Neuroscience to Investigate Social Mechanisms Involved in Human-Robot Interactions

Youssef Hmamouche, Magalie Ochs, Laurent Prévot, Thierry Chaminade

Neuroscience to Investigate Social Mechanisms Involved in Human-Robot Interactions

Youssef Hmamouche
youssef.hmamouche@lis-lab.fr
Aix Marseille Université, CNRS, LIS, UMR7020
Marseille, France

Magalie Ochs
magalie.ochs@lis-lab.fr
Aix Marseille Université, CNRS, LIS, UMR7020
Marseille, France

Laurent Prévot
laurent.prevot@univ-amu.fr
Aix Marseille Université, CNRS, LPL, UMR7309
Aix-en-Provence, France
Institut Universitaire de France
Paris, France

Thierry Chaminade
thierry.chaminade@univ-amu.fr
Aix-Marseille Université, CNRS, INT, UMR7289
Marseille, France
Corresponding author

ABSTRACT

To what extent do human-robot interactions (HRI) rely on social processes similar to human-human interactions (HHI)? To address this question objectively, we use a unique corpus. Brain activity and behaviors were recorded synchronously while participants were discussing with a human (confederate of the experimenter) or a robotic device (controlled by the confederate). Here, we focus on two main regions of interest (ROIs), that form the core of "the social brain", the right temporoparietal junction [rTPJ] and right medial prefrontal cortex [rMPFC]. An new analysis approach derived from multivariate time-series forecasting is used. A prediction score describes the ability to predict brain activity for each ROI, and results identify which behavioral features, built from raw recordings of the conversations, are used for this prediction. Results identify some differences between HHI and HRI in the behavioral features predicting activity in these ROIs of the social brain, that could explain significant differences in the level of activity.

KEYWORDS

human-machine interactions, conversation, functional MRI, multimodal signals processing, feature selection, prediction

ACM Reference Format:
Youssef Hmamouche, Magalie Ochs, Laurent Prévot, and Thierry Chaminade. 2020. Neuroscience to Investigate Social Mechanisms Involved in Human-Robot Interactions. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3395035.3425263

1 INTRODUCTION

Recently, "second-person neuroscience" was proposed as a new way for "the study of real-time social encounters in a truly interactive manner" [22], and the approach described in this paper addresses this endeavor. It is part of a larger project that previously involved recording of the data and preparing the corpus for sharing with the community. This unique corpus combines synchronized behavioral and neurophysiological recordings during natural human-human and human-robot natural interactions (HHI and HRI). The participant's brain activity was continuously scanned with functional magnetic resonance imaging (fMRI) while having natural conversations with a human (confederate of the experimenter) or a robotic head (controlled by the confederate, unbeknownst of the participant who believed the robot to be autonomous).

The methodological challenge is to understand relationships between complex behaviors and activity in the social brain during unconstrained interactions. It poses one major difficulty: what to use as explanatory variable? The approach used here was described previously in [5]. In a nutshell, behaviors are recorded and processed to build time-series, used in turn for the analysis of neurophysiological data. This approach allows the analysis to go further than the simple comparison between HHI and HRI already published for this corpus [3].

Finding pertinent relationships between behavioral time-series and neurophysiological time-series is another challenge for which a new approach is presented here, based on a branch of machine learning, namely time-series forecasting [13]. The method is used to identify which behavioral time-series are required to predict brain activity in specific brain regions, and their respective weights. We present a preliminary result obtained using this methodology focusing on two core regions of the social brain, in the the Medial PreFrontal Cortex (MPFC) and TemporoParietal Junction (TPJ). Given the increase of activity between HHI and HRI reported previously in these areas, our hypothesis is that different behaviors will be associated to the activity in these areas during these two conditions. Alternatively, the same behaviors associated with different levels of activity will suggest a top-down "switch" akin to the "intentional stance" for these areas [8].

2 DESCRIPTION OF THE CORPUS

2.1 Justification of the paradigm

The theoretical grounds underlying the choice of experimental paradigm, the procedures to acquire and prepare the corpus both for

ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-8002-7/20/10…$15.00
https://doi.org/10.1145/3395035.3425263

behavioral and neurophysiological data, and the bases of the machine learning machinery developed for its analysis have already been published. We will summarize the main points for this paper to be self-contained, but suggest to read the references provided in each section for more details. In order to investigate natural social interactions, it is essential that participants are unaware of the real purpose of the experiment. For this purpose, a cover story was developed responding to a complex set of specifications. The experiment is described as a neuromarketing experiment. A company wants to know if discussing with a fellow or an Artificial Intelligence about the images of a forthcoming advertising campaign is enough to guess the objective of the campaign [5]. The cover story is credible, there is a common, but loose, goal for the conversation, the conversations are truly bidirectional, and there is a legitimacy for talking with a robot.

Our objective isn't to test how an autonomous agent works, but how individuals' behavior changes depending on the nature of the agent. For this reason, participants believed the robot to be autonomous, but to avoid erratic responses during the recordings, a Wizard of Oz procedure was favored in which the confederate selects prerecorded responses expressed by the retroprojected conversational robotic head (Furhat robotics[1]) on a touchscreen.

2.2 FMRI experiment and data processing

Participants (n=25 in the full corpus) came to the MRI center and were presented with a fellow (a gender-matched confederate of the experimenter) and the robot. The cover story was presented and the participant installed in the scanner. The participant underwent four sessions of approximately 8 minutes of scanning comprising six experimental trials as follows: an image is presented, then a 1-minute discussion takes place, alternatively with the human and the robot. At the end of the recordings, we recorded 12 trials of 1-minute with the human and the robot for each participant.

fMRI data processing followed standard procedures and has been described in details [3]. fMRI data analysis first relied on the general linear model implemented in the toolbox Statistical Parametric Mapping [18], imported into the Conn toolbox [24] that automatizes the extraction of BOLD time series in regions of interest. A continuous time-series of 385 points covering the 8 minutes (repetition time: 1.205 seconds) were extracted for each Region of Interest (ROI), each session and each participant.

2.3 Regions of interest

For this paper, we investigated a question related to the theoretical framework of philosopher Daniel Dennett, namely the intentional stance [8]. It claims that when interacting with humans, we adopt a specific stance as we ascribe mental states to this rational agent. In contrast, we adopt a different stance when interacting with robots, despite the fact that they can appear and behave as humans. It is still debated in robotics today [25], and neuroimaging offers an unique opportunity to address this question as we measure objective neurophysiological markers of cognitive processes.

The intentional stance entails the ascription of mental states such as beliefs, desires, knowledge to others. It parallels a theoretical framework at the core of social cognitive neuroscience, namely Theory of Mind (ToM), that describes one's ability to ascribe hidden mental states, such as intentions, desires or beliefs, to oneself and to others. The neural bases of this ability has been intensively investigated with neuroimaging techniques, identifying two key areas in the cortex that are differentially activated when ascribing mental states versus physical states. In the first significant finding, participants in a Positron Emission Tomography (PET) scanner believed they played stone-paper-scissors with a human or a computer [10], and the only significant difference was an increase of activity in the medial prefrontal cortex (MPFC). The same region, as well as the right temporoparietal junction (TPJ), also had reduced activity when the partner was believed to be a robot endowed with with AI [6]. The very rude contrast HHI versus HRI we performed on this corpus [3] identified increased response in the right TPJ. The statistical threshold for the contrast HHI versus HRI was lowered ($p < 0.001$ uncorrected, unpublished data) to identify a cluster in the MPFC. We selected the two "social brain" ROIs from the Brainnetome parcellation of the human brain [9] that contained these rTPJ and rMPFC clusters.

There is a possible implication of the superior temporal gyrus (STS) in mentalizing, but this is controversial given the central role of this region in language (see for example [21]). We decided to include the left STS, given the left hemisphere dominance for language processing in right handed participants, to elucidate which aspects of behaviors are more pertinent for this region. Finally, we also included a control region in which the signal is not related to mental processing, a mask of brain white matter (WM).

3 METHODOLOGY

We implemented a machine learning process for the prediction of brain activity based on conversational behavioral data. A particularity of our approach compared to related work looking for features that cause the activation of a brain area ([4, 7, 15, 26]) is that it allows to quantify the predictive weights of the behavioral features. Therefore, it provides additional information about the association between behavioral features and the activity of a brain area.

The models of the proposed process are trained based on data of the corpus described in Section 2. The main steps are:

- Extracting behavioral features time-series from raw behaviors recorded in different modalities during conversations.
- Resampling and synchronizing behavioral features with respect to the BOLD signal frequency.
- Using feature selection method to select a subset of relevant features for each ROI.
- Predicting the discretized BOLD signal of each ROI based on the selected features.

3.1 Features extraction

The multimodal recordings of behaviour contains 3 types of raw data: video, speech (audios and transcriptions) and eye-tracking. Note that we don't have the video of the participants, as they were inside the fMRI scanner during the experiment. Therefore, we have speech recorded for both the participant and the conversant (human or robot), the videos of the conversant only, and eyetracking recordings of the participant only. The aim of feature extraction is to derive high-level behavioral features from this multimodal raw

[1] https://www.furhatrobotics.com; [1]

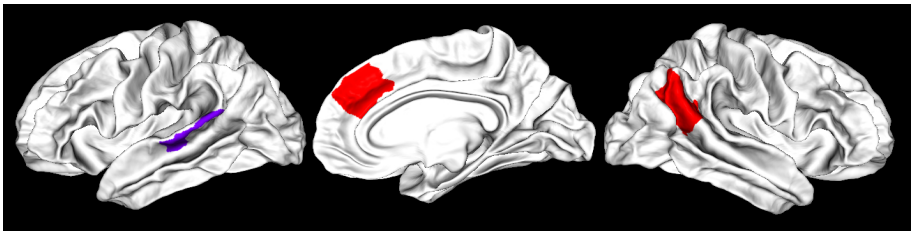

Figure 1: The three regions of interest investigated in this paper, from left to right. In purple, the left posterior STS region (lSTS). In red, the right medial prefrontal cortex (rMPFC) and right temporoparietal junction (rTPJ).

data, features that describe the behavioral events happening during the conversation. These features are presented in Supplementary Table, and in more details in [20]. For speech, we built time-series IPU (Inter-Pausal Units[2]) by IPU (*e.g.*, for lexical richness) or word by word (*e.g.*, for Feedback and Discourse Markers). For facial features, we used Openface software ([2]) to detect facial landmarks, action units), head pose and gaze coordinates from the video of the conversant. Eyetracking features were extracted directly from the output of Eyelink1000 system. From this data we calculated other features characterizing where the participant was looking (Face, Eyes, Mouth), with the use of the detected landmarks points of the conversant combined with the gaze coordinate of the participant. Features were re-sampled to have the same number of observations.

3.2 Feature selection

We use feature selection techniques before applying prediction models to improve the prediction by selecting the most relevant predictors, and at the same time to find the smallest set of features predicting the ROI. There are two categories of feature selection techniques. The first are Wrapper methods, *i.e.*, methods that use the prediction model recursively to eliminate features with low importance scores. Here, ***ModelSelect*** uses the classifier by executing first the model with all features, then selecting the top-k features based on their importance scores. The second category are filter methods, that select variables based on the relationships between the variables and independently from the prediction model used. We use two filter methods: **Ranking based on mutual information (*MI-rank*)** selects variables without the need of the classifier. It ranks variables based on their Shannon mutual information with the target variable. **K-medoid based method (*k-medoids*)** is an extension of the previous method. It first groups close predictive variables into clusters using the k-medoids algorithm (maximizing intra-class mutual information, and minimizing mutual information between classes), then selects one variable from each group based on the mutual information with the target variable. This method is designed to work with multimodal data, as grouping variables before ranking them partly solves the problem of dependencies between variables belonging to the same modality.

3.3 Prediction setup

Human-human and human-robot data are treated separately in order to compare them. For each agent, the obtained data consist of 13248 observations. The training data consist of 18 participants from 24, the test data are the observations of the remaining 6 participants (*i.e.*, 25% of all data). We applied the ADASYN algorithm on the training data [12] to address the problem of imbalanced data. The problem of imbalanced data occurs in our case because some ROIs are activated rarely during conversations, depending on the participants and their behaviors. This algorithm generates new observations by considering the distribution of the data. Then, we performed a 10-fold-cross-validation on training data to find the appropriate parameters of the classifier based on the F-score measure. Finally, we train the model with the selected parameters and evaluations are made on test data.

[2]An IPU is a speech block bounded by pauses and coming from a single speaker [16].

3.4 Prediction formulation

we first binarize the BOLD signal measured by fMRI as we want to predict whether a brain area is active or not as the original signal is continuous. The discretization method uses the mean of observations as a threshold after normalizing data of each participant.

The BOLD signal recorded as fMRI response to a behavioral event follows the Hemodynamic Response Function (HRF)[11]. This function peaks with a delay around five seconds that can vary somehow depending, for example, on the brain area. Other factors, such as saturation of the signal (see detailed examples in chapter 14 of [18]), make it difficult to consider a deconvolution of the BOLD signal. To handle this variability, our prediction model attempts to predict the discretized bold signal at time t based on sequence of observations of behavioral features that happened previously between $t-4s$ and $t-7s$. Therefore, lagged variables of each feature are considered. This is a temporal classification problem that can be expressed as follows:

$$Y(t) = f(X_1^{t-\tau_1:t-\tau_2}, \ldots, X_k^{t-\tau_1:t-\tau_2}) + U(t),$$

where Y is the discretized BOLD signal, $\{X_1(t), X_2(t), \ldots, X_k(t)\}$ are k behavioral variables. $X_i^{t-\tau_1:t-\tau_2}$ are the lagged variables of the i^{th} behavioral feature X_i, $\tau_1 = 7s$, $\tau_2 = 4s$, and $U(t)$ represents the error of the model.

We already used this methodology in previous works [13, 14], where we tested several classifiers (Logistic regression, SVM, Long Short Term Memory network) to predict brain areas involved in speech perception and production and face perception. We found that the Random Forest classifier outperforms the others. Here, we only use the Random Forest classifier, and we try to improve the predictions with feature selection.

	ROI	Feature selection	Codes of the selected features	Features importance	F-score
HHI	lSTS	*mi-rank*	[**1-c, 10-c**, 4-c, 5-c]	[**0.71, 0.17**, 0.1, 0.02]	0.71
	rTPJ	*k-medoids*	[**1-p, 1-c, 10-c, 34**, 29, 7-p]	[**0.33, 0.21, 0.17, 0.14**, 0.09, 0.07]	0.62
	rMPFC	*k-medoids*	[**1-c, 25**, 21, 26, 16, 1-p, 28, 19, 27, 21, 38, 9-c, 14]	[**0.19, 0.13**, 0.09, 0.09, 0.08, 0.07, 0.07, 0.05, 0.04, 0.03, 0.03, 0.02, 0.02]	0.60
HRI	lSTS	*mi-rank*	**[1-c]**	[**1.0**]	0.68
	rTPJ	*k-medoids*	[**1-c, 35**, 1-p, 16, 12-p,11-p, 9-p, 29]	[**0.62, 0.15**, 0.06, 0.05, 0.03, 0.02, 0.02, 0.02]	0.62
	rMPFC	*modelSelect*	[**37, 19, 21**, 26, 18, 25, 26]	[**0.22, 0.21, 0.12**, 0.06, 0.06, 0.06, 0.05]	0.62

Table 1: Results of prediction evaluations for Human-Human and Human-Robot Interactions (HHI and HRI *resp.*). For each region of interest (ROI) indicating the codes of the selected features (see Supplementary Table), and their importance scores (Bold: Importance > 0.1). Features code ending with "-p" are for the scanned participant, with "-c" are for the confederate.

4 RESULTS AND DISCUSSIONS

The results are presented in Table 1. It contains the best F-scores obtained, as well as the best feature selection method and the associated features and their importance. To reproduce the results, the input features and the code source are available online[3]. The results show that filter feature selection methods perform better than the wrapper method, which was selected once as best method. And as we expected, in one hand, the *mi-rank* method allows to have the best predictions for lSTS area in both HHI and HRI, as it involves one main modality. *K-medoids* works better for rTPJ and rMPFC, which involve features from multiple modalities.

Results indicate that the system is able to predict brain activity in lSTS, rTPJ and rMPFC higher than chance, while it is not possible for the white matter (F-scores not different from 0.5), as expected given the absence of task-related activity in this tissue. A second remark relates to the contrast between lSTS on the one hand and rTPJ and rMPFC on the other hand. The former was selected as being an area strongly devoted to language, while some speculate about its involvement in social cognition. The MI Rank was sufficient for significant prediction in both HHI and HRI, and speech activity of the confederate strongly dominates the list of features used for predicting, even being the only feature in the case of HRI. In the case of HHI, only one other confederate linguistic features represents than 10% of importance. Perceiving voices is enough to activate the STS [19]. Altogether, these results strongly comfort that the left posterior superior temporal sulcus region is devoted to speech perception. Concerning the "social cognition" areas, a number of preliminary remarks can be made. Importantly, increased activity for for the contrast HHI compared to HRI were found in the two ROIs [3]. which has been interpreted as the neural consequence of the adoption of an intentional stance in HHI, but not HRI. The feature selection method and the F-scores are similar for the two areas and the two interlocutors. Such similarity is in line with the hypothesis that these two areas have both similar and complex (6 to 13 features used for prediction) functions. Considering the rTPJ, speech activity (feature 1 for participant and confederate) is responsible for more than 50% of the predictive power for both HHI and HRI, while the other features (focusing on feature with importance > 0.1 indicated in bold in Table 1) are different between HHI and HRI. For HHI, we found one speech (Type-token ratio) and one facial (smiles) from the conversant, while the participant's saccades are used in HRI. Important features to predict the activity of the rMPFC differ between HHI and HRI. The conversant's speech and mouth movements represent 22% of the predictions for the former, while the participant's speed of gaze and the conversant's head rotations and action unit 2 together provide 55% of the prediction importance for the later. Altogether, there are differences between the features identified in HHI and HRI at this preliminary stage. For example, for both rTPJ and rMPFC, features used to predict HHI all relate to the conversant, emphasizing its importance in social cognition, while for HRI, the most important feature is from the participant's behavior (eye movements), which is not particularly pertinent for social cognition. Further work is needed to better compare the results for the two agents.

[3]https://github.com/Hmamouche/NeuroTSConvers

5 CONCLUSION

Here we have described an innovative methodology to investigate which behavioral features are required to explain the activity in brain areas devoted to social cognition, recorded while participants were having uncontrolled conversations with a human or a humanoid robot. We focused on four regions on the basis of simple assumptions: white matter should not be predictable, the left STS should be related to language, and regions of the social brain should be related to complex social signals. Preliminary results fit our hypotheses. Some differences were found between the behavioral features used to predict activity in social brain areas, not so much in the language perception area lSTS between HHI and HRI. We previously proposed that we adopt a different stance based on the level of activity of these areas. Further work is needed to interpret the differences found with the current analysis, still under development.

6 ACKNOWLEDGMENTS

This research is supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and AAP-ID-17-46-170301-11.1 by the Aix-Marseille Université Excellence Initiative (A*MIDEX).

REFERENCES

[1] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström. 2012. Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction. In *Cognitive Behavioural Systems (Lecture Notes in Computer Science)*, Anna et al. Esposito (Ed.). Springer Berlin Heidelberg, 114–130.

[2] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. 59–66. https://doi.org/10.1109/FG.2018.00019

[3] Birgit Rauchbauer, Bruno Nazarian, Morgane Bourhis, Magalie Ochs, Laurent Prévot, and Thierry Chaminade. 2019. Brain activity during reciprocal social interaction investigated using conversational robots as control condition. *Philosophical Transactions of the Royal Society B: Biological Sciences* 374, 1771 (April 2019), 20180033. https://doi.org/10.1098/rstb.2018.0033

[4] Daniel Bone, Chi-Chun Lee, and Shrikanth Narayanan. 2014. Robust Unsupervised Arousal Rating: A Rule-Based Framework with Knowledge-Inspired Vocal Features. *IEEE transactions on affective computing* 5, 2 (June 2014), 201–213. https://doi.org/10.1109/TAFFC.2014.2326393

[5] Thierry Chaminade. 2017. An experimental approach to study the physiology of natural social interactions. *Interaction Studies* 18, 2 (Dec. 2017), 254–275. https://doi.org/10.1075/is.18.2.06gry

[6] Thierry Chaminade, Delphine Rosset, David Da Fonseca, Bruno Nazarian, Ewald Lutcher, Gordon Cheng, and Christine Deruelle. 2012. How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience* 6 (2012). https://doi.org/10.3389/fnhum.2012.00103

[7] Hsuan-Yu Chen, Yu-Hsien Liao, Heng-Tai Jan, Li-Wei Kuo, and Chi-Chun Lee. 2016. A Gaussian mixture regression approach toward modeling the affective dynamics between acoustically-derived vocal arousal score (VC-AS) and internal brain fMRI bold signal response. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5775–5779. https://doi.org/10.1109/ICASSP.2016.7472784

[8] D. C. Dennett. 1987. *The intentional stance*. MIT Press, Cambridge, Mass.

[9] Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R. Laird, Peter T. Fox, Simon B. Eickhoff, Chunshui Yu, and Tianzi Jiang. 2016. The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cerebral Cortex* 26, 8 (May 2016), 3508–3526. https://doi.org/10.1093/cercor/bhw157

[10] Helen L. Gallagher, Anthony I. Jack, Andreas Roepstorff, and Christopher D. Frith. 2002. Imaging the Intentional Stance in a Competitive Game. *NeuroImage* 16, 3 (July 2002), 814–821. https://doi.org/10.1006/nimg.2002.1117

[11] C Gössl, L Fahrmeir, and DP Auer. 2001. Bayesian modeling of the hemodynamic response function in BOLD fMRI. *NeuroImage* 14, 1 (2001), 140–148.

[12] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 1322–1328.

[13] Youssef Hmamouche, Magalie Ochs, Laurent Prevot, and Thierry Chaminade. 2020. Exploring the Dependencies between Behavioral and Neuro-physiological Time-series Extracted from Conversations between Humans and Artificial Agents. In *9th International Conference on Pattern Recognition Applications and Methods*. SCITEPRESS - Science and Technology Publications, Valletta, Malta, 353–360. https://doi.org/10.5220/0008989503530360

[14] Youssef Hmamouche, Laurent Prevot, Magalie Ochs, and Chaminade Thierry. 2020. BrainPredict: a Tool for Predicting and Visualising Local Brain Activity. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 703–709.

[15] André Knops, Bertrand Thirion, Edward M. Hubbard, Vincent Michel, and Stanislas Dehaene. 2009. Recruitment of an Area Involved in Eye Movements During Mental Arithmetic. *Science* 324, 5934 (June 2009), 1583–1585. https://doi.org/10.1126/science.1171599

[16] Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Twelfth Annual Conference of the International Speech Communication Association*.

[17] Magalie Ochs, Sameer Jain, and Philippe Blache. 2018. Toward an automatic prediction of the sense of presence in virtual reality environment. In *Proceedings of the 6th International Conference on Human-Agent Interaction*. ACM, 161–166.

[18] William D. Penny, Karl J. Friston, John T. Ashburner, Stefan J. Kiebel, and Thomas E. Nichols. 2011. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier.

[19] Cyril R. Pernet, Phil McAleer, Marianne Latinus, Krzysztof J. Gorgolewski, Ian Charest, Patricia E.G. Bestelmeyer, Rebecca H. Watson, David Fleming, Frances Crabbe, Mitchell Valdes-Sosa, and Pascal Belin. 2015. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage* 119 (Oct. 2015), 164–174. https://doi.org/10.1016/j.neuroimage.2015.06.050

[20] Birgit Rauchbauer, Youssef Hmamouche, Bigi Brigitte, Laurent Prévot, Magalie Ochs, and Thierry Chaminade. 2020. Multimodal corpus of bidirectional conversation of human-human and human-robot interaction during fMRI scanning. In *Proceedings of the twelfth international conference on Language Resources and Evaluation, LREC 2020*. European Language Resources Association (ELRA).

[21] Elizabeth Redcay. 2008. The superior temporal sulcus performs a common function for social and speech perception: Implications for the emergence of autism. *Neuroscience & Biobehavioral Reviews* 32, 1 (Jan. 2008), 123–142. https://doi.org/10.1016/j.neubiorev.2007.06.004

[22] Leonhard Schilbach, Bert Timmermans, Vasudevi Reddy, Alan Costall, Gary Bente, Tobias Schlicht, and Kai Vogeley. 2013. Toward a second-person neuroscience. *Behavioral and Brain Sciences* 36, 4 (July 2013), 393–414. https://doi.org/10.1017/s0140525x12000660

[23] Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research* 13, Jun (2012), 2063–2067.

[24] Susan Whitfield-Gabrieli and Alfonso Nieto-Castanon. 2012. Conn: A Functional Connectivity Toolbox for Correlated and Anticorrelated Brain Networks. *Brain Connectivity* 2, 3 (June 2012), 125–141. https://doi.org/10.1089/brain.2012.0073

[25] Agnieszka Wykowska, Thierry Chaminade, and Gordon Cheng. 2016. Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, 1693 (May 2016), 20150375. https://doi.org/10.1098/rstb.2015.0375

[26] Tal Yarkoni, Deanna M. Barch, Jeremy R. Gray, Thomas E. Conturo, and Todd S. Braver. 2009. BOLD Correlates of Trial-by-Trial Reaction Time Variability in Gray and White Matter: A Multi-Study fMRI Analysis. *PLOS ONE* 4, 1 (Jan. 2009), e4257. https://doi.org/10.1371/journal.pone.0004257

Modality	Features	Code	Description	Details
Linguistic features	Speech-activity	1	Is the interlocutor speaking?	Based on time-aligned IPU transcript.
	Overlap	2	Both interlocutors speaking?	idem.
	Laughter	3	Laughter occurrences	Based on word-level time-aligned transcripts
	Filled pauses	4	Filled_pauses occurrences	Based on word-level time-aligned transcripts : *"euh", "heu", "hum", "mh"*
	Feedback	5	Conversational Feedback occurrences	Based on word-level time-aligned transcripts : *'oui' (yes), 'ouais' (yeah, 'non'(no), 'ah', 'd'accord'(right), 'ok'* + Laughters
	Discourse Markers	6	Occurrence of words used to keep speech organized	Based on word-level time-aligned transcripts : *'alors'(so), 'mais'(but), 'donc'(therefore), et('and'), 'puis'(then), 'enfin'(finally), 'parceque'(because), 'ensuite'(after)*
	Spoken Particles	7	Occurrence of (final) spoken particle items	Based on word-level time-aligned transcripts : 'quoi', 'hein','ben','bon'(well), mais (but), 'bref' (in short)]
	Interpersonal	8	Merge of inter-personal linguistic features	Merge of (Filled-pauses, Feedback, Discourse Markers, Spoken Particles and Laughter)
	Turn Latency	9	Time to take the turn	Based on time-aligned IPU transcript.
	Type-token ratio	10	Type-token ratio	Based on time-aligned transcript: (number of different tokens) / (total number of tokens).
	Lexical-richness	11	Lexical richness measure	Based on time-aligned transcript: (number of adjectives + number of adverbs) / (total number of tokens) [17].
	Polarity and Subjectivity	12-13	Sentiment analysis metrics	Based on time-aligned transcript [23].
Facial features	gaze-angle-x, gaze-angle-y	14, 15	Gaze angle coordinates	Based on conversant video frame by frame.
	pose-T(x, y,z), pose-R(x, y,z)	16,17	Head rotation and translation estimation	-
	Head-translation-energy	18	Kinetic energy of head translation	-
	Head-rotation-energy	19	Kinetic energy of head rotation	-
	AU01_r, AU02_r, AU06_r, AU26_r	20-24	Facial Action Units involved in smiles, surprise and speech production *resp.*	-
	Mouth-AU	25	Facial movements related to mouth.	Sum (AU10_r, AU12_r, AU14_r, AU15_r, AU17_r, AU20_r, AU23_r, AU25_r, AU26_r)
	Eyes-AU	26	Facial movements related to eyes.	Sum (AU01_r, AU02_r, AU04_r, AU05_r, AU06_r, AU07_r, AU09_r)
	Total-AU	27	Global representation of all facial movements.	Sum of all action units.
	Emotions	28-33	('Happiness', 'Sadness','Surprise', 'Fear', 'Anger', 'Disgust')	Probabilities detected from conversant video frame by frame.
	Smiles	34	Smile's probability estimation.	idem.
Eyetracking features	Saccades	35	Occurence of Saccades	Based on gaze coordinates of the participant, recorded using the Eyelink1000 system.
	Vx, Vy	36, 37	Speed of the gaze coordinates.	-
	Face	38	Occurrences of looks on the face.	-
	Eye	39	Occurrences of looks on the eye.	-
	Mouth	40	Occurrences of looks on the mouth.	-

Supplementary table: behavioral features extracted.