Identifying Causal Relationships Between Behavior and Local Brain Activity During Natural Conversation

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

[1] Aix Marseille Université, Université de Toulon, CNRS, LIS, UMR7020, Marseille, France
[2] Aix Marseille Université, CNRS, LPL, [3] Institut Universitaire de France, Paris, France
[4] Aix Marseille Université, CNRS, INT, UMR7289, Marseille, France

firstname.lastname@univ-amu.fr

Abstract

Characterizing precisely neurophysiological activity involved in natural conversations remains a major challenge. We explore in this paper the relationship between multimodal conversational behavior and brain activity during natural conversations. This is challenging due to Functional Magnetic Resonance Imaging (fMRI) time resolution and to the diversity of the recorded multimodal signals. We use a unique corpus including localized brain activity and behavior recorded during a fMRI experiment when several participants had natural conversations alternatively with a human and a conversational robot. The corpus includes fMRI responses as well as conversational signals that consist of synchronized raw audio and their transcripts, video and eye-tracking recordings. The proposed approach includes a first step to extract discrete neurophysiological time-series from functionally well defined brain areas, as well as behavioral time-series describing specific behaviors. Then, machine learning models are applied to predict neurophysiological time-series based on the extracted behavioral features. The results show promising prediction scores, and specific causal relationships are found between behaviors and the activity in functional brain areas for both conditions, *i.e.*, human-human and human-robot conversations.

Index Terms: multimodal signals processing, natural conversation, machine learning, human-human and human-machine interactions, Functional MRI

1. Introduction

Identifying dependencies between behavior and brain activity during conversations is an essential step towards understanding the brain bases of conversational speech. We study here those dependencies in two conditions: human-human vs. human-robot interaction. Besides exploring differences in neurophysiological processes of an subject conversing with a human or with a robot, it also enables us to scrutinize brain activities related to conversation in the case of impoverished social context. A strength of our approach is the use of fMRI that provides invaluable data to localize brain activity and with a methodology that allows a relatively fine grained sampling in the time domain as well. The result is a unique data set of natural language conversations recorded in fMRI, providing synchronized neurophysiological and behavioral signals [1].

This dataset is unique in that participants' behaviour is unconstrained and therefore different from fMRI dataset generally acquired in highly controlled conditions. The classical approach of contrasting two or more well controlled experimental conditions therefore cannot be used. Existing works (*cf.* Section 2) have a major drawback: They use one or a small number of behavioral signals that are derived from very controlled tasks. Our contribution consists in handling complex raw multimodal behavioral signals acquired during relatively unconstrained conversations, and deriving from them high-level features as predictive variables to predict fMRI responses in well-defined functional regions of interest (ROI).

In this paper, we present a framework for (i) predicting fMRI responses based on behavioral signals in localized brain areas recorded during conversations, and (ii) identifying causal relationships between conversational behavior and brain activity. The proposed approach consists of two main steps:

1. *Feature extraction*: high level (verbal and non-verbal) features are extracted from raw behavioral data. They are grounded on speech produced by the two interlocutors, as well as eye-tracking signals of the participant and the video of the human or artificial interlocutor.
2. Applying *feature selection and prediction* based on the extracted features and the fMRI time series.

Evaluations are performed on 24 participants (*cf.* Section 3). The best prediction result allows us to identify the most relevant features for each brain area, yielding on the way a discussion comparing the cases where the interlocutor is a human or a robot.

After presenting related work in the next section, we describe the fMRI experiment and data sets acquisition in section 3. Then, we present our approach in section 4 and our results in section 5.

2. Related Work

Several approaches have been proposed in the literature to predict brain activity based on behavior. In [2], the authors investigate the effect of adding the visual speech to auditory speech signals in increasing the activity of auditory cortex areas. The results show significant increase in the activation of the studied regions of interests (ROIs) based on ANOVA analysis. In [3], the fMRI neural activation associated to meanings is predicted based on a large text data. The brain regions studied are in the sensory-motor cortex. The model used consists of transforming the text into semantic features, 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 multi-subject approach, *i.e.*, by concatenating data from multiple subjects. For example, in [4], the goal was to predict voxels activity from cortical areas, measured via the Blood-Oxygen-Level Dependent (BOLD) 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 the 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 the brain activity. For example, the speech reaction time has been used to predict activity in specific brain regions [5]. In [6], the prosodic mimicry generated by computers has been used to study its effect on interpersonal relations in human–computer interaction. In [7], the acoustically-derived vocal arousal score [8] is used to predict the BOLD signal using the Gaussian mixture regression model. In [9], the authors predict the BOLD signal in the posterior parietal cortex based on eye movement data using a multivariate regression model. More general approaches try to predict the brain activity of various areas using different types of signals at the same time. For example, in [10], 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.

These works analyse dependencies between behavior and specific functional brain areas. However, one or few modalities are included to describe the behavior used to predict the brain activity. In addition, the methods used are generally based on correlation analysis or multiple regression.

However, machine learning methods can be particularly relevant for this kind of questions, such as prediction models based on decision trees and artificial neural networks, aside with feature selection techniques. In our case, we propose a framework that consists in extracting high-level features from raw data, then applying feature selection and prediction with different classifiers to predict discretized neuro-physiological signals from multimodal behavioral signals composed of audios, videos and eye-tracking recording.

3. Datasets acquisition and processing

The data is collected from an fMRI experiment described in [1], and illustrated in Figure 1. It involves 24 participants, and consists of four sessions, each containing six conversations of 60 seconds, three with a human and three with a conversational robot alternatively. An *"advertising campaign"* provides a cover story: participants are informed 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 [11] (controlled by the confederate in a *Wizard-of-Oz* mode, unbeknown to the participant), is about one single image of the purported *"advertising campaign"*.

3.1. Processing fMRI signals

Standard fMRI acquisition procedures were used, described in details in [1]. BOLD signal 3-dimensional images are recorded in the whole brain every 1.205 seconds. Standard SPM12 preprocessing procedures are used [12], 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 [13] 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 [14], 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 artifacts with the PhysIO toolbox [15], and finally extracting BOLD signal in the white matter and cerebrospinal fluid and using the 5 first eigen variate of the time-series as nuisance representing signal fluctuations in non-cortical brain tissues. A 275-area parcellation based on functional and anatomical connectivity patterns [16] defines ROIs for the whole brain, and specific regions are chosen based on their anatomical location. Continuous time-series (385 time points) are extracted for each ROI and each session and participant representing the mean activity after denoising.

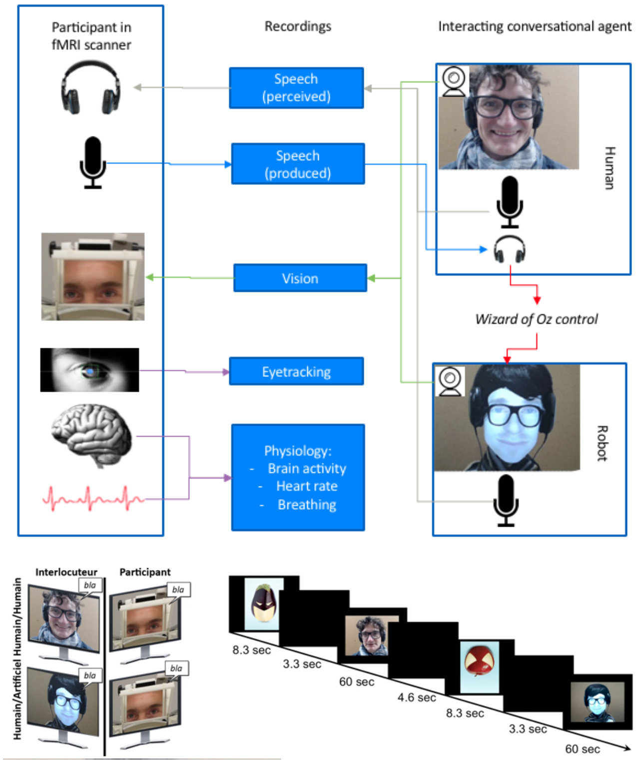

Figure 1: *The experimental design.*

4. The Proposed Approach

We propose a framework for (i) predicting local brain activity based on multimodal behavioral signals during bidirectional conversations, and (ii) identifying dependencies between them. The idea is to mimic how brain areas activate in a bidirectional conversation depending on what the brain receives and produces. A schema of this framework is shown in Figure 2. It consists of two main modules. A *predictors extractor* module which takes as input raw multimodal signals, then extracts high level features using methods and pre-trained models. Next, it resamples them to have time series with the same number of observations. Finally, the time series are represented as sequences

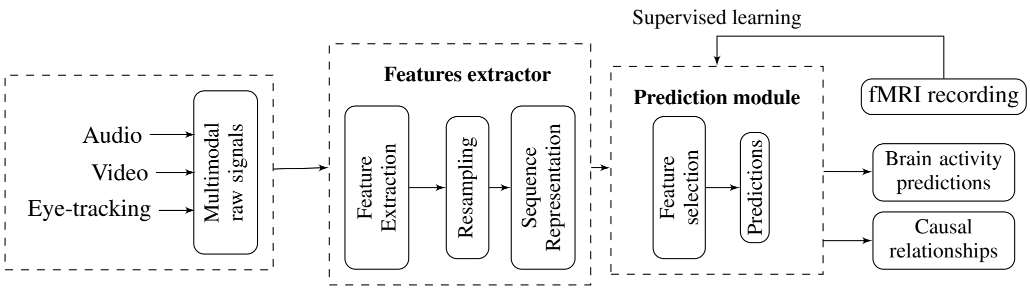

Figure 2: *A schema of the used multimodal prediction process.*

using the same number of time-steps according to Equation 1. The *prediction* module applies feature selection and learns supervised classifiers from the processed features and the BOLD signals. In the rest of this section, we detail the two parts of the proposed framework.

4.1. Extracting features from multimodal signals

This step involves processing 3 types of signals. Facial features are directly extracted from video of the interlocutor, speech features are extracted from manual transcripts, and eye-tracking features are extracted from participant's data recorded inside the fMRI scanner. In total, more than 40 features are extracted. For the sake of space, we just present the name of these features per modality in Table 1. For more details, all features are available online[1], and described in [17].

Raw signals	Extracted Features
Audios	Speech activity, interpersonal speech items (particles items and discourse markers [18]), Overlap, Reaction Time, Filled-breaks, Feedbacks, Laughters, Lexical Richness, Sentiment analysis (polarity and subjectivity).
Videos	Facial Action Units, Head pose coordinates, Gaze coordinates, smiles, emotions.
Eye-tracking data	Coordinates of the gaze movement, and binary variables categorizing *resp.* the presence of saccades, and where the participant is looking at interlocutor's face, eyes of mouth.

Table 1: *Summary of the extracted behavioral features.*

The extracted time series have two types (discrete and continuous), and have different frequencies, while the used machine learning models require variables with the same number of observations. We adopted two resampling methods. For continuous features, we resample them by considering their relative average between each two successive time points. For discrete features, we consider the number of occurrences.

4.2. The prediction module

Before starting the prediction formulation, let us describe the dynamic link BOLD signal and the behavior events from a neuroscience point of view. The fMRI response to a behavioral event is characterized by the Hemodynamic Response Function (HRF), which peaks around 5 seconds after a trigger event [19]. We focus in this work on predicting if a brain area is active or not. Therefore, we discretize first the BOLD signal in each ROI. The BOLD time series of each participant are normalized, then discretized using thresholding. The thresholds used are chosen after cross-validating the prediction module with different values close to the mean activation of each ROI, then selecting the appropriate value in terms of the prediction scores.

Let $Y(t)$ be the discretized BOLD signal of a given brain area, and $X(t) = \{X_1(t), X_2(t), \ldots, X_k(t)\}$ be k-dimensional time series representing the behavioral variables. One could express each value of the BOLD signal at time t as a function of the predictive features at time $t - 5s$. Considering the fact that the delay varies around $5s$ depending on behaviors, brains areas, and subjects, our approach consists in considering more points near to this delay in order to cover its variability. We express the BOLD signal at time t as a function of the lagged variables of each feature between times $t - 7$s and $t - 4$s. This model can be written as follows:

$$Y(t) = f(X_1^{t-\tau_1:t-\tau_2}, \ldots, X_k^{t-\tau_1:t-\tau_2}) + U(t), \quad (1)$$

where f is the function that we aim at determining, $X_i^{t-\tau_1:t-\tau_2}$ 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.

5. Evaluations and results

5.1. Prediction Procedure

In the current study, we focus on 6 ROIs: the left and right Fusiform Gyrus (FG) involved in face perception, the left and right Motor Cortex (MC) which support speech production, and the left and right Superior Temporal Sulcus (STS) involved in speech perception. Two evaluations are performed independently in two conditions: human-human and human-machine interactions. For each condition, the obtained data contain 13248 observations[1] in which the data associated to 4 participants ($\approx 17\%$) are kept as test set. The classifiers used are from the Scikit-learn library [20]: Support-Vector Machine (SVM), Random Forest (RanForest), and the Logistic Regression (LogReg). Since the predictive variables are in form of sequences, we also used the Long Short Term Memory (LSTM) network from the Tensorflow library [21]. A baseline classifier is used with 3 strategies: a stratified way by generating random predictions regarding the distribution of the training data, a uniform way by generating random predictions uniformly, and the last one based on the most frequent label.

Brain activation does not follow the same distribution in all ROIs. For example some ROIs may be activated rarely during conversations depending on the situation. This may cause

[1] The processed datasets, the implementations, and the detailed results are available in `https://github.com/Hmamouche/NeuroTSConvers`, last accessed on 28/07/2020.

ROIs	F-scores (Human-human)					F-scores (Human-machine)				
	LogReg	LSTM	RanForrest	SVM	baseline	LofReg	LSTM	RanForrest	SVM	baseline
Left FG	0.59	0.65	**0.66**	0.57	0.54	0.63	**0.69**	**0.69**	0.62	0.55
Left MC	0.70	0.67	**0.72**	0.71	0.54	0.68	0.64	**0.72**	**0.72**	0.55
Left STS	0.70	0.70	**0.72**	0.71	0.53	0.65	0.65	**0.66**	**0.66**	0.51
Right FG	**0.64**	0.61	**0.64**	0.59	0.53	**0.65**	**0.65**	**0.65**	0.60	0.54
Right MC	0.71	0.65	**0.73**	**0.73**	0.53	0.68	0.64	**0.73**	**0.73**	0.53
Right STS	0.68	0.68	**0.69**	0.65	0.53	0.66	0.60	**0.68**	**0.68**	0.54

Table 2: *Prediction results on test data. The F-scores of classifiers are provided for both human-human and human-robot interactions.*

imbalanced data, and can affect the classification quality. To handle this issue, we applied the ADASYN algorithm on the training data [22], which generates synthetic observations taking into account the distribution of the data. Then, a 10-fold-cross-validation is applied on training data to find the appropriate parameters of the classifiers and avoiding over-fitting. This is conducted on all classifiers except the LSTM network, because it takes a huge amount of time. For this specific model, we applied one training-test pass directly with a fixed architecture composed of one LSTM hidden layer and a fully connected output layer containing one neuron to provide one prediction each time using the sigmoid activation function. The network is trained using the ADAM optimization algorithm [23], and the binary cross-entropy is used as loss function.

Feature selection is performed with the help the classifiers themselves by ranking features based on their weights. To evaluate the predictions, three classification measures are considered, the weighted recall, precision and F-score. We focus here on the weighted F-score.

5.2. Results

Table 2 shows the prediction results for both human-human and human-robot data. Overall, the Random Forest provides the best predictions with F-scores between 0.64 and 0.73. The *Student's t-test* is applied to test the equality (null hypothesis) of the means of the F-scores between the best and the baseline classifiers obtained by the 10-fold-cross-validation. This test is one of recommended methods to compare the performance of machine learning algorithms [24]. The obtained p-values (the probability of the null hypothesis) are shown in Table 3. For MC and STS, the p-values are less than 0.01. The best F-scores are significantly better than the baseline F-scores at a significance threshold less than 1%. The right FG area are the most difficult to predict, where the p-value is equal to 0.005.

ROIs	T-test pvalues	
	Human-human	Human-robot
Left MC	1.34e-06	2.61e-07
Right MC	4.04e-07	1.59e-07
Left STS	4.62e-07	8.27e-10
Right STS	2.05e-08	4.41e-07
Left FG	0.00041	0.00091
Right FG	0.00508	0.00075

Table 3: *The obtained p-values of the statistical test (T-test) between F-scores of best and baseline classifiers.*

5.3. Discussion

The obtained results show promising F-scores, especially for MC and STS areas. However, it is not easy to get very accurate predictions perhaps because brain activities might depends on other factors that can not be recorded during this fMRI experiment.The statistical test used to compare the best classifier and the baseline can be applied to compare the results of a classifier with two different subsets of features. We performed that by comparing the F-scores of the best classifier with two conditions: the first one using the best set of features obtained via feature selection, and the second one using non related features, which belong to the set of non-selected features that are supposed to have no effect on the ROIs. We obtained very low p-values. Finally, we discuss the obtained dependencies between behavior and the brain activity. They are the output of the combination between feature selection and prediction, by selecting, for each ROI, the smallest subset of features yielding the prediction scores. These dependencies can be divided into two categories:

(i) Existing dependencies: our results confirm existing hypothesis related to MC and FG areas. For left and right MC, only the speech activity of the participant is selected for both human-human and human-robot interactions. For left and right FG, the selected features are: variables characterizing where the participant is looking (face, eye or mouth), head movements, facial expressions of the interlocutor, and participant's saccades and eyes movements speed.

(ii) New dependencies: for left and right STS areas, in addition to existing hypothesis which include speech perception, we found three modalities involved in perception. The first one includes linguistic features of the interlocutor: speech activity, reaction time, lexical richness, sentiments (polarity and subjectivity). The second one includes facial features of the interlocutor: head movements, and facial expressions. The third modality includes a binary variable that represents the existence or not of the participant's saccades.

6. Conclusion

In this paper, we developed a framework for identifying causal relationships between brain activity and behavioral features for bidirectional non-controlled conversations. Evaluations are made on ROIs involved in speech perception and production, and face perception. The results show that the obtained predictions are significantly better than those obtained with the baseline model, and the obtained dependencies confirm hypothesis about the relationship between behaviors and the functional brain areas. Importantly, new dependencies are found for the STS area.

7. References

[1] Rauchbauer Birgit, Nazarian Bruno, Bourhis Morgane, Ochs Magalie, Prévot Laurent, and Chaminade Thierry, "Brain activity during reciprocal social interaction investigated using conversational robots as control condition," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 374, no. 1771, p. 20180033, Apr. 2019.

[2] K. Okada, J. H. Venezia, W. Matchin, K. Saberi, and G. Hickok, "An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex," *PLoS ONE*, vol. 8, no. 6, Jun. 2013.

[3] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting Human Brain Activity Associated with the Meanings of Nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, May 2008.

[4] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, pp. 453–458, Apr. 2016.

[5] T. Yarkoni, D. M. Barch, J. R. Gray, T. E. Conturo, and T. S. Braver, "BOLD Correlates of Trial-by-Trial Reaction Time Variability in Gray and White Matter: A Multi-Study fMRI Analysis," *PLOS ONE*, vol. 4, no. 1, p. e4257, Jan. 2009.

[6] N. Suzuki, Y. Takeuchi, K. Ishii, and M. Okada, "Effects of echoic mimicry using hummed sounds on human–computer interaction," *Speech Communication*, vol. 40, no. 4, pp. 559–573, 2003.

[7] H.-Y. Chen, Y.-H. Liao, H.-T. Jan, L.-W. Kuo, and C.-C. Lee, "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, pp. 5775–5779.

[8] D. Bone, C.-C. Lee, and S. Narayanan, "Robust Unsupervised Arousal Rating: A Rule-Based Framework with Knowledge-Inspired Vocal Features," *IEEE transactions on affective computing*, vol. 5, no. 2, pp. 201–213, Jun. 2014.

[9] A. Knops, B. Thirion, E. M. Hubbard, V. Michel, and S. Dehaene, "Recruitment of an Area Involved in Eye Movements During Mental Arithmetic," *Science*, vol. 324, no. 5934, pp. 1583–1585, Jun. 2009.

[10] J. F. DeSouza, S. Ovaysikia, and L. K. Pynn, "Correlating Behavioral Responses to fMRI Signals from Human Prefrontal Cortex: Examining Cognitive Processes Using Task Analysis," *Journal of Visualized Experiments : JoVE*, no. 64, Jun. 2012.

[11] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive Behavioural Systems*, ser. Lecture Notes in Computer Science, A. e. a. Esposito, Ed. Springer Berlin Heidelberg, 2012, pp. 114–130.

[12] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier, Apr. 2011.

[13] J. Ashburner, "A fast diffeomorphic image registration algorithm," *NeuroImage*, vol. 38, no. 1, pp. 95–113, Oct. 2007.

[14] S. Whitfield-Gabrieli and A. Nieto-Castanon, "Conn: A Functional Connectivity Toolbox for Correlated and Anticorrelated Brain Networks," *Brain Connectivity*, vol. 2, no. 3, pp. 125–141, May 2012.

[15] L. Kasper, S. Bollmann, A. O. Diaconescu, C. Hutton, J. Heinzle, S. Iglesias, T. U. Hauser, M. Sebold, Z.-M. Manjaly, K. P. Pruessmann, and K. E. Stephan, "The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data," *Journal of Neuroscience Methods*, vol. 276, pp. 56–72, 2017.

[16] L. Fan, H. Li, J. Zhuo, Y. Zhang, J. Wang, L. Chen, Z. Yang, C. Chu, S. Xie, A. R. Laird, P. T. Fox, S. B. Eickhoff, C. Yu, and T. Jiang, "The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture," *Cereb Cortex*, vol. 26, no. 8, pp. 3508–3526, 2016.

[17] B. Rauchbauer, Y. Hmamouche, B. Brigitte, L. Prévot, M. Ochs, and T. Chaminade, "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), 2020.

[18] D. Schiffrin, *Discourse markers*. Cambridge University Press, 1987, no. 5.

[19] C. Gössl, L. Fahrmeir, and D. Auer, "Bayesian modeling of the hemodynamic response function in bold fMRI," *NeuroImage*, vol. 14, no. 1, pp. 140–148, 2001.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[21] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[22] H. He, Y. Bai, E. A. Garcia, and S. Li, "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, pp. 1322–1328.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[24] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.