Interpretable Prediction of Brain Activity during Natural Social Interactions using Multimodal Behavioral Signals

Youssef Hmamouche, Magalie Ochs, Laurent Prevot, Chaminade Thierry

Interpretable Prediction of Brain Activity during Natural Social Interactions using Multimodal Behavioral Signals

Youssef Hmamouche[1], Magalie Ochs[2], Laurent Prévot[3], Thierry Chaminade[4*],

1 International Artificial Intelligence Center of Morocco, University Mohammed VI Polytechnique, Ben Guerir, Marocco
2 LIS UMR 7020, CNRS, Aix Marseille Université, Université de Toulon, Marseille, France
3 LPL UMR 7309, CNRS, Aix Marseille Université, Marseille, France
4 INT UMR 7289, CNRS, Aix Marseille Université, Marseille, France

* Corresponding author: thierry.chaminade@univ-amu.fr

Abstract

We present an analytical framework aimed at predicting the local brain activity of individuals during a conversation with another human or a humanoid robot based on multimodal recordings of their behavior. In this framework, we first extract high-level features from the raw behavioral recordings of both interlocutors. Then, classifiers are applied to predict binarized brain activity from these features using a dynamic prediction model. Here, we focus on brain regions involved in social interactions, both speech processing involved areas in conversations and information integration areas, in order to validate our framework. This framework not only predicts local brain activity significantly better than random, but it also identifies the behavioral features required for this prediction depending on the brain area under investigation and on the nature of the conversational partner. In the left Superior Temporal Sulcus, perceived speech is the most important behavioral feature for predicting brain activity, regardless of the agent, while multiple features, which differ between the human and robot interlocutors, contribute to prediction in regions involved in social signal integration, such as the TemporoParietal Junction. This framework allows us to study how multiple behavioral signals from different modalities are integrated in individual brain regions during inherently complex unconstrained natural social interactions.

Introduction

Investigating the causes of brain activity during natural social interactions is a difficult problem given that multiple cognitive processes are at play in such complex behavior. Meanwhile, brain activity follows non-linear dynamics potentially influenced by factors that are difficult to measure, such as internal thoughts and other psychological factors. However, it is important to evaluate which behavioral event, and more importantly combinations of such events, significantly influence the activity in local brain areas to better understand brain-behavior relationships in natural interactions.

In this article, we tackle the problem of finding dependencies between neurophysiological time series in Regions Of Interest (or ROIs, which are intermediate functional units between neurons and the whole brain, corresponding to a patch of cortical surface measuring several mm^3). The analysis framework can be described with three main

steps. We (*i*) use a non-invasive neuroimaging technique (here, fMRI) to record human brain activity during a natural social interaction as well as synchronized behavior of the interacting agents across multiple modalities, such as audio or video; then (*ii*) high-level features are extracted from the raw recordings captured during the interaction using knowledge from cognitive neuroscience that informs us about the behaviors that can be extracted; finally (*iii*) machine learning methods are used to find relationships between the recorded fMRI signal and the behavioral features. To validate the analysis framework proposed, we focus in this paper on conversation as the natural social interaction, on brain areas involved in speech perception on the one hand, and on areas known to respond to several modalities of behaviors related to social cognition in controlled experimental settings on the other hand.

Testing causal relationships between time series can be performed using causality tests. Those tests are generally used to test the causality of one or multiple variables on a target variable. Many of those tests are based on prediction models, such as the Granger causality test and its alternatives [1]. The problem here is that we have a large set of multimodal predictive variables, and we don't know which subset of variables we have to include in our models given a large number of possibilities. The challenge is then, first, to find the subset of behavioral features that has an impact on a given neurophysiological time series, and then to test their prediction. Therefore, we propose to first build a dynamic prediction model with a feature selection and identify the most relevant features in terms of their prediction scores. An advantage of this approach is that the model can be used not only to detect relationships between variables but also to make predictions.

The majority of existing approaches test *a priori* hypotheses between brain activity and behavior [2–5], and rely on multivariate regressions to handle the problem of brain activity prediction based on behavior. Here, we propose to use advanced machine learning algorithms to predict the neurophysiological response based on recorded behaviors without relying on strong *a priori* hypotheses, allowing us to identify new relationships between specific aspects of complex behaviors on the one hand, and neurophysiology on the other hand. In practice, we build a framework that first extracts several multimodal features from raw behavioral signals and then automatically finds the smallest set of features that have a significant impact on the prediction of localized brain activity recorded with fMRI.

For (*i*), we use an existing corpus of natural conversations recorded during fMRI experiments with 24 participants, providing synchronized neurophysiological and behavioral signals [6]. This corpus is unique in the sense that participants' behavior is unconstrained and therefore different from classical fMRI datasets generally acquired in highly controlled conditions. This corpus includes human-human and human-robot conversations, which allows the analysis of the variability of relations between local activity and behavior not only in terms of the brain area under investigation, but also of the social context of the interaction operationalized by the nature on the conversational agent, human or artificial [7]. With regards to (*ii*), a number of recent publications explain how high-level behavioral features have been extracted from raw recordings, either conversational features extracted from raw audio recordings after transcription and annotation [8] or visual features extracted from the video recordings of the interlocutor [9]. Other features used here rely on the recording of the eye movements of the participant and an exhaustive list of all raw recordings can be found in [10]. Here we mainly focus on the machine learning aspect of the framework (*iii*) that has been implemented for predicting fMRI brain activity based on multimodal raw signals of human-human or human-robot conversation. It brings two major contributions:

1. A *methodology for predicting discretized fMRI responses from behavioral signals*

of bidirectional conversation is presented in sections Analysis Framework and Implementation.

2. The *identification of dependencies between behavior and brain activity in specific brain areas* is presented in section Results and discussed in Discussion.

To summarize our contributions, the problem presented in this paper is specific in the sense that the data are unique and have never been used in related work for a prediction task similar to ours. The prediction methodology is also different since we are dealing with time series with the same frequency as the fMRI signals, and we have not found any related work that is similar to our research in terms of prediction approach.

Related work

Prediction using multiple modalities is a challenging task due to the diversity of the available signals to process. Classical prediction models are more simple in the sense that the input features belong to the same modality, and thus have similar structures. Concerning multimodal prediction, many problems arise, such as how to synchronize signals of different types and frequencies, how to represent data from each modality, and the fusion methodology to use for including all signals into one prediction model.

Multimodal approaches

There are several real applications that involve multimodal signals to predict a given feature. Multimodal data are very useful in emotion recognition. In [11], a system is presented to predict depression using audio, visual, and linguistic features. Similarly, in [12], the goal was to classify emotional states based on multimodal signals including audio, video, and physiological sensor signals. In both papers, the approaches used are based on extracting multiple features from each modality, then fusing them in one model that performs a classification task based on the extracted features. This approach seems very logical since it enables explaining the results from the variables extracted so as to make interpretation possible. On the other hand, the approach that consists in building a one-step prediction model using the raw data as input may be efficient in terms of prediction accuracy but lacks interpretability. Multimodal data are also very common in the field of human social interaction. In [13], a multimodal approach is proposed to predict back-channel feedback related to bidirectional human-human interaction. These back-channel represent signs indicating the continuity of the interaction. The used model is probabilistic, and it is based on Hidden Markov Model or Conditional Random Fields. The model takes as input three types of features, the prosody, the spoken words, and the eye gaze coordinates. In [14], the SEMAINE multimodal corpus is provided in the context of human social interaction. It consists of several bidirectional conversations with 20 participants containing both visual and speech recordings. This corpus was recorded to provide researchers with a multimodal data with two main aims: first, to analyse the interactions from a social cognition point of view; and second, to teach machines to interact with humans. For more details about multimodal approaches, an important survey about multimodal approaches has been presented in [15]. It contains a general discussion of multimodal strategies for learning prediction models from information from different sources. The authors discuss the main and general steps for multimodal machine learning and present several real applications involving multimodal data. The authors describe a set of fundamental steps for building a machine learning multimodal system, where the most important step is data representation, which is related to extracting and summarizing useful features from the different modalities. Other steps are also discussed about coordinating between the modalities, aligning and fusing them in order to make

predictions. Those steps are very important, but each application will require specific processing depending on the underlying modalities and the predictive variables.

Multimodal approaches for brain activity prediction

Regarding the issue of predicting brain activity based on behavior, several approaches have been proposed in the literature. In [16], the authors investigate the effect of adding visual information to auditory speech signals on the activity of auditory cortex areas. The results show a significant increase in the activation of the studied regions of interest (ROIs) based on ANOVA analysis. In [2], the fMRI neural activation associated with the semantic is predicted based on a large text dataset. The brain regions studied are in the sensory-motor cortex. The model used consists of transforming the text into semantic features, and then building a regression model that expresses the fMRI brain activity as a linear combination of semantic features. The authors show a prediction accuracy of 0.62 or higher, but on each participant independently. This issue has also been addressed with a multi-subject approach, by concatenating data of multiple participants. For example, in [5], the goal was to predict voxel activity from cortical areas, measured via the BOLD (Blood Oxygen Level Dependent) signal based on the speech signal. The data used has been collected from an fMRI experiment on 7 subjects. The methodology adopted is based, first, on constructing semantic features from natural language, then, a dimension reduction using PCA (Principal Component Analysis) is applied to reduce the number of predictive variables, and a model is learned based on multiple linear regression with regularization in order to predict the BOLD signal. Finally, the obtained prediction results and the principal components of the predictive variables are both combined to classify brain areas according to the semantic features categories. Other types of behavioral signals have been investigated by evaluating the effect of a single feature on brain activity. For example, the speech reaction time has been used to predict activity in specific brain regions [3]. In [4], the acoustically-derived vocal arousal score [17] is used to predict the BOLD signal using the Gaussian mixture regression model. In [18], the authors predict the BOLD signal in the posterior parietal cortex based on eye movement data using a multivariate regression model. More general approaches have been tried to predict the brain activity of various areas using different types of signals at the same time. For example, in [19], correlations are analyzed using linear regression between the BOLD signal and behavioral features computed from observed facial expressions, speech reaction time, and eye-tracking data.

Discussion

In the related works presented above, dependencies between behavior and specific functional brain areas have been investigated. However, only one or a few modalities have been taken into account. In addition, the methods used are generally based on correlation analysis or multiple regression. However, finding relevant predictive features using feature selection techniques with machine learning methods, such as prediction models based on artificial neural networks, can be particularly relevant to this research question. In this article, we propose a framework that consists in extracting high-level features from raw multimodal behavioral data consisting of audios, videos and eye-tracking recordings, then applying feature selection and prediction with different classifiers to predict discretized neuro-physiological signals in circumscribed brain regions from multimodal behavioral signals.

.

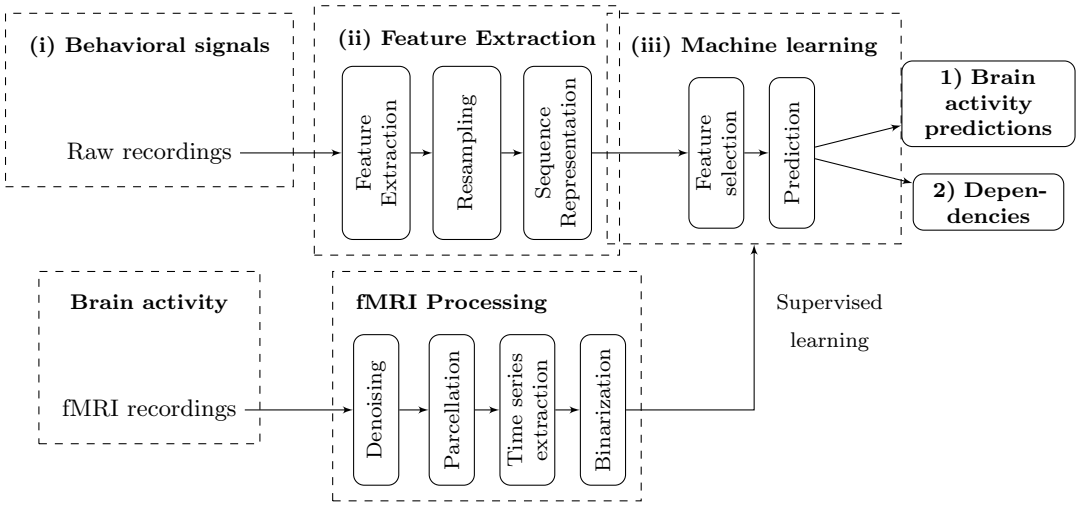

Figure 1. An illustration of the analysis framework. It is composed of three main steps. The first one (*i*) includes input signal acquisition and processing. The second step (*ii*) is for feature extraction and resampling. In this step, high-level and interpretable features are extracted from input signals, and since the signals are multi-modal, they are resampled and sequenced. The final step (*iii*) consists in training/testing machine learning models to predict brain activity time series in order to find dependencies between the extracted features and the target variables.

Analysis Framework

Overview

The analysis framework presented is rooted in a meta-model using machine learning tools which allows to 1) predict brain activity from multimodal behavioral signals recorded in complex natural social interactions, and 2) to identify dependencies between brain activity and the high-level features extracted from the raw signals. The proposed framework is illustrated in Fig 1. In this section, we present an overview of its main steps. The methodology presented may be used with other input signals and other types of brain activity, such as EEG and MEG, but would require additional assumptions and processes, taking into account for example the non-independence of successive data points.

The output of the framework consists of the predictions of regional brain activity and the features selected for the prediction, and in particular the importance score of the selected behavioral features. The importance of the features is generally missing in most of the existing works that merely identify features that trigger the activation of a brain area [3, 4, 17, 18]. An additional output is also when the regional activity can't be predicted, given the brain region of interest, the experimental condition scrutinized and the features used for the prediction.

The first step is to extract features from raw input signals for each modality separately in order to construct time series that will be used as predictive variables. But they require resampling in order to obtain time series with the same number of observations as the input signals have different recording frequencies. Finally, the predictive variables also need to be restructured as sequences, since we are trying to predict the next brain activity based on the past values of the behavioral features. Then, feature selection methods and prediction models are applied to predict the discretized BOLD signals, and also to find the smallest subset of features leading to the best possible predictions for each brain area. Thus, the system does not require *a priori* hypotheses about

the relationship between brain activity and behavior. Instead, it allows the finding of new causal relationships between conversational behavior and brain activity, which can be interpreted from a neuroscience and social cognition point of view in order to identify complex multimodal cognitive mechanisms. Our approach consists in using a new prediction model based on a specific temporal function to predict the discretized brain activity time series based on the behavioral features.

Feature extraction (*ii*)

Integrating multimodal signals within the same deep learning architecture is a delicate task compared to classical unimodal architectures. With multimodal signals, different networks must coordinate to have efficient and robust predictions. Our idea is to transform all the input modalities into interpretable sequences, *i.e.*, features that describe behaviors that can be described verbally, such as a binary time series describing whether one person is speaking or not. Such a strategy allows us to have a unified classification network that not only works for all modalities together but can also work with any single one.

The aim of feature extraction in our case is to compute high-level features from raw recordings as the latter are not easy to introduce in machine learning models and not easily interpretable.

The importance of these features is demonstrated by the fact that they are more effective to explain brain activity in higher-order integrative brain areas, that are of major interest when investigating complex social interactions. While raw recordings that consist of audios and videos are likely to affect in priority primary brain areas in the auditory and visual cortex.

A similar approach was used for the emotions extracted from voice signals, showing a shift from early feed-forward processing of stimulus categories to later processing of the salience of the stimuli [20]. This step, based on domain knowledge, provides the input behavioral features that are known to affect brain activity and are amenable to improvements as knowledge itself progresses, providing new features that can be extracted from raw signals, used in the prediction model and allowing to evaluate their predictive power. Iteratively, such an approach could help refine the types of representations used in the brain. As a result, an obvious limitation, but also a strength, of the approach, is that the current results only provide a snapshot constrained by the raw signals used and the features extracted in the current implementation, but that can later be compared when new features become available and added to this extraction step.

Machine learning (*iii*)

Feature selection

Feature selection is performed on the variables representing the temporal sequences of the extracted recorded features. The goal here is to identify the most relevant set of variables in terms of the prediction accuracy for each variable that will be predicted. In the following, we detail three different methods that we have used for feature selection.

Wrapper feature selection method

This method uses the prediction model itself to perform the selection of the appropriate variables. First, the prediction model is executed with all features. Then, the top-k features are selected based on their weights provided by the model. This method is simple in terms of implementation, but its main drawback is that the selected features are specific to the prediction model used, which means that if we change the model, the selected features can also change.

Ranking based on mutual information

The filter method selects variables without the need of a prediction model. It ranks variables based on their Shannon mutual information (MI) with the target variable. The Shannon MI works only for discrete variables, for example, the bivariate MI between two variables X, Y can be expressed as follows:

$$I(X;Y) = \sum_{x,y} P(x,y)\, log\left(\frac{P(x,y)}{P(x)\,P(y)}\right) \tag{1}$$

In our case, we can't use formulation (1) directly because some features are continuous. Two solutions are possible to overcome this problematic, the first and most common one consists in discretizing the continuous features, and then to use the Shannon MI. The second one is to use a continuous estimation of the Shannon MI, as proposed in [21]. The authors present an estimation of MI based on k-nearest neighbors approach, which relies an idea similar to the one for estimating the continuous entropy [22]. This MI estimator can be expressed as follows:

$$\hat{I}(X;Y) = \Gamma(k) + \Gamma(n) - \frac{1}{n}\sum_{i=1}^{n}(\Gamma(n_x(i)+1) + \Gamma(n_y(i)+1)), \tag{2}$$

where Γ is the Gamma function, n is the number of observations of X and Y, k is the number of neighbors to consider, $n_x(i)$ is the number of points where the distance from x_i is strictly less than $d_i/2$ where d_i is the distance from X_i to its k^{th} neighbor.

Clustering based method

This method is another filter-based feature selection method. It corresponds to an extension of the previous method in the sense that it considers the mutual information between predictive variables in addition to their mutual information with the target variable. It first groups close predictive variables into clusters using the *k-medoids* algorithm based on the principle of maximizing mutual information within clusters and minimizing mutual information between clusters. Then, it selects one variable from each cluster based on the mutual information with the target variable. This method is designed to work with multimodal data, since the fact of grouping variables before ranking them seems a relative solution to the problem of dependencies between variables that belong to the same group.

Predictions

Our approach is based on a multimodal fusion. This step creates a shared representation of the features irrespective of their modality. The sequenced behavioral features obtained previously are fed into deep network classifiers and then fully connected layers. The classifiers used to belong to two types. The first one includes the Random Forest (RF), Support vector Machine (SVM), and Logistic Regression (LReg).. The second type includes models based on neural networks including a fully connected network and an LSTM (Long Short Term Memory) network. For both networks, the backpropagation (through time for LSTM) is applied to train the network over 50 iterations using the stochastic gradient descent (SGD) algorithm. The multimodal architecture with the LSTM network is illustrated in Figure 4.

To measure the robustness of our predictions, we added a baseline classifier that generates random predictions (named Rand in the rest of this article). We trained this classifier after tuning a parameter representing the strategy of prediction generation. Three strategies are considered: A stratified way by generating predictions regarding the distribution of the training data, a uniform way by generating predictions uniformly, and

the third way is based on the most frequent class. The strategy that provides the best prediction results in the training stage is the one that is kept for testing the baseline classifier. The classifications metrics used are the F-score and the Recall. Recall is the percentage of examples classified as positive, among the total number of positive examples, while Precision is the percentage of true positive examples among the examples classified as positive. The F-score is more balanced since it considers both false positives and false negatives. Note that Precision can be inferred from the results since the F-score is the weighted average of Precision and Recall.

Finally, Student's *t*-tests are performed to test the equality (null hypothesis) of the average F-scores between the best and the baseline classifier obtained in the training step. Student's *t*-tests are the most recommended and used statistical tests to compare machine learning models [23].

Implementation

Here, we describe how the general framework described in the previous section was used to analyse a specific dataset. More precisely, the proposed analysis framework is used to 1) predict brain activity from multimodal signals of bidirectional human-human and human-robot conversations, and 2) to identify dependencies between brain activity and high-level features extracted from raw behavioral signals. The dataset is fully relevant to the issues the analysis framework is supposed to address, namely, physiological and multi-modal behavioral recordings performed during complex behaviors, between which we want to identify dependencies. The complex behaviors under scrutiny are unconstrained conversations between participants and a fellow human and a robot. Inputs of the framework are, on the one hand, brain activity measured in fMRI via the BOLD signal, and on the other hand, behavioral signals of bidirectional conversations between a participant whose brain activity is scanned inside the fMRI machine and an interlocutor, either a human or a robot, located outside the fMRI machine (here, three types of conversational signals: speech, video, and eyetracking).

Corpus (i)

The data used in this work were collected with fMRI recordings described in previous work [6], and are available at [24]. It consists of four sessions for 24 participants, each containing six conversations of 60 seconds, three with a human and three with a conversational robot in alternating order. An *advertising campaign* provides a cover story to make sure that the participants are unaware that the actual focus of the experiment is to record a corpus of natural social interactions: participants are told that they should guess what is the message carried by images in which fruits appear either as *superheroes* or *rotten*. Each conversation between the participant and either a confederate of the experimenter or a FURHAT conversational robot [25] (controlled by the confederate in a *Wizard-of-Oz* mode, unknown to the participant), is about one single image of the purported *advertising campaign*. The project received ethical approval from the Comité de Protection des Personnes (CPP) Sud-Marseille 1 (approval number 2016-A01008-43). Written consent was obtained from all participants. The input raw data consist of 3 type of signals: video of the interlocutor (human or robot), speech (raw audio recordings and manual transcriptions) of both the participant and the interlocutor, and eye-tracking recordings of the participant. Note that the videos of the participants are not recorded, as they were inside the scanner during the fMRI experiment.

fMRI data preparation

The fMRI data requires processing. First, whole brain recording is processed to remove as much of the noise it contains as possible. It is then parcellated in regions of interest to summarize the activity of functionally homogeneous areas. The mean data is extracted in each region of interest, single trials extracted from the continuous recordings of the session, and finally binarized. In addition, the BOLD response lags from the behavioral sources with a delay modelled by the hemodynamic response function that needs to be taken into account.

Processing fMRI signals

Standard fMRI acquisition procedures were used, described in detail in [6]. BOLD signal 3-dimensional images are recorded in the whole brain every 1.205 second (repetition time). Standard SPM12 preprocessing procedures are used [26], including correction for time delays in slice acquisition ("slice timing"), image realignment, magnetic field inhomogeneities correction, normalization to the standard MNI space using the DARTEL [27] procedure for coregistration of individual participants' anatomy, and finally spatial smoothing with a 5-mm full-width half-maximum 3-dimensional Gaussian kernel. Extraction of the BOLD signal in regions of interest is performed using the conn toolbox [28], and includes several denoising procedures, firstly a linear detrending using a high-pass filter with a threshold of 128 seconds, secondly using realignment parameters to calculate nuisance regressors related to participants' movement during scanning, thirdly taking heartbeat and breathing recordings to remove physiological artefacts with the PhysIO toolbox [29], and finally extracting BOLD signal in the white matter and cerebrospinal fluid and using the 5 first eigenvariates of the time series as nuisance representing signal fluctuations in non-cortical brain tissues.

Brain Regions Of Interest (ROIs)

A 275-area parcellation based on functional and anatomical connectivity patterns ([30], `https://atlas.brainnetome.org/bnatlas.html`) defines ROIs for the whole brain. Continuous time series (385 time points) are extracted for each ROI and each session and participant represents the mean activity within the ROI after denoising. In this paper, we focus on specific regions (Fig 2 and Table 1) involved in social cognition (TemporoParietal Junction, Precuneus, ventral and dorsal medial PreFrontal Cortex) including speech perception (in the caudal superior temporal region) and emotional processing (Amygdala). Two different types of control ROIs were also included, one corresponding to the white matter, where signal fluctuations are not supposed to reflect neuronal information processing and shouldn't be predictable, and another in the Primary Visual Cortex where predictions should rely on visual information instead of auditory or social information.

Binarization

Importantly, the raw BOLD is a continuous measure. A binarization of the signal into 0 (inactive) and 1 (active) is required as our approach consists in predicting whether a brain region is active or not. Such binarization of the BOLD signal has been used repeatedly when machine learning approaches are applied to fMRI signals, *e.g.* [31–35]. Here, we use a binarization method proposed by Ostu et al. [36] reproducing previous approaches based on filtering the BOLD signal of each brain area into two states (active, or non-active) based on its average.

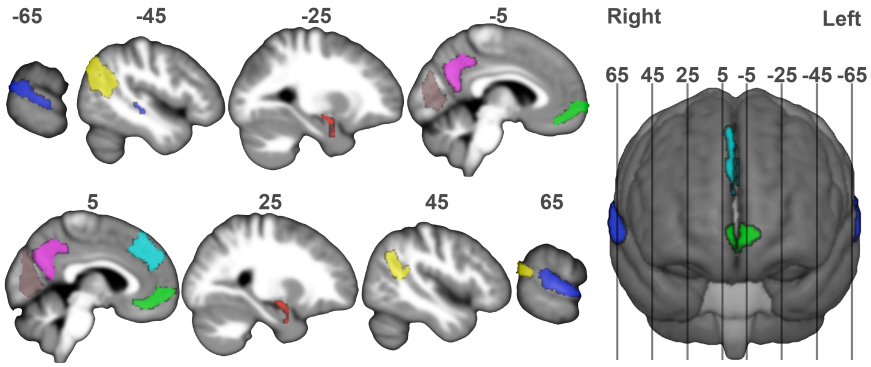

Figure 2. Regions of interest (ROIs) under investigation superimposed to sagittal sections of the average of the participants' brains normalized to the Montreal Neurological Institute (MNI) template space. Sections' location (in mm from the midbrain section, negative for the left hemisphere) are indicated by numbers above the sections and on the reference three-dimensional render seen from the front on the right panel. ROI order as in Table 2 for clarity: Primary Visual Cortex (V1): brown; Superior Temporal Sulcus (STS): Blue; TemporoParietal Junction (TPJ): Yellow; Precuneus (Pre): Pink; Amygdala (Amy): Red; VentroMedial PreFrontal Cortex (VMPFC): Green; DorsoMedial PreFrontal Cortex (DMPFC): cyan.

Table 1. The regions of interest (ROIs).

Abbreviations	Brain areas	Brainnetome atlas
l,r V1	left and right Primary Visual Cortex	191,192
l,r STS	left and right Superior Temporal Sulcus	75,76
l,r TPJ	left and right TemporoParietal Junction	143,144
l,r Pre	left and right Precuneus	153,154
l,r Amy	left and right Amygdalae	213,214
l,r VMPFC	left and right VentroMedial PreFrontal Cortex	41, 42
r DMPFC	right DorsoMedial PreFrontal Cortex	12
WM	White Matter	

Compensating for the hemodynamic response delay

The BOLD signal follows the Hemodynamic Response Function (HRF), which characterizes the Blood-oxygen-level-dependent (BOLD) signal to a single behavioral event that is recorded [37]. This function peaks with a delay of around five seconds, but that can vary somehow around this value depending on the brain area, the participant, as well as other factors that are not well known. To handle this variability, we express the discretized bold signal at time t based on a sequence of 4 previous consecutive observations of behavioral features between, which span the duration between $t - \tau_1$ and $t - \tau_2$, where $\tau_1 = 7.2s$ and $\tau_2 = 3.6s$ (Fig 3).

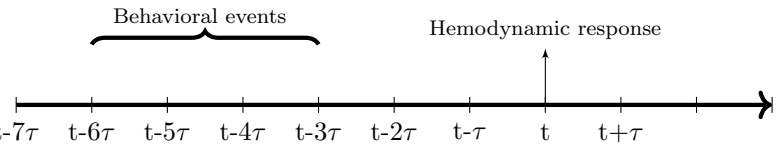

Figure 3. The time delay corresponding to the hemodynamic response to behavioral events is taken into account in the model by considering four consecutive behavioral time points happening between 7.2 and 3.6 seconds before the time t where the hemodynamic response is evaluated.

The number of observations in each sequence depends on the re-sampling rate of behavioral features with respect to the target variable. In the current study, we re-sample them with the same frequency of the BOLD signal, that is, one observation each $\tau = 1.2$s. We also tried to re-sample the BOLD signal based on interpolation to have an observation of each $0.6s$ to double the number of temporal features. The performance results show that this re-sampling does not improve the predictions while requiring more computational time. Therefore, we focus in this work on the same re-sampling rate for both BOLD and behavioral features.

Let y_t be the discretized time series associated to the BOLD signal, and $x_t = \{x_{1,t}, x_{2,t}, \ldots, x_{k,t}\}$ are k behavioral time series, representing the predictive features, where $x_{i,t}$ is the i^{th} variable at time t. We use the following notations to represent the sequences of behavioral variables in a concise way:

$$x_i^{t-\tau_1:t-\tau_2} = (x_{i,t-6\tau}, x_{i,t-5\tau}, x_{i,t-4\tau}, x_{i,t-3\tau})\,. \tag{3}$$

As mentioned before, we aim at predicting y at time t based on a sequence of x between $t - 8s$ and $t - 4s$, with a time-step $\tau = 1.2s$. This is a temporal classification problem, and with the previous notations, our dynamic model can be expressed as follows:

$$y_t = f\left(x_1^{t-\tau_1:t-\tau_2}, \ldots, x_k^{t-\tau_1:t-\tau_2}\right) + e_t, \tag{4}$$

where f is function of the model, and e_t represents its error vector.

Behavioral features processing (ii)

The aim of feature extraction is to compute high-level features from multimodal raw behavioral data, that may describe specific social and conversational factors involved in a conversation. Our feature extraction approach is based on interpretable features that are computed automatically from raw multimodal signals. From the raw recordings, we extract high-level features described in Table 2. They are also described with additional details in [6]. The extraction process is in itself performed using deep learning models that rely on multiple types of networks:

- Computer vision-based networks for emotion recognition, face and eye movements, *etc.*
- Classical and recurrent neural networks for audio and text data: sentiment analysis, semantic and structural features from the text, spectral features, *etc.*
- Classical Time series analysis for eyetracking signals.

Facial features are directly extracted from the video of the interlocutor, speech features are extracted from manual transcriptions of the participant and interlocutor's recorded conversations, and eye-tracking features are extracted from the participant's gaze recorded inside the fMRI scanner. The features are constructed as time series for

Table 2. Description of the extracted multimodal features.

	Features names	Description	Details
Linguistic features	Speech Activity	The interlocutor speaking?	Based on time-aligned IPU transcript.
	Overlap	Both interlocutors speaking?	*idem.*
	Laughter	Laughter occurrences	Based on word-level time-aligned transcripts.
	Filled-pauses	Filled-Pauses occurrences	Based on word-level time-aligned transcripts : *'euh', 'heu', 'hum', 'mh'.*
	Feedback	Conversational Feedback occurrences	Based on word-level time-aligned transcripts : *'oui' (yes), 'ouais' (yeah), 'non' (no), 'ah', 'd'accord' (right), 'ok'* + Laughters.
	Discourse-Markers	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), 'parce que' (because), 'ensuite' (after).*
	Spoken-Particles	Occurrence of (final) spoken particle items	Based on word-level time-aligned transcripts : *'quoi', 'hein','ben','bon' (well), 'mais' (but), 'bref' (in short).*
	Interpersonal	Merge of inter-personal linguistic features	Merge of (Filled-pauses, Feedback, Discourse Markers, Spoken Particles and Laughter).
	Turn-Latency	Time to take the turn	Based on time-aligned IPU transcript.
	SpeechRate	Speaking speed.	Based on time-aligned IPU transcript.
	Type-Token-Ratio	Lexical richness measure	Based on time-aligned transcript: (number of different tokens) / (total number of tokens).
	Lexical-Richness	Lexical richness measure [38].	Based on time-aligned transcript: (number of adjectives + number of adverbs) / (total number of tokens).
	Polarity & Subjectivity	Sentiment analysis metrics [39].	Based on time-aligned transcript, and a pre-trained KNN classifier.
Facial features	Head-Tx, Head-Ty, Head-Tz	Head translation	Based on head pose estimated using Openface.
	Head-Rx,Head-Ry, Head-Rz	Head rotation	*idem.*
	Head-T-energy	Kinetic energy of head translation	*idem.*
	Head-R-energy	Kinetic energy of head rotation	*idem.*
	AU-mouth	Sum of facial movements related to mouth.	Based on Facial Action Units (AUs) existence detected by Openface library.
	AU-eyes	Sum facial movements related to eyes.	*idem.*
	AU-all	Sum of all action units.	*idem.*
	Direct-gaze	Percentage of direct gaze direction of the conversant	*idem.*
	Emotions	('Neutral', 'Happiness','Sadness', 'Surprise', 'Fear', 'Anger', 'Disgust')	Based on a pre-trained CNN classifier.
	Smiles	Smile's detection.	Based on a CNN classifier from Opencv library.
Eyetracking	Saccades	Occurrence of Saccades	Based on gaze coordinates of the participant, recorded using the Eyelink1000 system.
	Gaze-speed	Gaze Speed.	*idem.*
	Gaze-movement-energy	Gaze movements energy	*idem.*
	Face-looks,Eyes-looks,Mouth-looks	Number of looks in face, eyes and mouths *respectively.*	Based on participant's gaze coordinates and interlocutor's detected landmarks.

each conversation. For example, in the case of speech, we analyzed Inter-Pausal Units (IPU[1]) by IPU (for example to compute the lexical richness) or word by word (*e.g.*, to compute Feedback and Discourse Markers). For videos, Openface 2.0 toolkit [41] is used to detect facial action units, landmarks, head pose estimation and gaze coordinates. Eyetracking coordinates of the participant are recorded using the Eyelink1000 system. We added other features characterizing where the participant is looking (Face, Eyes, Mouth), by combining the detected landmarks points of the conversant and the gaze coordinates.

In total, more than 40 features are extracted. Table 2 contains the names of the extracted features per modality with a description of how each feature is extracted. For more details, all features are available online[2], and are also described in [6].

To make the results more interpretable, we create higher-order features by grouping individual features (see Table 2) into meta-features that each represent one aspect of behavior that is relevant to social cognition (see Table 3). This underlying assumption is that it is not the importance score of each individual feature that is relevant, but the importance scores of all features pertaining to similar aspects of social cognition. For example, all individual features measuring the head movements of the interlocutor (translations, rotations, energy) are pooled together as "Head-movements of the interlocutor" or Head-Movement-I.

Table 3. Regrouping the extracted features into social meta-features (suffix "-P" for the participant, and "-I" for the interlocutor).

Meta-features	Original features
Head-movement-I	Head-Rx-I, Head-Ry-I, Head-Rz-I, Head-Tx-I, Head-Ty-I, Head-Tz-I, Head-translation-energy-I, Head-rotation-energy-I
Facial-movement-I	AU-all-I, AUs-mouth-I, AUs-eyes-I, Neutral-I
Emotions-I	Angry-I, Disgust-I, Fear-I, Happy-I, Sad-I, Surprise-I, Smiles-I
Eyetracking-P	Gaze-speed-P, Gaze-movement-energy-P, Saccades-P
Social-gaze	Face-looks-P, Mouth-looks-P, Eyes-looks-P, Direct-gaze-I
Interpersonal-I	FilledBreaks-I, Feedbacks-I, Discourses-I, Particles-I, Laughters-I, Interpersonal-I, Polarity-I, Subjectivity-I, Turn-Latency-I
Interpersonal-P	FilledBreaks-P, Feedbacks-P, Discourses-P, Particles-P, Laughters-P, Interpersonal-P, Polarity-P, Subjectivity-P, Turn-Latency-P
SpeechActivity-I	SpeechActivity-I, Overlap-I, SpeechRate-I
SpeechActivity-P	SpeechActivity-P, Overlap-P, SpeechRate-P
Linguistic-Complexity-I	LexicalRichness-I, TypeToken-Ratio-I
Linguistic-Complexity-P	LexicalRichness-P, TypeToken-Ratio-P

The extracted features require further processing in order to homogenize their structure. This processing includes resampling the obtained features with respect to the

[1]An IPU is a speech block of a single speaker bounded by pauses [40].
[2]`https://github.com/Hmamouche/NeuroTSConvers`

frequency of the fMRI signal and a concatenation of the time series of all trials and participants.

Predictions (*iii*)

The framework proposed in the previous section Analysis Framework is a general meta-model that can be used in a variety of ways depending on the dataset and under investigation. Now, we describe how it was applied to the dataset described above.

Human-human and human-robot data are evaluated separately in order to compare results from the two conditions. For each condition, the obtained data consist of 13248 observations. We fix the training set to 18 participants from 24, and we apply the ADASYN algorithm [42] on this set to address the problem of imbalanced data. This algorithm generates new observations by considering the distribution of the data.

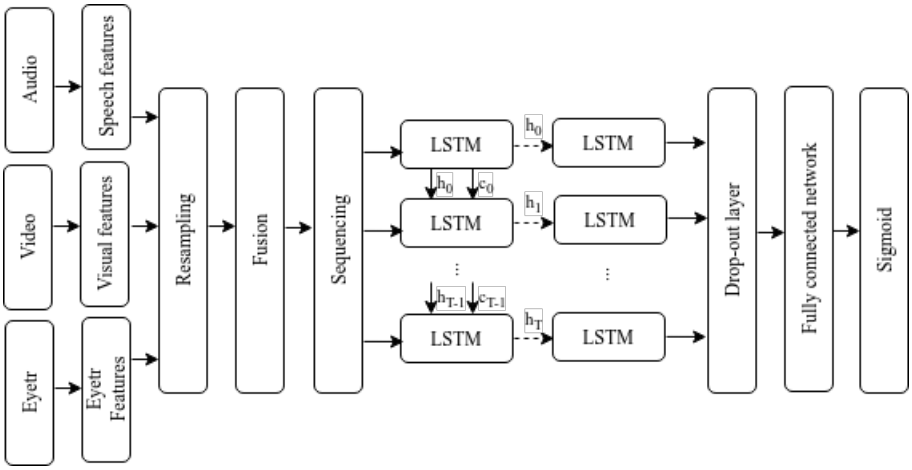

Figure 4. Illustration of our multimodal prediction approach. Feature extraction is performed based on specific methods and pre-trained models. Then the obtained time series are re-sampled and concatenated, then they are fed to a network composed of LReg and fully connected layers.

Then, we performed a 9-fold-cross-validation on training data to find the appropriate parameters of the classifier based on the F-score measure. We choose 9 folds instead of the classical 10 folds because we use data from 18 participants out of 24 in the training set and we want to evaluate the models on data of participants unseen by the models in order to avoid over-fitting. The models are then tested on data of 6 participants, that is, 25% of all data. Student's *t*-tests are performed to test the equality (null hypothesis) of the average F-scores between the different models and the baseline classifier obtained in the training step via the 9-fold cross-validation, with a significance threshold of $p \leq 0.05$.

Results

In this section, we present the obtained prediction results. They include performance scores of the used classifiers, as well as interpretable results about the most relevant features for each ROI.

Statistically significant predictions

Table 4. F-scores obtained by the best model in the different ROIs for the HHI (Human-Human Interaction) and HRI (Human-Robot Interaction) conditions. The *p*-values are provided for the best model. Bold indicates significant *p*-values at the threshold 0.05.

	F-scores HHI			F-scores HRI		
ROIs	**Best**	**Rand**	***p*-value**	**Best**	**Rand**	***p*-value**
lV1	0.58	0.50	*0.143*	0.56	0.50	*0.999*
rV1	**0.59**	0.53	***0.005***	0.58	0.51	*0.221*
lSTS	**0.71**	0.50	***≤0.001***	**0.69**	0.51	***≤0.001***
rSTS	**0.71**	0.49	***≤0.001***	**0.71**	0.52	***≤0.001***
lTPJ	**0.61**	0.50	***≤0.001***	**0.59**	0.51	***0.036***
rTPJ	**0.64**	0.50	***≤0.001***	**0.63**	0.54	***≤0.001***
lPre	**0.60**	0.51	***≤0.001***	0.57	0.50	*0.503*
rPre	**0.62**	0.50	***≤0.001***	**0.57**	0.50	***0.036***
lAmy	0.59	0.49	*0.081*	0.61	0.51	*0.114*
rAmy	**0.58**	0.48	***0.004***	0.64	0.51	*0.560*
lVMPFC	0.55	0.49	*0.190*	0.54	0.50	*0.216*
rVMPFC	**0.58**	0.50	***0.015***	0.54	0.51	*0.858*
rDMPFC	**0.66**	0.54	***≤0.001***	0.66	0.54	*0.157*
WM	0.54	0.51	*0.436*	0.51	0.51	*0.796*

F-score and recall measures are calculated for each condition (human-human and human-robot interactions), each region of interest as well as the different models used are given in supplementary material (Tables 7 and 8 for the F-score and the Recall respectively). Here, we focus on the models that yielded the best F-scores, reporting in Table 4 the recall score obtained for both conditions in all ROIs. Prediction of ROI activity of human-human interactions are significant for all ROIS except the lV1, lAmy and lVMPFC, while for human-robot interactions, there are only 5 areas with significant prediction F-scores (lSTS, rSTS, rPre, lTPJ, and rTPJ).

Interpretable prediction results

In this part, we present detailed results concerning the selected features that lead to the best predictions for each brain area. We also show their importance scores in order to discuss the impact of each social modality on the studied brain areas.

Table 5. Meta-features (with importance score $\geq$ 0.10) used to significantly predict ROIs activity in conditions of human-human interaction (HHI). Recall score is obtained from the best model classifier (see Table 8).

ROI	Recall score	Meta-features	Importance scores
rV1	0.63	Facial-movement-I	0.45
		Eyetracking-P	0.38
		Head-movement-I	0.16
lSTS	0.71	SpeechActivity-I	0.85
		Interpersonal-I	0.12
rSTS	0.70	SpeechActivity-I	0.52
		SpeechActivity-P	0.21
		Social-gaze	0.15
lTPJ	0.62	SpeechActivity-P	0.33
		SpeechActivity-I	0.25
		Head-movement-I	0.21
		Social-gaze	0.11
rTPJ	0.70	SpeechActivity-P	0.67
		Interpersonal-P	0.33
lPre	0.63	Eyetracking-P	0.58
		Head-movement-I	0.16
		Facial-movement-I	0.15
		SpeechActivity-P	0.11
rPre	0.63	Head-movement-I	0.38
		SpeechActivity-P	0.17
		Facial-movement-I	0.15
		Social-gaze	0.11
rAmy	0.63	SpeechActivity-P	0.26
		SpeechActivity-I	0.23
		Interpersonal-I	0.15
		Linguistic-Complexity-I	0.15
rVMPFC	0.61	Eyetracking-P	0.45
		Head-movement-I	0.34
rDMPFC	0.72	SpeechActivity-P	1.00

Table 6. Meta-features (with importance score $\geq$ 0.10) used to significantly predict ROIs activity in conditions of human-robot interaction (HRI). Other details as in Table 5.

ROI	Recall score	Meta-features	Importance scores
lSTS	0.69	SpeechActivity-I	1.00
rSTS	0.70	SpeechActivity-I	0.75
		SpeechActivity-P	0.25
lTPJ	0.62	Eyetracking-P	0.19
		SpeechActivity-I	0.16
		Facial-movement-I	0.15
		SpeechActivity-P	0.14
		Interpersonal-I	0.13
rTPJ	0.67	SpeechActivity-P	1.00
rPre	0.62	SpeechActivity-P	0.88
		Eyetracking-P	0.11

The results include recall score, and the set of the best predictive meta-features and their respective scores, to reach the prediction. Note that a threshold of 10% is used for reporting meta-features to focus on the most important ones. Tables 5 and 6 contain these results for human-human and human-robot interactions respectively.

Discussion

Results indicate that the proposed analysis framework is able to predict brain activity higher than chance in many brain areas but not in the white matter, as expected given that the white matter is not known to contain large BOLD responses. Even if a such response can be found with very specific settings, it is unlikely to be associated with the subtle and complex behavioral features we are investigating here. In addition, we used a whole-brain white matter mask to extract the white matter signal, so that any localized response is lost by the averaging of signal coming from different fibres, and it was also regressed out during the denoising of the fMRI data. This confirms the validity of the significant results. Importantly, predictions are significantly greater than chance for both human and robot conditions (HHI and HRI respectively) in the Superior Temporal Sulcus (STS) and TemporoParietal Junction (TPJ) bilaterally. The former was selected as being an area strongly devoted to language perception, as this region of the cortex also contains "voice patches", regions responding specifically to the perception of voices [43]. Therefore, perceiving speech from others is likely to activate these patches in a simple on/off fashion. Indeed, the highest F-scores are obtained in this region in both hemispheres and both interaction conditions. Furthermore, on the left side, dominant for language, the interlocutor's speech (SpeechActivity-I) is the only predictor (with an importance score of 1) used to reach a recall score of 0.69 in the HRI condition (see Table 6). Yet, there is ongoing speculation about a more general involvement of this part of the cortex in social cognition, for example in visual perception [44]. The second meta-feature used for prediction on the right STS region in HRI is the speech produced by the participant (with an importance score of 0.25), indicating a complex interaction between producing and perceiving speech in the non-dominant hemisphere. But in HHI, other meta-features are identified , pertaining to language in the left hemisphere (Interpersonal-I combines linguistic features associated with social aspects of the interaction) but also related to visual social cues (social gaze is related to the scanned participants' visual exploration of the interlocutor, including mutual gaze). In other words, while in both HRI and HHI, the second predictor for the right STS is the speech of the participant, these two speech-related meta-features are completed with social ones bilaterally for the human interlocutor, but not for the robot. Altogether, this results derived from a natural discussion with a natural and artificial agent fit largely with expectations, with specialized left STS for speech perception, right STS being also involved in speech production, and non-verbal social behavioral meta-features when the interlocutor is a human.

A second series of observations can be made for areas associated with social cognition in which we obtained significant F-scores. These areas include rTPJ, lTPJ, and rPre for both HHI and HRI, and VMPFC, Amy and rDMPFC for HHI only. First, their predictions score are smaller than those of the left and right STS. Then, in a number of these areas, notably the left TemporoParietal Junction (TPJ) for both HHI and HRI and the left and right Precuneus (lPre & rPre), right Amygdala (rAmy) for HHI only, a combination of several meta-features, combining different types of signals, including social ones, are combined to provide significant predictions. These two remarks support the hypothesis that these areas are associating more complex multimodal signals than the STS. Given that we acknowledge that defining all features (and meta-features) relevant for social cognition is exploratory, we do expect activity in these areas to be less predictable. Interestingly, the speech activity of the participant is the only predictor for the right TPJ in HRI, while the speech activity for the interlocutor was the only meta-feature for right STS despite the closeness of these two areas on the cortical surface (see figure 2. This result rules out the possibility that non-specific relations could be identified due to the proximity of our regions of interest. It should also be noted that a number of meta-features unrelated to speech are used to significantly predict

activity in the left TPJ for both human and robot interlocutors, including visual signals (head movements in HHI, facial movements in HRI), highlighting a complex function of combining information from different modalities while experimental paradigms that focus on individual sources of information are usually preferentially associated with right hemisphere responses. It is possible that complex associations of multiple modalities are not correctly captured by classical experimental paradigms focusing on more controlled, and unimodal, aspects of social cognition.

The right Primary Visual Cortex (rV1) corresponds to the input of retinal information in the cortex. Indeed, its response is explained by the head and facial movements of the interlocutor as well as the participant's eye movements, all related to the visual input. It was not predicted in HRI, possibly because the robot's movements were very limited compared to the human's. The medial PreFrontal Cortex region is only associated with speech production in HHI, which is more consistent with its role in motor control language [45] than in mentalizing [46]. Its absence in HRI could also be explained by the fact that speech was also lessened in HRI compared to HRI. Predictions in areas involved in emotional aspects of social cognition, VentroMedial PreFrontal Cortex and Amygdala, yield very similar results: activity can only be predicted for human-human interaction and in the right hemisphere. Several meta-features are used for predictions, that include Interpersonal-I for the amygdala and eye and head movements from the participant and interlocutor respectively for the VentroMedial PreFrontal Cortex. Altogether, these results concerning higher-order associative areas (that are shared to a large extend by the Precuneus too, known to have a more complex response in HHI than HRI in this experiment [47])- increased predictions for the human interactions, limited to the right hemisphere, and relying on complex associations of meta-features - all indicate that the analysis approach used here is valid to identify relations between complex multimodal recordings of natural behaviors and local brain activity, but it is less statistically powerful when relations between multimodal behavioral features and brain activity are limited - either by the limited association between the behavior and cognition in the case of emotion processing in the left hemisphere or by the poverty of the interaction in the case of the robot.

This final remark is confirmed by the p-values of the statistical test performed of the mean of F-scores of the best classifier and the baseline (random predictions generator) using the Students t-test; the results show that the p-values obtained for HRI are greater than those of HHI for all brain areas investigated. This implies that the brain activity in HRI are less predictable than in HHI, as was predicted given that interacting with a human is more natural than interacting with a robot

Conclusion

In this paper, we present a new framework for predicting fMRI brain activity from behavioral signals and identifying dependencies between them during natural conversations between two agents. Evaluations are made on a corpus containing several human-human and human-robot conversations of 24 participants. We focused on brain areas involved in speech perception and in social interaction. We obtained predictions significantly better than those obtained with a random classifier in many of the brain regions investigated. The obtained dependencies confirm existing hypotheses about the relationship between circumscribed aspects of behaviors and functional brain areas but also allows to address new questions about the multimodal integration of behaviors in brain areas involved in higher-order aspects of cognition. The framework's output provides new and precise results, by finding the best possible model, the associated subset of relevant features and a quantification of their impact for each brain area. In addition, they allow comparing the difference between brain activation in the cases of human-human and human-robot

interaction in terms of the predictive behavioral features for each brain area. The next step of this work is to test the framework on all brain areas. Understanding why certain regions cannot be predicted is a puzzling finding requiring further investigation and in particular the addition of new behavioral features in the future.

References

1. Hlaváčková-Schindler K. Equivalence of granger causality and transfer entropy: A generalization. Applied Mathematical Sciences. 2011;5(73):3637–3648.

2. Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, et al. Predicting Human Brain Activity Associated with the Meanings of Nouns. Science. 2008;320(5880):1191–1195. doi:10.1126/science.1152876.

3. Yarkoni T, Barch DM, Gray JR, Conturo TE, Braver TS. BOLD Correlates of Trial-by-Trial Reaction Time Variability in Gray and White Matter: A Multi-Study fMRI Analysis. PLOS ONE. 2009;4(1):e4257. doi:10.1371/journal.pone.0004257.

4. Chen HY, Liao YH, Jan HT, Kuo LW, Lee CC. 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; 2016. p. 5775–5779.

5. Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature. 2016;532(7600):453–458. doi:10.1038/nature17637.

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

7. Chaminade T. An experimental approach to study the physiology of natural social interactions. Interaction Studies. 2017;18(2):254–275. doi:10.1075/is.18.2.06gry.

8. Hallart C, Maes J, Spatola N, Prévot L, Thierry C. Comparaison linguistique et neuro-physiologique de conversations humain humain et humain robot. Revue TAL. 2021;61(3):69–93.

9. Chaminade T, Spatola N. Perceived facial happiness during conversation correlates with insular and hypothalamus activity for humans, not robots. Frontiers in Psychology. 2022;6. doi:10.3389/fpsyg.2022.871676.

10. Rauchbauer B, Hmamouche Y, Bigi B, Prévot L, Ochs M, Chaminade T. 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. Marseille, France: European Language Resources Association; 2020. p. 668–675. Available from: `https://aclanthology.org/2020.lrec-1.84`.

11. Gupta R, Malandrakis N, Xiao B, Guha T, Van Segbroeck M, Black M, et al. Multimodal prediction of affective dimensions and depression in human-computer interactions. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge; 2014. p. 33–40.

12. Brady K, Gwon Y, Khorrami P, Godoy E, Campbell W, Dagli C, et al. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge; 2016. p. 97–104.

13. Morency LP, de Kok I, Gratch J. A probabilistic multimodal approach for predicting listener backchannels. Autonomous Agents and Multi-Agent Systems. 2010;20(1):70–84.

14. McKeown G, Valstar MF, Cowie R, Pantic M. The SEMAINE corpus of emotionally coloured character interactions. In: 2010 IEEE International Conference on Multimedia and Expo; 2010. p. 1079–1084.

15. Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence. 2018;41(2):423–443.

16. Okada K, Venezia JH, Matchin W, Saberi K, Hickok G. An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex. PLoS ONE. 2013;8(6). doi:10.1371/journal.pone.0068959.

17. Bone D, Lee CC, Narayanan S. Robust Unsupervised Arousal Rating: A Rule-Based Framework with Knowledge-Inspired Vocal Features. IEEE transactions on affective computing. 2014;5(2):201–213. doi:10.1109/TAFFC.2014.2326393.

18. Knops A, Thirion B, Hubbard EM, Michel V, Dehaene S. Recruitment of an Area Involved in Eye Movements During Mental Arithmetic. Science. 2009;324(5934):1583–1585. doi:10.1126/science.1171599.

19. DeSouza JFX, Ovaysikia S, Pynn LK. Correlating Behavioral Responses to fMRI Signals from Human Prefrontal Cortex: Examining Cognitive Processes Using Task Analysis. Journal of Visualized Experiments : JoVE. 2012;64. doi:10.3791/3237.

20. Giordano BL, Whiting C, Kriegeskorte N, Kotz SA, Gross J, Belin P. The representational dynamics of perceived voice emotions evolve from categories to dimensions. Nature Human Behaviour. 2021;5(9):1203–1213. doi:10.1038/s41562-021-01073-0.

21. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys Rev E. 2004;69:066138.

22. Kozachenko L, Leonenko NN. Sample estimate of the entropy of a random vector. Problemy Peredachi Informatsii. 1987;23(2):9–16.

23. Yu W, Ruibo W, Huichen J, Jihong L. Blocked 3× 2 cross-validated t-test for comparing supervised classification learning algorithms. Neural computation. 2014;26(1):208–235.

24. INT, LPL. convers; 2020. Available from: `https://hdl.handle.net/11403/convers/v2`.

25. Al Moubayed S, Beskow J, Skantze G, Granström B. Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction. In: Esposito Aea, editor. Cognitive Behavioural Systems. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2012. p. 114–130.

26. Penny WD, Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Elsevier; 2011.

27. Ashburner J. A fast diffeomorphic image registration algorithm. NeuroImage. 2007;38(1):95–113. doi:10.1016/j.neuroimage.2007.07.007.

28. Whitfield-Gabrieli S, Nieto-Castanon A. Conn: A Functional Connectivity Toolbox for Correlated and Anticorrelated Brain Networks. Brain Connectivity. 2012;2(3):125–141. doi:10.1089/brain.2012.0073.

29. Kasper L, Bollmann S, Diaconescu AO, Hutton C, Heinzle J, Iglesias S, et al. The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. Journal of Neuroscience Methods. 2017;276:56–72. doi:10.1016/j.jneumeth.2016.10.019.

30. Fan L, Li H, Zhuo J, Zhang Y, Wang J, Chen L, et al. The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. Cereb Cortex. 2016;26(8):3508–3526. doi:10.1093/cercor/bhw157.

31. Ezaki T, Watanabe T, Ohzeki M, Masuda N. Energy landscape analysis of neuroimaging data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2017;375(2096):20160287. doi:10.1098/rsta.2016.0287.

32. Deco G, Kringelbach ML. Hierarchy of Information Processing in the Brain: A Novel 'Intrinsic Ignition' Framework. Neuron. 2017;94(5):961–968. doi:10.1016/j.neuron.2017.03.028.

33. Watanabe T, Hirose S, Wada H, Imai Y, Machida T, Shirouzu I, et al. A pairwise maximum entropy model accurately describes resting-state human brain networks. Nature Communications. 2013;4(1):1370. doi:10.1038/ncomms2388.

34. Weistuch C, Mujica-Parodi LR, Razban RM, Antal B, van Nieuwenhuizen H, Amgalan A, et al. Metabolism modulates network synchrony in the aging brain. Proceedings of the National Academy of Sciences. 2021;118(40):e2025727118. doi:10.1073/pnas.2025727118.

35. Ezaki T, Fonseca dos Reis E, Watanabe T, Sakaki M, Masuda N. Closer to critical resting-state neural dynamics in individuals with higher fluid intelligence. Communications Biology. 2020;3(1):52. doi:10.1038/s42003-020-0774-y.

36. Otsu N. A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics. 1979;9(1):62–66.

37. Gössl C, Fahrmeir L, Auer D. Bayesian modeling of the hemodynamic response function in BOLD fMRI. NeuroImage. 2001;14(1):140–148.

38. Ochs M, Jain S, Blache P. 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; 2018. p. 161–166.

39. Smedt TD, Daelemans W. Pattern for python. Journal of Machine Learning Research. 2012;13(Jun):2063–2067.

40. Levitan R, Hirschberg J. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: Twelfth Annual Conference of the International Speech Communication Association; 2011.

41. Baltrusaitis T, Zadeh A, Lim YC, Morency L. OpenFace 2.0: Facial Behavior Analysis Toolkit. In: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018); 2018. p. 59–66.

42. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE; 2008. p. 1322–1328.

43. Pernet CR, McAleer P, Latinus M, Gorgolewski KJ, Charest I, Bestelmeyer PEG, et al. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. NeuroImage. 2015;119:164–174. doi:10.1016/j.neuroimage.2015.06.050.

44. Allison T, Puce A, McCarthy G. Social perception from visual cues: role of the STS region. Trends in Cognitive Sciences. 2000;4(7):267–278. doi:https://doi.org/10.1016/S1364-6613(00)01501-1.

45. Picard N, Strick PL. Motor areas of the medial wall: a review of their location and functional activation. Cerebral cortex (New York, NY : 1991). 1996;6(3):342–353. doi:10.1093/cercor/6.3.342.

46. Bird CM, Castelli F, Malik O, Frith U, Husain M. The impact of extensive medial frontal lobe damage on 'Theory of Mind' and cognition. Brain. 2004;127(4):914–928. doi:10.1093/brain/awh108.

47. Spatola N, Chaminade T. Precuneus brain response changes differently during human–robot and human–human dyadic social interaction. Scientific Reports. 2022;12(1):14794. doi:10.1038/s41598-022-14207-9.

Supplementary information

Detailed prediction results

The detailed prediction F-scores and recalls with all the models tested in this analysis tested on human-human and human-robot datasets are provided in Tables 7 and 8.

Table 7. The F-scores obtained by each model.

ROIs	F-scores HHI						F-scores HRI					
	LReg	**LSTM**	**FNN**	**RF**	**SVM**	**Rand**	**LReg**	**LSTM**	**FNN**	**RF**	**SVM**	**Rand**
lV1	**0.58**	0.57	0.56	0.57	0.57	0.50	0.52	0.53	0.53	**0.56**	0.55	0.50
rV1	0.59	0.59	0.55	**0.60**	**0.60**	0.53	0.56	0.57	0.56	**0.58**	**0.58**	0.51
lSTS	**0.71**	0.70	0.70	**0.71**	0.68	0.50	0.68	0.68	0.66	**0.69**	0.68	0.51
rSTS	0.70	0.69	0.69	**0.71**	0.68	0.49	0.70	**0.71**	0.70	0.67	0.63	0.52
lTPJ	0.58	0.58	0.59	**0.61**	0.60	0.50	**0.59**	0.58	0.55	0.58	0.56	0.51
rTPJ	0.63	0.63	0.61	**0.64**	**0.64**	0.50	0.62	0.62	0.59	**0.63**	**0.63**	0.54
lPre	0.55	0.58	0.55	0.60	0.59	0.51	0.52	0.54	0.53	**0.57**	0.56	0.50
rPre	0.57	0.58	0.58	**0.62**	0.60	0.50	0.49	0.49	0.50	0.56	**0.57**	0.50
lVMPFC	0.47	0.50	0.55	0.54	0.55	0.49	0.47	0.52	**0.54**	**0.54**	**0.54**	0.50
rVMPFC	0.54	0.55	0.57	0.57	**0.58**	0.50	0.47	0.45	0.49	**0.54**	**0.54**	0.51
rDMPFC	**0.66**	0.60	0.57	0.65	0.65	0.54	**0.66**	0.64	0.58	0.62	0.64	0.54
lAmy	0.58	0.55	0.54	**0.59**	0.53	0.49	0.55	0.58	0.55	0.58	**0.61**	0.51
rAmy	0.53	0.57	**0.58**	0.57	0.54	0.48	**0.64**	0.61	0.58	0.61	0.61	0.51
WM	0.47	0.46	0.49	**0.54**	0.53	0.51	0.46	0.46	0.46	**0.51**	0.38	**0.51**

Table 8. The Recall scores obtained by each model.

ROIs	Recalls HHI						Recalls HRI					
	LReg	**LSTM**	**FNN**	**RF**	**SVM**	**Rand**	**LReg**	**LSTM**	**FNN**	**RF**	**SVM**	**Rand**
lV1	**0.62**	**0.62**	0.57	0.58	0.61	0.49	**0.59**	**0.59**	**0.59**	0.56	0.55	0.49
rV1	**0.63**	0.59	0.55	0.60	0.62	0.52	**0.63**	0.61	0.60	0.60	0.58	0.50
lSTS	**0.71**	**0.71**	**0.71**	0.70	0.68	0.49	**0.69**	0.68	0.67	**0.69**	0.68	0.50
rSTS	**0.70**	0.70	0.69	**0.71**	0.67	0.48	0.70	**0.71**	0.69	0.66	0.63	0.51
lTPJ	**0.62**	**0.62**	0.60	0.61	0.60	0.49	**0.62**	0.59	0.56	0.58	0.55	0.50
rTPJ	**0.66**	0.64	0.63	0.64	0.65	0.49	**0.67**	**0.67**	0.63	0.63	0.65	0.52
lPre	**0.63**	0.62	0.60	0.62	0.59	0.50	**0.62**	0.55	0.54	0.57	0.59	0.49
rPre	**0.63**	0.61	0.59	**0.63**	0.62	0.49	0.61	**0.62**	**0.62**	0.58	0.58	0.49
lVMPFC	0.57	0.56	**0.58**	0.54	0.54	0.49	0.59	0.52	0.56	0.54	**0.57**	0.50
rVMPFC	**0.61**	0.58	0.59	0.58	0.58	0.49	0.57	0.57	**0.61**	0.54	0.57	0.50
rDMPFC	**0.72**	0.68	0.57	0.65	0.65	0.51	0.69	**0.71**	0.61	0.61	0.66	0.51
lAmy	**0.65**	0.54	0.55	0.59	0.52	0.48	**0.66**	0.63	0.59	0.63	0.62	0.50
rAmy	**0.61**	0.59	0.59	0.57	0.59	0.48	**0.64**	0.60	0.58	0.61	0.61	0.50
WM	**0.61**	**0.61**	0.60	0.54	0.54	0.51	**0.61**	**0.61**	**0.61**	0.51	0.43	0.50

Table 9. Model and feature selection yielding the best results on ROIs with significant predictions for HHI conditions.

ROI	Model	Feature selection	F-score	Recall score
lV1	SVM	Model-rank	0.60	0.62
lSTS	RF	Model-rank	0.61	0.71
rSTS	RF	MI-rank	0.71	0.70
lTPJ	RF	Model-rank	0.61	0.61
rTPJ	SVM	k-medoids	0.64	0.65
lPre	RF	Model-rank	0.60	0.62
rPre	RF	MI-rank	0.62	0.63
rVMPFC	SVM	Model-rank	0.58	0.58
rDMPFC	LSTM	MI-rank	0.66	0.72
rAmy	FNN	k-medoids	0.58	0.59

Table 10. Model and feature selection yielding the best results on only ROIs with significant predictions for HRI conditions.

ROI	Model	Feature selection	F-score	Recall score
lSTS	RF	Model-rank	0.69	0.69
rSTS	LSTM	MI-rank	0.71	0.71
lTPJ	LReg	Model-rank	0.59	0.62
rTPJ	SVM	k-medoids	0.63	0.65
rPre	SVM	MI-rank	0.57	0.58