Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 668–675
Marseille, 11–16 May 2020

Multimodal corpus of bidirectional conversation of human-human and human-robot interaction during fMRI scanning

Birgit Rauchbauer [1,2,3], Youssef Hmamouche[2,4], Brigitte Bigi[2], Laurent Prévot[2], Magalie Ochs[5], Thierry Chaminade[1]

[1]Aix Marseille Université, CNRS, INT, Marseille, France
[2]Aix Marseille Université, CNRS, LPL, Aix-en-Provence, France
[3]Aix Marseille Université, CNRS, LNC, Marseille, France
[4]Aix Marseille Université, CNRS, LIS, Marseille, France
[5] Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
birgit.rauchbauer@univ-amu.fr; youssef.hmamouche@lis-lab.fr; laurent.prevot@univ-amu.fr; magalie.ochs@lis-lab.fr; thierry.chaminade@univ-amu.fr

Abstract

In this paper we present investigation of real-life, bi-directional conversations. We introduce the multimodal corpus derived from these natural conversations alternating between human-human and human-robot interactions. The human-robot interactions were used as a control condition for the social nature of the human-human conversations. The experimental set up consisted of conversations between the participant in a functional magnetic resonance imaging (fMRI) scanner and a human confederate or conversational robot outside the scanner room, connected via bidirectional audio and unidirectional videoconferencing (from the outside to inside the scanner). A cover story provided a framework for natural, real-life conversations about images of an advertisement campaign. During the conversations we collected a multimodal corpus for a comprehensive characterization of bi-directional conversations. In this paper we introduce this multimodal corpus which includes neural data from functional magnetic resonance imaging (fMRI), physiological data (blood flow pulse and respiration), transcribed conversational data, as well as face and eye-tracking recordings. Thus, we present a unique corpus to study human conversations including neural, physiological and behavioral data.

Keywords: conversation, physiology, artificial agents

1. Introduction

Reciprocal interactions with others allow humans to establish and maintain social bonds. We investigated human interaction using the most ubiquitous form of communication through language in *conversation.* Our objective was to characterize conversations through a multimodal approach for a comprehensive investigation of human interaction. As a control condition we used interaction with a conversational robot. This allowed us to change the social nature of the interlocutor, while preserving the reciprocal dynamics of social exchange. The social nature of the interaction with human derives from the assumption that humans adopt an intentional stance towards other humans, but not artificial agents, such as conversational robots. Adopting an intentional stance assumes that the interaction partner is capable of having mental states (Dennett, 1989). This allows the ascription of intentions and beliefs to explain the behavior of interaction partner (i.e., mentalizing (Frith & Frith, 1999)). We thus hypothesized that interaction with a human, as compared to a robot, would elicit both mentalizing mechanisms and the motivation to establish a relationship (i.e. social motivation).

To investigate this, we collected multimodal data including neuro-physiological responses of functional magnetic resonance imagining (fMRI), as well as physiological data and observable behaviors such as gaze, facial expressions and verbal productions during conversation. All these modalities have been shown sensitive to social and emotional engagement in interactions. We compared conversations between participants and another human (Human-Human Conversation, HHC) or a conversational robot (Human-Robot Conversation, HRC). This was done via life uni-directional videoconference and bi-directional audio connection between participants inside the scanner and a human or a conversational robot outside the scanner room.

Analyses of neural data collected with fMRI revealed hypothesized activated brain areas related to mentalizing and social motivation during HHC as compared to HRC (for details see (Rauchbauer et al., 2019)). Rauchbauer et al. (2019) thus introduced a novel and innovative paradigm to study real-life reciprocal interaction, using HRC as a control condition to study the social aspects of human interaction.

The aim of the present paper is to introduce the unique multimodal corpus of data collected in this study to a larger audience. The corpus comprises, apart from neural data, also conversational, physiological, eye-gaze and face-tracking data. This multimodal data is made available in data repositories mentioned in the present article. It presents a unique combination of variables involved in real-life, natural conversation. This data may be used to combine for example, conversational features with neural data to gain a deeper insight into which characteristics of a conversation activate brain areas related, for example, to social motivation. In making the corpus available online, we also encourage other researchers to use this unique and innovative corpus to characterize and study human interaction, as well as interaction with a conversational robot from a multimodal perspective. The corpus allows to answer various research questions and may allow to predict a modality of conversation through one or a combination of recorded modalities. In this paper we will introduce in detail the experimental design, data recorded.

2. Experimental design

We recorded data of twenty-five native French-speaking participants (17 female, mean age 26.76 (SD=7.96)). The participants engaged in a natural, life conversation alternating between another human (who, unknown to the participant, was working with the experimenter as a confederate) and a conversational, robotic head from Furhat robotics (https://www.furhatrobotics.com/; (Al Moubayed, Beskow, Skantze, & Granström, 2012)) during functional magnetic resonance imaging (fMRI).

Participants were welcomed to the study and told the “cover story” of the experiment, which provided a fake rationale for the experiment and its set-up. Participants were informed that the study was sponsored by an advertising company to test the key message of a new campaign. The message was to be discovered through a conversation with another participant and a conversational robot, who also had information on the campaign and was able to speak autonomously (see Figure 1 for experimental design, taken from (Rauchbauer et al., 2019)). Participants were informed about the study design, which presented images of anthropomorphized fruits and vegetables of the forthcoming advertisement campaign. These fruits and vegetables were designed to look like superheroes or were rotten (see Figure 2, taken from (Rauchbauer et al., 2019)). The participants were told that they could talk naturally about the presented images with the other agent (alternating between the human and the robot), who would be outside the fMRI scanner room and connected via life-video stream and bidirectional audio. They were informed that each conversation would last one minute after which a new image would be presented, and the conversation partner would change. The robot was a conversational robotic head, who, as participants were told, had information on the advertisement campaign and could talk autonomously. Unknown to the participants, the robot was controlled by the human confederate and the robot’s arguments were pre-written conversations based on a behavioral study (Chaminade, 2017). Participants were shown the conversational robotic head before being brought into the scanner room. At the end of the study, the participants were debriefed in an open format. Participants could voice their impression of the interaction with both the human and the robot. Also, it was verified that participants had believed in the autonomous speech of the robot and the cover story.

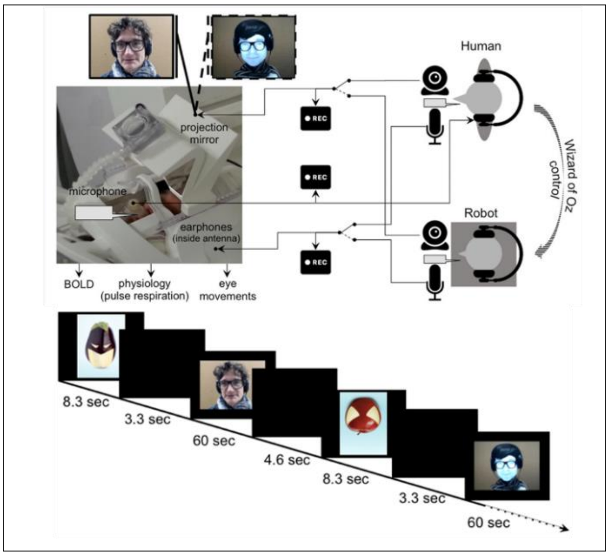

Figure 1: Experimental set-up. Panel above: Shown is the communication between the scanned participant and the other conversation agent, either the confederate or the robot, as well as the recording modalities; Panel below: the timeline of the experiment showing the alternation between the stimuli and conversation periods, as well as the relative timing. The fruit pictures correspond to the images used in the cover story; the robot and confederate pictures illustrates episodes of live conversations (Image taken from (Rauchbauer et al., 2019)).

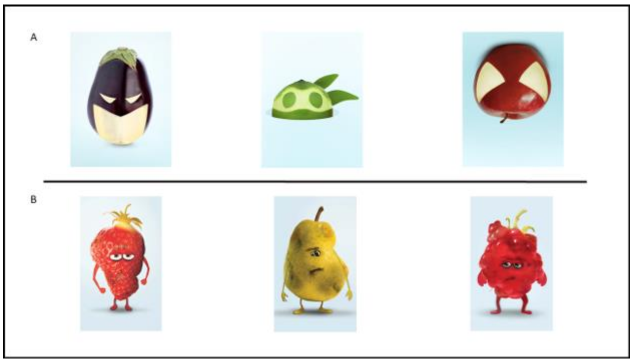

Figure 2: Images presented in experiment. Fruits and vegetables depicted as (A) superheroes (in Sessions 1 & 3) and (B) rotten (in sessions 2 & 4) (Image taken from (Rauchbauer et al., 2019)).

Interaction type	**Sessions 1 & 3** **Superheroes**	**Sessions 2 & 4** **Rotten**
Human-Human	Eggplant	Raspberry
Human-Robot	Apple	Pear
Human-Human	Lemon	Strawberry
Human-Robot	Eggplant	Raspberry
Human-Human	Apple	Pear
Human-Robot	Lemon	Strawberry

Table 1: Order of images, respectively conversations. Shown is the order of presented images and respective conversational topics.

2.1 Human and robotic conversational agent

The human interlocutor (i.e.; the confederate) and the participant were always gender matched. Participants were told that the human interlocutor had already participated in the scanning part of the experiment. The participant and the confederate met briefly before the experiment.

The conversational robotic head had a gender matched voice and face, which was retro-projected on a semi-transparent plastic mask, in the shape of a human face. The robot was developed by Furhat robotics (https://www.furhatrobotics.com/; (Al Moubayed, Beskow, Skantze, & Granström, 2012)) including options for autonomous or pre-written speech (using a Wizard of Oz (WOZ) procedure). For the experiment, as mentioned above, arguments were pre-written from human conversations recorded in a previous behavioral experiment using the same stimuli (Chaminade, 2017). This was done to highlight the artificial nature of the conversational robot, as compared to the social nature of the human. It also allowed the human confederate to control the statements of the robot in real-time from outside the scanner. This was done by using the WOZ user interface on a tablet connected to the robot's intranet, which showed buttons corresponding to the pre-programmed statements. Thirty statements were used for each of the six images (see Table 2 for examples). To illustrate, the button "pear" would launch statements such as "This is a pear", and the button "sad" would launch the statement "It looks sad" (all conversation in French). Feedbacks from the robot included general non-specific feedbacks ("yes", "no", "maybe") identical across images, and statements related to specific images, as the example to the pear shows. The presentation of the images and the task to identify an advertising message allowed to control for content of the conversation.

Bienvenue	***Welcome***
Bonjour	*Hello*
Salut	*Hi*
Comment ça va?	*How are you?*
Bon	*Good*
Merci	*Thank you*
Je m'appelle Furhat	*My name is Furhat*
Poire	***Pear***
C'est une poire	*It's a pear*
C'est une poire jaune	*It's a yellow pear*
La poire semble triste	*The pear seems sad*
Elle n'a pas l'air contente	*It does not look happy*
Elle semble malheureuse	*It seems unhappy*
La poire a l'air malade	*The pear looks sick*
Elle paraît faible.	*It seems weak.*
Elle semble fatiguée.	*It seems tired.*
La poire semble triste	*The pear seems sad*
Elle n'a pas l'air contente	*It does not look happy*
Elle semble malheureuse	*It seems unhappy*

Table 2: Examples of pre-written feedback for the conversational robot in the original French; English translation in italics;

Modality of data	**Extracted variables per conversation**
Neural data (fMRI)	Begin and end of conversations from logfiles
	BOLD signal for group and time series analysis
Speech and transcribed data	Number of IPUs and tokens
	Time series and mean duration of conversational features: IPUs, overlap of speech, reaction time, lexical richness, filled breaks, lexical feedback items, discourse markers, particle items, laughter, sentiment analysis (subjectivity and polarity)
Eye tracking data	Time series on fixation (face, mouth, eyes), saccades, speed, blinks
Face tracking data	Facial landmarks, Head Pose, Facial Action Units, Emotions
Physiological data	Blood flow pulse, respiration

Table 3: Multimodal data of the present study, presenting the modality of the data and the extracted variables per modality per conversation

The experimental set up for fMRI was a within-subject block design with the CONVERSATIONAL AGENT (Human or Robot) as the experimental factor. This established HHC and HRC. The experimental paradigm was set up in four sessions, consisting of six one-minute conversation per CONVERSATIONAL AGENT. Both the order of the stimuli and the agents were not randomized but kept constant across participants. In terms of the images the first and the third block always presented "superhero" images and second and the fourth block images of "rotten fruits" (see Figure 3, taken from (Rauchbauer et al., 2019)); each of the images was presented twice in each block, once for the conversation with the human and once with the robotic agent. The CONVERSATIONAL AGENTS alternated, starting with the human confederate.

2.2 Detailed protocol and data recording

Images were presented for 8.3 seconds (see Figure 1)). After a 3.3-second-long black screen, live conversation with bi-directional audio and uni-directional live video from outside the scanner started. Due to technical constraints of the fMRI machine no video was possible from inside the scanner to outside, but data of participants' eye movements was collected. After the one-minute conversation with one of the conversational agents, a black screen was presented for 4.6 seconds. The conversation was always started off by the participant in the scanner. We recorded three minutes of conversation per conversational agent in each block (6 minutes total in each block). Thus, for each participant twenty-four minutes of conversation were recorded. The bidirectional audio set-up enabling the live conversation between the scanner and the outside consisted of an active noise-cancelling MR compatible microphone (FORMI-III+, optoacoustics), to cancel out scanner noise and insert earphones from Sensimetrics. Webcams outside the scanner room allowed for uni-directional video, which was projected onto a mirror on the antenna of the participant in the scanner. Participants' direction of gaze on the projection mirror was recorded (Eyelink 1000 system, SR Research). The experimental paradigm and data recording were controlled via Labview (National Instrument). In addition to fMRI data, audio, eye movements, respiration and blood flow pulse of the participants, and video and audio data from the confederate (human and robotic) was recorded. Transcribed data of the conversation is available on Ortolang

(https://www.ortolang.fr)
(https://hdl.handle.net/11403/convers

2.2.1 fMRI data

For details on the standard fMRI data collection and the preprocessing please refer to Rauchbauer et al. (2019). Data acquisition of Blood Oxygen-Level Dependent (BOLD) signal 3-dimensional images were obtained via whole brain scans every 1.205 seconds. fMRI preprocessing was carried out using SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). This included correcting for time delays due to scanning the whole brain in consecutive slices ("slice timing"), realignment of images, corrections of magnetic field inhomogeneities, and normalization of the participant's individual function and anatomical data to standard brains from the Montreal Neurological Institute (MNI) with DARTEL (https://neurometrika.org/node/34) (Ashburner, 2007) and spatial smoothing using a 5-mm full-width half-maximum 3-dimensional Gaussian kernel. We extracted BOLD signal of regions of interest using the conn toolbox (https://web.conn-toolbox.org/) (Whitfield-Gabrieli & Nieto-Castanon, 2012). This includes denoising procedures such as linear detrending applying a high-pass filter (threshold of 128 seconds), the use of realignment parameters for the calculation of nuisance regressors due to participants' movement inside the scanner and the removal of physiological artifacts (blood flow pulse and respiration) using the PhysIO toolbox (https://www.tnu.ethz.ch/de/software/tapas/documentations/physio-toolbox.html) (Kasper et al., 2017). This also comprises the extraction of the BOLD signal in white matter and cerebrospinal fluid and, in non-cortical brain tissue, the use of the 5 first eigenvariates of the time-series as nuisance, which represents signal fluctuations. Logfiles of the BOLD signal acquisition, containing information on parameters of the scanning session are in JavaScript Object Notation (.json). The logfiles in text format containing the onset of conditions for fMRI analyses have been uploaded on OpenNeuro (Gorgolewski, Esteban, Schaefer, Wandell, & Poldrack, 2017) (https://openneuro.org/): https://openneuro.org/datasets/ds001740 including all raw data acquired during fMRI scanning. fMRI group data can be found on Neurovault (https://neurovault.org/): /collections/ASGXRWEM/.

2.2.2 Speech and Transcribed data

Conversational data was preprocessed to extract noise from the conversation in the scanner (for participant data) using a noise reduction filter by Sox (http://sox.sourceforge.net). For this a float value was set individually for each participant between 0.01 and 0.5. Denoised data was furthermore pre-segmented into Inter-Pausal Units (IPUs), also using a coefficient set individually for each participant, with a float value between 0.2 and 0.95. This coefficient allowed the automatic determination of volume threshold for periods of "silence" and "IPU" in each analysis window and applies to the mean of the distribution of the root mean square (RMS) of the audio file. IPUs were defined as blocks of speech in between silences of minimum duration 200 ms (Blache, Bertrand, & Ferré, 2009). IPUs were also extracted for the confederates' audio files, yet no denoising was necessary, since these were recorded outside of the scanner room. Visual inspection of successful denoising and segmentation into IPUs was done using Praat (Boersma, 2002). Files were furthermore uploaded into SPPAS, version 1.9.9 (www.sppas.org/;(Bigi, 2015)) for transcription. Transcribed files of the participants' and the two interlocutors speech have been deposited on the data repository Ortolang (https://www.ortolang.fr) (https://hdl.handle.net/11403/convers) (see examples 1 and 2 for excerpt of transcription).

Automatic Text normalization was performed using the SPPAS software tool (Bigi, 2011). From the normalized data the number of IPUs and tokens in the conversation were extracted (see Figure 3). Figure 3 shows that the number of IPUs (t = 24.461, p ≤ .001) and tokens (t = 29.386, p ≤ .001) differed significantly between the human and the robot confederate. The number of tokens also differed between the HHC and HRC for the participant (t = 5.858, p ≤ .001). Furthermore, we also calculated the time series of conversational features as well as their mean duration from the participants' and the interlocutors' speech separately using a Python Script. Extracted conversational features were IPUs, overlap of speech between the two interlocutors (i.e.; both speakers speaking at the same time) and reaction time of the start of conversation after the partner had finished his speech turn. Reaction times have positive values in case of a pause between turn taking, and negative values in case of overlapping speech of interlocutors. Lexical richness was computed considering the amount of spoken tokens and adjectives plus adverbs divided by a total number of extracted tokens (including adjectives, adverbs, auxiliary words, conjunction, determiners, nouns, prepositions, pronouns, verbs) (Ochs, Jain, & Blache, 2018). Furthermore we considered filled breaks (i.e., utterances like "mmh" during pauses of active speech) (Swerts, 1998) and lexical feedback items (Prévot, Bigi, & Bertrand, 2013; Prévot, Gorish, & Mukherjee, 2015), representing expressions to communicate perception and understanding, as well as reactions to what the conversational partner had said (E.g.; "yes", "no", "okay", etc.) (Gravano, Hirschberg, & Beňuš, 2011). Additionally, we extracted discourse markers, used to organize the ongoing discourse, such as "so" or "therefore" (Schiffrin, 1987), particle items, which express the speaker's mood (Barnes, 1995) and also laughter (Ellis, 1997). We further calculated subjectivity and polarity of speech with the Pattern library (Smedt & Daelemans, 2012). This sentiment analysis is an automated process that extracts positive and negative feelings, emotions and opinions from speech. Polarity captures the expression of positive and negative feelings or opinions, such as anger (negative feeling) or happiness (positive feeling). Values range from 1 (expression of positive feeling) to -1 (expression of negative feeling). Subjectivity refers to the expression of subjective (as opposed to objective) content with scores between 0 (for objective content) and 1 (for subjective or personal content). An example for an objective opinion would be "The strawberry is red", whereas a subjective opinion would be "The strawberry is ugly".

Analysis of conversational features was carried out per conversational agent and participant in the separate conditions (Human-Human Conversation participant (HHC_part), Human-Human Conversation confederate (HHC_conf), Human-Robot Conversation participant (HRC_part), Human-Robot Conversation robot

(HHC_robot)). This conversational data will be used jointly with the fMRI, physiological and behavioral data to characterize the multimodality of real-life conversations. As such, it could be hypothesized that certain conversational features, such as laughter or discourse markers, could predict activation in brain areas related to social motivation during HHC. This extensive corpus allows to investigate such questions.

IPU 1: 1.15 - 3.95

bon franchement franchement t'as déjà cueilli des fruits quand t'étais petit
well frankly frankly did you already pick fruits when you were a kid

IPU_1: 4.28 - 5.61
j'en j'en ai dans mon jardin hein j'ai le cerisier on est train de faire la récolte de cerises en ce moment
I I have some in my garden, huh, I have the cherry tree we are doing the cherry harvest right now

IPU 2: 8.23 - 9.65
ah ouais en plus en plus c'est la saison
oh yep moreover moreover it is the season

IPU_2: 9.95 - 10.17
ouais
yep

IPU_3: 10.06 - 11.79
c'est la saison elles sont elles sont bonnes en ce moment
this is the season they are they are good right now

IPU_ 3: 11.92 - 14.83
elles sont elles sont petites elles sont brillantes comme la pomme
they are they are small they are bright like the apple

IPU_4: 15.28 - 18.07
ouais ah et toi donc toi t'as des très très beaux fruits
yep ah and you so you have very very beautiful fruits

Example 1: Excerpt and visualization of transcribed human-human conversation. Original merged text file from *Praat* was simplified for visualization purposes; shown are Inter-Pausal Units (IPUs) with timestamps. Transcribed speech by the participant is presented in green, and by the confederate in blue; English translation is presented right below the corresponding statements in italics;

IPU_1: 1.07 - 3.52
là encore une fois il s'agit de l'aubergine
here again it's about the eggplant

IPU_2: 3.84 - 4.71
euh
hum

IPU_1: 4.28 - 5.61
elle ressemble à Batman
it looks like Batman

IPU_3: 5.83 - 7.53
oui en effet elle ressemble à Batman
yes indeed it looks like Batman

IPU_4: 7.84 - 12.97
ce que j'avais pas remarqué précédemment c'est que le pédoncule au dessus de l'aubergine
what I had not noticed before is the peduncle on top of the eggplant

IPU_5: 13.20 - 17.41
lui donne comme une coupe de cheveux assez étrange qui ressemble pas à Batman
gives it like a weird haircut that does not look like Batman

IPU_2: 16.21 - 16.65
Oui
yes

IPU_3: 16.88 - 17.57
tu as raison
you are right

Example 2: Excerpt and visualization of transcribed human-robot conversation. Original merged text file from *Praat* was simplified for visualization purposes; shown are Inter-Pausal Units (IPUs) with timestamps. Transcribed speech by the participant are presented in green, and by the conversational robot in blue; English translation is presented right below the corresponding statements in italics;

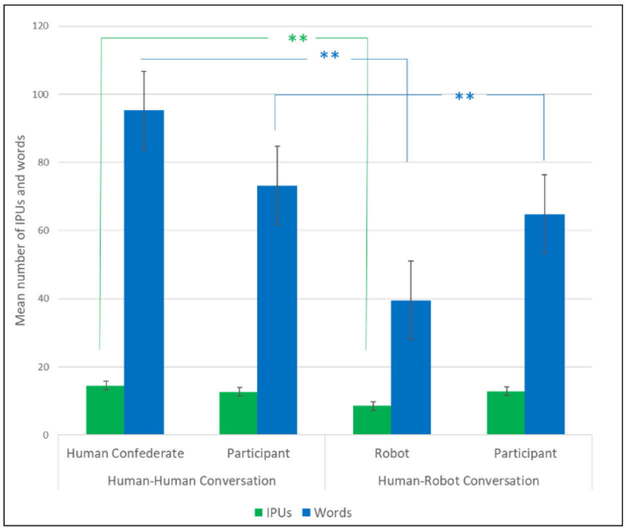

Figure 3: Mean number of Intra-Pausal Units (IPUs) and tokens. Figure displays the mean number of IPUs and tokens per Human-Human Conversations for the human confederate and the participant and per Human-Robot Conversation for the conversational robot and the participant; Error bars denote standard deviation of the mean; ** p ≤ .001

2.2.3 Face and eye-tracking data

Eye-tracking data was collected from the participant while lying supine in the scanner, synchronized with fMRI data collection with an Eyelink 1000 Plus Long Range Mount (SR Research Ltd., Mississauga, Ontario, Canada, https://www.sr-research.com/products/eyelink-1000-plus/) with a temporal resolution of 1000 Hz. The eye position of the left eye was recorded using pupil and corneal tracking. 5-point gaze-calibration was conducted at the beginning of the experiment, validated, and, if necessary, recalibrated before each run. Eye tracking raw data consists of files in European data format (.edf). These files were transformed into an American Standard Code for Information Interchange format (.ascii) using C. The files contain messages indicating the fixation, saccades and blinks along a standard coordinate system (x, y, z), as well as information on start and end time of the conversation. Raw data was processed using python for synchronization with video data frequency and for separation of speed, saccades, fixation and blinks per conversation. The coordinate system for Eyelink, as for OpenFace (see below) is based on the standard image coordinate system, defined by pixel location.

Videos recorded from the human confederate and the robot were used for face-tracking analyses, recording frequency was 30Hz. We used OpenFace (https://cmusatyalab.github.io/openface/; (Amos, Ludwiczuk, & Satyanarayanan, 2016)) to analyze each video separately (see for example Figure 4). The output format of OpenFace is a .csv (comma separated value) containing 1800 observations which is equivalent to the number of images in the video (videos of conversation of 60 sec; 30 images per second (i.e., 30 Hz)). The .csv output file contains the 68 facial landmarks, 17 FAUs, 3 features of gaze movements and 6 features of head pose rotations and translations, which were extracted using pre-trained models from OpenFace. Facial landmarks represent important facial regions, such as eyes, nose, mouth, jaw, eyebrows and face outline using shape prediction models. For this the face is localized in an image and salient facial structures detected. Pre-trained facial landmark detectors estimate the location of facial structures on a coordinate system (x,y). Detection of gaze movements show, along coordinates, where the confederate is looking. This is also done by applying pre-trained models to the images (Wood et al., 2015). This includes description of gaze angles in a change of looking from, for example, left to right, with an angle of 0 if the person is looking straight (https://github.com/TadasBaltrusaitis/OpenFace/wiki/Output-Format). The head pose measurements describe the head location in relation to the camera. Rotation is described in pitch, roll and yawn detection along coordinates (x, y, z). Pitch is rotation movement around the X-axis, describing up and downward head movement. Roll describes a rotation around the Z-axis, which is a tilting movement of the head. Yaw is a rotation of the head around the Y-axis, describing a right and left head movement. Translation is a head movement along the X, Y and Z axis. The Facial Action Coding System (FACS) describes facial movements. It encodes facial muscular movements upon appearance and deconstructs it into FAUs producing the expression. FAUs can represent the presence and intensity of facial muscular movements. Data will be uploaded on Openneuro (Gorgolewski et al., 2017) (https://openneuro.org/).

Incorporating face and eye-tracking data into multimodal analysis of conversation will give insights, for example, into the interplay between conversational features and facial expressions. Using the example from above, combining face tracking data and conversational features to predict activation in brain areas related to social motivation, may offer a comprehensive picture into how human interactions are shaped.

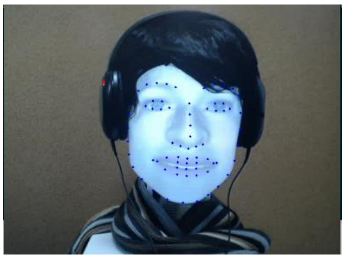

Figure 4: Example of face tracking data analysis of robot face using OpenFace.

2.2.4 Physiological data

Physiological data was recorded with the SIEMENS scanner's own system. A photoplethysmography unit was positioned on the left-hand index fingertip to record pulse oximetry and a breathing belt was positioned at the chest level. Both were connected wirelessly through Bluetooth. Data was acquired continuously at the frequency of 200Hz. Data format being proprietary, a preprocessing is needed using a specific toolbox (PhysIO toolbox; https://www.tnu.ethz.ch/de/software/tapas/documentations/physio-toolbox.html)(Kasper et al., 2017). The toolbox reads the data and synchronizes it with the acquisition. For the corpus, the output of this processing is saved as a matrix in a comma-separated value file (named subj01_sess01_physio.csv for session 1 of subject 1) which has three columns corresponding, respectively, to the time stamps of the observation, the cardiac signal and the respiration signal. With a session duration of 462.73 seconds and 200Hz recordings, a maximum of 92546 time points can be recorded, yet some are missing. Figure 5 indicates that for most participants, less than 1‰ of the data is missing, three outliers have around 3.5‰ missing data and one outlier (subj25) has 1.29% of missing data. It is therefore important to take into account the time stamps to keep the synchrony between the time series.

Physiological data is very noisy, in particular because of the high magnetic field and because participants were prone to move peripheral limbs while discussing. The preprocessing toolbox can process the data further, including automatic enhanced peak detection for the blood pulse signal and discarding of erroneous recordings for the respiration (detached or saturated breathing belt for example). These steps are used to output nuisance covariates for the fMRI analysis such as cardiac and respiratory phases (described in section fMRI data), but as these can be recalculated from the raw data and are only meaningful in the context of fMRI analysis, they haven't been included in the currently shared corpus. The raw physiological data will be uploaded on Openneuro (Gorgolewski et al., 2017) (https://openneuro.org/). Apart from acting as nuisance regressors to denoise fMRI data, physiological data can also be informative in the context of social dimensions of emotions (e.g., Britton, Taylor, Berridge, Mikels, & Liberzon, 2006) during a conversation. This illustrates that adding physiological data in the multimodal analysis of conversation can give a comprehensive view on social and emotional aspects of human interaction.

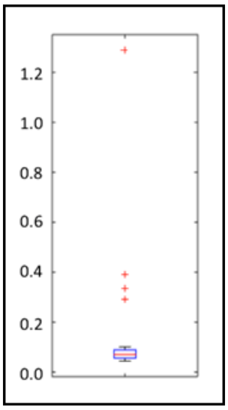

Figure 5: Boxplot showing percentage of missing data (number of missing points divided by maximum number of data points in %) in our corpus. Note one striking outlier.

3. Discussion

This project aims to investigate the specificities of human interactions through conversations by using a human-robot interaction as a control condition. The HRC allows to keep the reciprocity of interaction intact, while changing the social motivation and adoption of an intentional stance in the interacting human. In a first study we showed that brain regions implicated in social motivation are involved in HHC. Yet, to characterize human interaction further, in the present paper we introduce the comprehensive corpus collected in the project. This multimodal corpus of natural conversations, including neural, physiological, behavioral and transcribed conversational data allows to combine these different modalities for a full picture of human conversation.

With an innovative experimental set-up, we created a bi-directional conversation between a human participant in a fMRI scanner and a human confederate or a conversational robot outside of the scanner. This allowed us to compare the social nature of a human interaction, characterized by the adoption of an intentional stance and social motivation compared to an interaction with a robot (non-social nature). A cover story framing the study as a neuromarketing experiment with the goal to extract the message behind a new advertisement campaign provided the topic of conversation.

fMRI data was recorded from the scanned participant to identify objective neural responses during conversation with another human or a conversational robot (for details see (Rauchbauer et al., 2019)). Conversational data of the interlocutors has been transcribed and conversational features, as well as time-series and descriptive statistics of IPUs, filled breaks, feedbacks, discourse markers, particles and laughter extracted to characterize the difference of conversations between two humans and a human and an artificial agent and to align conversational with neural data. Furthermore, the multimodal corpus included physiological (blood flow pulse and respiration), as well as eye-gaze (from participant) and face-tracking data (from the confederates) to comprehensively describe natural conversation. Data analysis of this corpus aims to integrate the multimodal markers of conversation to comprehensively characterize human conversation and conversation with an artificial agent. Thus, the presented corpus allows, for the first time, to include multimodal conversational data in the investigation of bidirectional interaction.

4. Conclusion

We investigated natural bi-directional conversations between two humans and a human and a conversational robot. In this innovative study we collected a multimodal corpus of neural, physiological, eye- and face-tracking and transcribed conversational data. We present a unique approach to studying real-life conversations from a multimodal perspective. This allows comprehensive investigation of human conversation.

5. Acknowledgements

Research supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and AAP-ID-17-46-170301-11.1 by the Excellence Initiative of Aix-Marseille University (A*MIDEX), a French "Investissement d'Avenir" programme and the Institute for Language, Communication and the brain (ILCB ANR-16-CONV-0002). BR is supported by the Fondation pour la Recherche Médicale (FRM, SPF20171039127).

6. Bibliographical References

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

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science, 6*.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage*, *38*(1), 95–113.

Barnes, B. K. (1995). Discourse particles in French conversation:(eh) ben, bon, and enfin. *French Review*, 813–821.

Bigi, B. (2011). A multilingual text normalization approach. In *Language and Technology Conference* (pp. 515–526). Springer.

Bigi, B. (2015). SPPAS-multi-lingual approaches to the automatic annotation of speech. *The Phonetician-International Society of Phonetic Sciences*, (111–112), 54–69.

Blache, P., Bertrand, R., & Ferré, G. (2009). Creating and exploiting multimodal annotated corpora: the ToMA project. In *Multimodal corpora* (pp. 38–53).

Springer.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*.

Britton, J. C., Taylor, S. F., Berridge, K. C., Mikels, J. A., & Liberzon, I. (2006). Differential subjective and psychophysiological responses to socially and nonsocially generated emotional stimuli. *Emotion*, *6*(1), 150.

Chaminade, T. (2017). An experimental approach to study the physiology of natural social interactions. *Interaction Studies*, *18*(2).

Dennett, D. C. (1989). *The intentional stance*. Cambridge, MA, USA: MIT press.

Ellis, Y. (1997). Laughing together: Laughter as a feature of affiliation in French conversation. *Journal of French Language Studies*, *7*(2), 147–161.

Frith, C. D., & Frith, U. (1999). Interacting minds--a biological basis. *Science*, *286*(5445), 1692–1695.

Gorgolewski, K., Esteban, O., Schaefer, G., Wandell, B., & Poldrack, R. (2017). OpenNeuro—a free online platform for sharing and analysis of neuroimaging data. *Organization for Human Brain Mapping. Vancouver, Canada*, *1677*(2).

Gravano, A., Hirschberg, J., & Beňuš, Š. (2011). Affirmative cue words in task-oriented dialogue. *Computational Linguistics*, *38*(1), 1–39.

Kasper, L., Bollmann, S., Diaconescu, A. O., Hutton, C., Heinzle, J., Iglesias, S., … Stephan, K. E. (2017). The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. *Journal of Neuroscience Methods*, *276*, 56–72. https://doi.org/https://doi.org/10.1016/j.jneumeth.2016.10.019

Ochs, M., Jain, S., & Blache, P. (2018). Toward an automatic prediction of the sense of presence in virtual reality environment. In *Proceedings of the 6th International Conference on Human-Agent Interaction* (pp. 161–166). ACM.

Prévot, L., Bigi, B., & Bertrand, R. (2013). A quantitative view of feedback lexical markers in conversational French. In *Proceedings of the SIGDIAL 2013 Conference* (pp. 87–91).

Prévot, L., Gorish, J., & Mukherjee, S. (2015). Annotation and classification of french feedback communicative functions. In *The 29th Pacific Asia Conference on Language, Information and Computation*.

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

Schiffrin, D. (1987). *Discourse markers*. Cambridge University Press.

Smedt, T. De, & Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, *13*(Jun), 2063–2067.

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, *30*(4), 485–496.

Whitfield-Gabrieli, S., & Nieto-Castanon, A. (2012). Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connectivity*, *2*(3), 125–141.

Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., & Bulling, A. (2015). Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3756–3764).

7. Language Resource References

Bigi B., (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech**.** In "the Phonetician" - International Society of Phonetic Sciences, ISSN 0741-6164, Number 111-112 / 2015-I-II, pages 54-69, retrieved from http://www.sppas.org/

Boersma, P. & Weenink, D. (2018): Praat: doing phonetics by computer [Computer program].Version 6.0.37, retrieved 14 March 2018 from http://www.praat.org/

Open Resources and Tools for LANGuage (Ortolang) (2015): Équipe ORTOLANG, 7.1.1 , ORTOLANG ISSN 2417–7482; https://www.ortolang.fr;

Gorgolewski, K., Esteban, O., Schaefer, G., Wandell, B., & Poldrack, R. (2017). OpenNeuro—a free online platform for sharing and analysis of neuroimaging data. Organization for Human Brain Mapping. Vancouver, Canada, 1677(2). https://openneuro.org/