001
Lheschl/LSTG/LINFop/LMTG/LTPsup/LSMG	-40	-32	10	4171	<.001	8.84	<.001
3) TI_h >TI_r (Original Model)							
Rheschl/RSTG	42	-28	14	912	.036	6.06	.045
Lheschl/LSTG/LIFGop	-36	-34	18	1589	.003	n.s	
RCUN/LCAL/RCAL/LCUN/LSOG/LLING	-10	-80	24	2241	<.001	n.s	
4) TI_h_n/o >TI_r_n/o (No overlap TI model)							
Lheschl/LSTG/LLING/RCAL/RCUN	-34	-62	6	3352	<.001	n.s.	

Table 1: Significant activations in selected contrasts (the other contrasts rendered not significant). For each cluster, MNI coordinates of the cluster's local maxima are given together with cluster size. Abbreviations (L/R prefixes for left and right hemisphere, n.s - not significant): heschl - Heschl's gyrus, STG - superior temporal gyrus, ROL - rolandic operculum, INS - insula, TPsup - superior temporal pole, INFop/orb/tr - inferior frontal gyrus pars opercolaris/orbitalis/triangularis, SMG - supramarginal gyrus, MFG - middle frontal gyrus, MTG - middle temporal gyrus, CUN - cuneus, CAL - calcarine sulcus, SOG - superior occipital gyrus, LING - lingual gyrus.

with a common decision with the agent or requiring a feedback from the agent. Therefore, during production the participants might have not been paying significant attention to the agent's reaction.

The results of this study showed that verbal interaction with a robot via audio-video connection can significantly differ from interaction with a human in terms of brain activity. We saw that taking into account conversational dynamics, the differences depending on the agent arise on perceptual rather than high-level cognitive levels. In the current data, the verbal and non-verbal behavior of the human and robot agents had significant differences, which may explain the obtained results. Our results suggest that while talking to the human agent, the participants were more involved in picking up the turn. Increased attention and natural fluid turn-taking are associated with higher conversational engagement (Yu, 2015), which in turn results in a more positive interaction experience. We assume that studying turn initiation can become a window to studying the users' inner experience during interaction with social robots. Thus, we assume that the design of social robots' turn-taking mechanism can benefit from emphasizing turn-ending behavioral and prosodic cues, which facilitates picking up the turn by the user.

In future research, we aim at implementing more naturalistic verbal and non-verbal cues in social robots with appropriate fluidity in turn-taking to investigate how it affects users' experience. To provide quick feedback to the user's utterance and mitigate the delays caused by manual operation of the robot's utterance, we are going to explore other methods of operating the robot, such as real-time teleoperation through virtual reality. In addition, we aim at exploiting other methods of brain data collection such as EEG. In the current study, we focused on fMRI since this method allows to spatially map the neural correlates of conversational dynamics, both on cortical and subcortical level. The results obtained in this study can help also researchers focus EEG studies on the identified regions of interest.

Conclusion

This paper examplifies that brain imaging can be a valuable method for studying human-robot interaction. We partially replicated the results of a previous study for comprehension episodes and contributed with additional findings regarding turn initiation that further help understand underlying differences between human-human and human-robot interactions. We found that listening to the robot required more relative activity than listening to the human; but importantly, preparing to speak after a human involved more activity than after a robot. We found significant differences on low-level perceptual rather than higher cognitive levels. We assume that these differences are explained by the robot's lack of natural turn-taking dynamics and non-verbal behaviors associated with higher conversational engagement. Improving these characteristics in a robot might close the gap in how our brain processes interactions with either robots or humans.

References

Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (2012). Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems* (pp. 114–130). Springer.

Beňuš, Š., Gravano, A., & Hirschberg, J. (2011). Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics*, *43*(12), 3001–3027.

Ben-Youssef, A., Varni, G., Essid, S., & Clavel, C. (2019). On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks. *International Journal of Social Robotics*, *11*(5), 815–828.

Bögels, S., Casillas, M., & Levinson, S. C. (2018). Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question. *Neuropsychologia*, *109*, 295–310.

Chaminade, T., Da Fonseca, D., Rosset, D., Lutcher, E., Cheng, G., & Deruelle, C. (2012). Fmri study of young adults with autism interacting with a humanoid robot. In *2012 ieee ro-man: The 21st ieee international symposium on robot and human interactive communication* (pp. 380–385).

Chaminade, T., Zecca, M., Blakemore, S.-J., Takanishi, A., Frith, C. D., Micera, S., ... Umiltà, M. A. (2010). Brain response to a humanoid robot in areas implicated in the perception of human emotional gestures. *PLoS one*, *5*(7), e11577.

Corrigan, L. J., Basedow, C., Küster, D., Kappas, A., Peters, C., & Castellano, G. (2015). Perception matters! engagement in task orientated social robotics. In *2015 24th ieee international symposium on robot and human interactive communication (ro-man)* (pp. 375–380).

Cross, E. S., Hortensius, R., & Wykowska, A. (2019). *From social brains to social robots: applying neurocognitive insights to human–robot interaction* (Vol. 374) (No. 1771). The Royal Society.

Cross, E. S., Riddoch, K. A., Pratts, J., Titone, S., Chaudhury, B., & Hortensius, R. (2019). A neurocognitive investigation of the impact of socializing with a robot on empathy for pain. *Philosophical Transactions of the Royal Society B*, *374*(1771), 20180034.

De Graaf, M. M., & Allouch, S. B. (2013). Exploring influencing variables for the acceptance of social robots. *Robotics and autonomous systems*, *61*(12), 1476–1486.

Duric, Z., Gray, W. D., Heishman, R., Li, F., Rosenfeld, A., Schoelles, M. J., ... Wechsler, H. (2002). Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proceedings of the IEEE*, *90*(7), 1272–1289.

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage*, *35*(4), 1674–1684.

Hall, J., Tritton, T., Rowe, A., Pipe, A., Melhuish, C., & Leonards, U. (2014). Perception of own and robot engagement in human–robot interactions and their dependence on robotics knowledge. *Robotics and Autonomous Systems*, *62*(3), 392–399.

Hegel, F., Krach, S., Kircher, T., Wrede, B., & Sagerer, G. (2008). Theory of mind (tom) on robots: A functional neuroimaging study. In *2008 3rd acm/ieee international conference on human-robot interaction (hri)* (pp. 335–342).

Hogenhuis, A., & Hortensius, R. (2021). Domain-specific and domain-general neural network engagement during human-robot interactions.

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1-2), 101–144.

Ishihara, T., Nitta, K., Nagasawa, F., & Okada, S. (2018). Estimating interviewee's willingness in multimodal human robot interview interaction. In *Proceedings of the 20th international conference on multimodal interaction: Adjunct* (pp. 1–6).

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? interaction and perspective taking with robots investigated via fmri. *PloS one*, *3*(7), e2597.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, *12*(5), 535–540.

Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: a survey. *International Journal of Social Robotics*, *5*, 291–308.

Leo, M., Medioni, G., Trivedi, M., Kanade, T., & Farinella, G. M. (2017). Computer vision for assistive technologies. *Computer Vision and Image Understanding*, *154*, 1–15.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6*, 731.

Ma, K., Wang, X., Yang, X., Zhang, M., Girard, J. M., & Morency, L.-P. (2019). Elderreact: a multimodal dataset for recognizing emotional response in aging adults. In *2019 international conference on multimodal interaction* (pp. 349–357).

Miura, N., Sugiura, M., Takahashi, M., Sassa, Y., Miyamoto, A., Sato, S., ... Kawashima, R. (2010). Effect of motion smoothness on brain activity while observing a dance: An fmri study using a humanoid robot. *Social neuroscience*, *5*(1), 40–58.

Özdem, C., Wiese, E., Wykowska, A., Müller, H., Brass, M., & Van Overwalle, F. (2017). Believing androids–fmri activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Social Neuroscience*, *12*(5), 582–593.

Rauchbauer, B., Hmamouche, Y., Bigi, B., Prevot, L., Ochs, M., & Thierry, C. (2020). Multimodal corpus of bidirectional conversation of human-human and human-robot interaction during fmri scanning. In *Proceedings of the 12th*

language resources and evaluation conference (pp. 661–668).

Rauchbauer, B., Nazarian, B., Bourhis, M., Ochs, M., Prévot, L., & Chaminade, T. (2019). Brain activity during reciprocal social interaction investigated using conversational robots as control condition. *Philosophical Transactions of the Royal Society B*, *374*(1771), 20180033.

Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J., & Joliot, M. (2020). Automated anatomical labelling atlas 3. *Neuroimage*, *206*, 116189.

Rosenthal-Von Der Pütten, A. M., Schulte, F. P., Eimler, S. C., Sobieraj, S., Hoffmann, L., Maderwald, S., ... Krämer, N. C. (2014). Investigations on empathy towards humans and robots using fmri. *Computers in Human Behavior*, *33*, 201–212.

Schaffer, D. (1983). The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics*, *11*(3), 243–257.

Song, J.-H., & Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. *Trends in cognitive sciences*, *13*(8), 360–366.

Spatola, N., & Chaminade, T. (2021). Precuneus brain response changes differently during human-robot and human-human dyadic social interaction.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. doi: https://doi.org/10.1073/pnas.0903616106

Streeck, J., Goodwin, C., & LeBaron, C. (2011). *Embodied interaction: Language and body in the material world*. Cambridge University Press.

Yu, Z. (2015). Attention and engagement aware multimodal conversational systems. In *Proceedings of the 2015 acm on international conference on multimodal interaction* (pp. 593–597).