Challenges in Linking Physiological Measures and Linguistic Productions in Conversations

Thierry Chaminade[1], **Laurent Prvot**[2,4], **Magalie Ochs**[3], **Birgit Rauchbauer**[1,2,5], **Nol Nguyen**[2]
[1]Aix Marseille Universit , CNRS, INT, Marseille, France
[2]Aix Marseille Universit , CNRS, LPL, Aix-en-Provence, France
[3]Aix Marseille Universit , CNRS, LIS, Marseille, France
[4]Institut Universitaire de France, Paris, France
[5]Aix Marseille Universit , CNRS, LNC, Marseille, France
firstname.lastname@univ-amu.fr

Abstract

We introduce here a new experimental set-up that provides temporally aligned linguistic and behavioral data together with physiological activity time-series recorded during social interactions. It brings the experimental approach closer to ecological social interaction. Such endeavour requires the aggregation of linguistic, physiological and neuro-cognitive information. Compared to measurement of activity grounded on existing linguistic material our setting presents some additional challenges as we are dealing with conversations. In addition to present the rationale, set-up and preliminary analyses, we discuss (i) the challenges caused by the spontaneous and interactional nature of the activity recorded ; (ii) the problem of balancing experimental set-up between the technical needs and the desire to keep some level of naturalness in the task ; and (iii) the difficulties in relating in a temporal way linguistic events with physiological signals that have their own biological dynamics.

Keywords: conversation, physiology, artificial agents

1. Introduction

We consider that *conversations* constitute a privileged framework to study social interactions. Our objective is to approach these highly sophisticated linguistic and social structures by scrutinizing neuro-physiologival responses of the participants as well as their observable behaviors such as gaze, facial expressions and verbal productions. Our ultimate goal is to characterize the participants' brain activity by means of fMRI but because of the huge challenges involved in using fMRI in conversational interactions, we started out with a simpler setting that featured some of the challenges, in particular in terms of data analysis. More precisely we recorded computer-mediated "skype like" conversations and we tracked a set of physiological parameters during these conversations : gaze (eye-tracking) and electrodermal activity. Electrodermal activity, as a physiological response, has a specific dynamic that needs to be handled while temporally relating the measures to the actual linguistic production in the corpus (Chaminade, 2017). In the study, we were interested in comparing different communication situation, in particular subjects were interacting either with another human or with an artificial agent. The significant differences we found between the two conditions is a interesting step showing the relevance of our measures and analyses.

2. Experimental set-up

A cover story provided a topic for the discussion as well as a common goal for the two interacting agents. The cover story consisted in presenting the experiment as a neuromarketing experiment, in which the pair of participants would discuss together through a videoconferencing system the message of a forthcoming advertising campaigns. In each campaign, the tested participant is presented with three images without text and then instructed to discuss it with either a natural (fellow human) or an artificial (embodied conversational agent or anthropomorphic robot) agent.

2.1. Physiological pilot set-up

Experimental conditions were defined by a 2 by 2 factorial plan. The first factor was the nature of the Agent the participant discussed with, a Human or an Artificial agent presented as autonomous; the second factor was the nature of the Interaction, either Live, through videoconferencing, or Video, using recordings of previous Live interactions as stimuli. The four conditions were therefore Human/Live, Artificial/Live, Human/Video, Artificial/Video. The virtual agent GRETA (Bevacqua et al., 2010; Pelachaud, 2015) used for the behavioral part of the project was used in a Wizard of Oz (WOZ) setting.

There was a room for the participant and another room for the human agent. Headphones were used so that the speech from both participants were acquired separately. In the Participant room, the recorded participant sat in front of a computer screen topped by the webcam and included microphone used for the Skype discussion. The Control computer was connected to the screen and the webcam and the participants headphones, which also controlled GRETA and WoZ. The eye-tracker cameras were below this screen. The left hand of the participant was fitted with a blood pulse sensor and two electrodes to record the electrodermal activity measurement guidelines (Roth et al., 2012), connected to Biograph box. The second room comprised a computer controlled by the Control computer, and was connected to the discussant screen, webcam and headphones.

2.2. Neurophysiological Experimental set-up

The participant is in the MRI scanner with earphones while the human agent is in another room with headphones, both facing a screen. Several aspects of behavior and physiological responses of this participant are recorded as continuous time series to form the corpus. Speech production, eye movements and skin conductance of the scanned participant are recorded with MRI-compatible devices. Recording of eye movement offers, in addition to eye-tracking data set, a live video output of the eyes of the scanned participant for the interlocutor. The human interlocutor is recorded by a webcam with incorporated microphone. The artificial agent is the conversational robotic platform Furhat. 12 repetitions of the 60-second discussions with the human interlocutor and each version of the artificial agents are recorded for each participant using the fact that there are several images in the advertising campaign that must be discussed separately. Each repetition consists in the presentation of one image for 10 seconds to the participant, followed by a 3-second black screen, and then sixty seconds during which the participant talks with the interlocutor, either a human participant or the artificial agent controlled by a WOZ. Around 20 participants are used for the analysis of an fMRI experiment, yielding 12 minutes of discussion per subject per agent, for an expected total of 4 hours per agent.

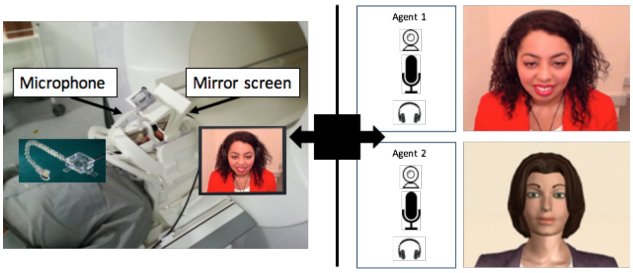

Figure 1: fMRI Experimental set-up

2.2.1. Equipment

fMRI requires dedicated MRI-compatible equipment: MR-compatible visual stimulation, eye-tracker, earphones and physiological response recorders (blood oxygenation and skin conductance, based on Siemens technology made available with the MRI scanner) (see Figure 1). An fMRI-compatible microphone with online de-noising has been acquired for completing the set-up, allowing for real-time discussion. The MRI center research ingeneer developed an interface to synchronize all this data. Preliminary analysis of the resulting speech recorded is very promising, with ASR system able to recognize significant parts of the input and no problem for manual transcription.

2.2.2. Remaining difficulties

There are two anticipated difficulties that call for specific attention. First, head movements interfere with analysis of the neurophysiological data, in contrast to the behavioral setup, where they constitute a very important variable to investigate interindividual coordination. As long as speech is concerned, the literature confirms that as only the lower jaw and vocal tract are concerned with vocalization, simple speech is compatible with fMRI acquisition provided that the rest of the head is held firmly to avoid movement. Inflatable cushions positioned on both sides of the head between the parietotemporal part of the skull and the MRI antenna (where the head rests) provide firm yet comfortable stabilization of the head.

The second difficulty is intractable and will be considered as a limitation of the study: speaking while lying supine and standing still in a noisy MRI scanner cant be considered as a natural way of speaking, even less as a natural social interaction. The use of live conversations with a human or an artificial agent will anyway provide a sufficient contrast in terms of social interaction to find differences hypothesized from the findings from more classical paradigms using simple perception or fake interactions. Surface recordings (EEG and MEG), while keeping freedom of movements and being more natural, don't allow recording the activity of deep brain structures or some cortical areas (e.g. depth of sulci), and spatial resolution is lower. In addition, muscles movement involved in speech present different challenges as they cause artifacts in these recordings.

3. Data sets, physiological pilot

The objective of the research is to characterize behavioural events that are temporally associated with physiological events. Preprocessing includes the precise synchronization of the behavioural and physiological time-series acquired independently, and the extraction of meaningful events from the time-series.

3.1. Electrodermal Activity

The example of the analysis of the electrodermal activity acquired in the behavioural pilot is used to illustrate the proposed approach more generally for the analysis of all physiological data. Using a Matlab toolbox for the analysis of electrodermal activity data (Ledalab (Benedek and Kaernbach, 2010)), the raw electrodermal recording is decomposed into phasic components and tonic responses. Tonic responses are deconvoluted in order to identify the timing and the intensity of the events responsible for each of the responses identified within the one-minute recording of each condition. The timing of the tonic electrodermal responses events is used to reconstruct a 30Hz time-series with delta functions indicating events onset (time-series [isElectrodermalEvent]).

3.2. Gaze Tracking

Eye tracking was recorded using standard procedures with from FaceLab5 from Seeing Machines. This system does not require physical constraint so that the participants remained free of their movements. Screen x and y voxel coordinates of the direction of the gaze and of the face on the screen were extracted. Eye closure and saccades were also extracted for filtering out unusable data. Time-series were downsampled from 60 to 30 Hz by decimation (removing one every other time point) to match the rest of the recorded times series. Moreover, video data was analyzed to extract facial features for each frame. A face recognition algorithm (Facial Feature Detection & Tracking; (Xiong and De la Torre, 2013)) was run frame by frame to identify the face present in the image. Screen x and y pixel coordinates of 49 keypoints on the face were extracted as well as the position and rotation of the face mask in relation to the screen normal vector. Face tracking results were combined with gaze tracking data to provide binary 30 Hz time series indicating gaze information for each frame. First, using face tracking coordinates, the position of the face, the eyes and the mouth on the screen were calculated for each frame and used to define regions of interest. Then, using gaze tracking coordinates, 30 Hz binary time series were created indicating whether or not the gaze was within these regions of interest (is the gaze on the screen [isData], is the gaze on the face [isFace], is the gaze on the eyes [isEyes], is the gaze on the mouth [isMouth]).

3.3. Linguistic features extraction

The audio files were manually transcribed and forced-aligned at the token level with SPPAS (Bigi and Hirst, 2012). This alignment allowed us to produce time series based on IPUs (stretches of speech separated by pauses of a given duration threshold, here we used 100, 200, 400 and 800ms) and tokens. These identified IPUs were used to construct three time series describing which agent is speaking [isParticipantSpeak], [isDiscussantSpeak], [isSilent].

Transcription had been realized in standard orthographic conventions without omitting truncated words, backchannels and other spontaneous speech phenomena such as disfluencies. We also determined whether the IPU is a feedback behavior and whether it hosts disfluencies, based on the transcription content.

4. Analysis

As explained above, multiple 30Hz time series were produced during preprocessing. Physiological events are considered as temporally associated with behavioural events, from which psychophysiological co-occurences will be investigated (Bach and Friston, 2013). A probabilistic approach was chosen under the assumption that it is adapted to the ecological type of relationships expected here, which are multidimensional (speech, face and eye movements, physiology) as well as noisy given the ecological design. The probability of a given behavior, the probability of having an electrodermal response, and the probability of having both the behavior and the electrodermal response was calculated in time windows of 100, 167, 200, 333 and 500ms. The posterior probability of certain behaviours (e.g. looking at the mouth of the interlocutor) giving rise to an electrodermal response was performed with a direct application of Bayes theorem. It is particularly well suited here to take into account that events were not controlled in terms of their probability and temporal distribution given the unconstrained nature of the conversational interaction.

Here, we present the analysis of linguistic and gaze behaviors in relation with electrodermal responses. Given the deconvolution of electrodermal activity and the physiological delays, synchronicity at the frequency used for data preprocessing (30 Hz, meaning co-occurrence of events within 33ms windows) is unrealistic. An exploratory approach was adopted, choosing time windows of 100, 167, 200, 333 and 500ms to calculate co-occurences. The effect of the two experimental factors on the posterior probabilities were assessed with linear statistics (ANOVA). Figure 2 presents the effect of these factors on the posterior probabilities associated with different behaviours. For example, panel 1 indicates how the posterior probability of observing a electrodermal response when the participant listens to the other agent is affected by the Agent, the Interaction, and the interaction between Nature and Interaction. While the results were quite consistent across the sizes of time windows, 200 ms always provided, when significant, the most significant effect. It is interesting to compare this to the conclusion of (Laming, 1968) that states that a simple reaction time to a visual stimulus, when no other task is required, is around 220ms. Significant effects of the agent are presented in figure 3: both when the eyes and the mouth are being watched, the posterior probability of co-occurrence of the behavioral and physiological event is higher for the Human compared to the artificial agent. In other words, natural behaviours in a conversation, such as looking at the eyes or the mouth of the discussant, is more likely when the other agent is human compared to artificial.

Finally, we investigated the differences in terms of interactivity across the different conditions from a linguistic perspective. The quantification is currently based on the ratio of IPUs directly involving *feedback* compared to the total number of IPUs in the interaction. The different conditions are leading to significantly different ratios in the expected directions, i.e. the more natural the interaction, the more feedback related IPUs are produced proportionally. More precisely, the nature of the agent brings significant differences for both live (p=0.004) and video (p=0.02) conditions, while the nature of the interaction shows significant differences for the virtual agent (p=0.04) and for human agent (p=0.04).

4.1. Neurophysiological-Linguistic data sets

The fMRI corpus will be investigated using classical approaches relying on the General Linear Model and implemented in SPM toolbox (Statistical Parametric Mapping) and region-of-interest (ROI) approach. Comparing brain responses to human and to ECA during interaction will already offer an interesting validation of how the different dimensions of social cognition are affected by the nature of the interaction partner.

The core of the project is to use fMRI corpus to estimate the timing of cognitive events through a reverse inference from brain activity. The important processing step is to transform 4D fMRI signal into binary or delta functions time series. The methodology proposed uses a similar procedure than for the skin conductance response, namely devolution of fMRI signal using the hemodynamic response function. Raw fMRI signal presents difficulties in comparison with skin conductance that impair a direct application of the method, namely its size (number of voxels), the relatively low signal to noise ratio and the low frequency signal trends. Classical fMRI preprocessing steps will therefore be applied, such as high-pass filtering and temporal and spatial smoothing using a gaussian kernel. To recreate the conditions of analysis used in the deconvolution of skin conductance, a region of interest (ROI) approach will be chosen. These regions will be chosen based on an in depth knowledge of their contribution to social cognition. ROIs will be identified on the basis of anatomy (e.g. (Wolfe et al., 2015) for the hypothalamus mask) and on the basis of existing coordinates and on the basis of localizer scans (eg voice localizer (Latinus and Belin, 2011)). We will obtain time-series (similar to the skin conductance time-series) for all ROIs. The result of the deconvolution will be binary (for sustained activity: is present for a certain duration) or delta function (for event-related: duration is null) time-series, identifying when during the course of the conversation do specific the cognitive events associated with the ROI analyzed (for example mentalizing for medial prefrontal cortex activity) occur. Physiological data will mainly be used as latent variables in a Dynamic Bayesian Network for identifying the timing of unidimensional (skin conductance) or multidimensional (fMRI) cognitive events during natural conversation. The outputs are therefore time-series tagging the moments when cognitive events take place during each trial of interaction with the ECA and the human.

5. Discussion

In this paper, we presented a new experimental set-up providing precisely aligned recordings of linguistic and physiological events for analysis. We have shown that the set-up allows for the recording of fine-grained linguistic phenomena, which enables an assessment of the level of interactivity of the dialogue. More crucially we have shown that the physiological measures obtained were correlated with various communicative behaviors. Therefore, we are now in the position to conduct more in-depth experiments and analyses in the field of Social Signal Processing in order to reveal temporal, and eventually causal, relationships between multi-modal linguistic behaviors and physiological activity. We also explained how we extend our pilot to work at neurophysiological level, in particular using functional magnetic resonance imaging (fMRI).

Moreover, we discuss the challenges caused by the spontaneous and interactional nature of the activity recorded. Of course the resulting analysis is somehow much more complex than for better controlled scenarios. In particular, a tricky problem that is likely to concern most of resources and projects attempting to link linguistic data with (neuro-)physiological measurements is to decide which analytical tool to use for relating the two types of data. The question of the temporal association will be also a general issue for this kind of project. The solution proposed here is only a first attempt to try to get a reliable association between the linguistic events and (neuro-)physiological signals. Finally, we illustrated the need to find the right balance set-up between what is desired, needed and required on the technical experimental side and the level of naturalness we would like to reach for our experiments. fMRI is probably among the most constraining experimental set-up, yet preliminary pilot in the scanner have shown that subjects manage to interact rather normally with someone outside the machine through our adapted communication device.

From a language sciences viewpoint, this project is a unique opportunity to correlate linguistic observations and models with other sources of evidence and in particular neurophysiological data. More precisely it allows to cross-validate verbal behaviors with neurophysiological recordings in the context of social interactions that are both spontaneous and controlled along a number of relevant dimensions thanks to the use of the artificial agent. In total, we will produce several hours of semi-controlled conversational data aligned with other behavioral information (gaze and face tracking) and physiological recordings (brain activity, skin conductance, respiration and peripheral blood pulse. We have also explored facial muscles activity and the head movements in (Ochs et al., 2017) from the participants of the behavioral experiment. This will constitute a unique corpus to further investigate the neurophysiological correlates of conversational activities allowing to address questions related to planning in interaction, face management and more generally about the interplay between cognitive, physiological and interactional constraints in language production.

6. Acknowledgments

The current work has benefited from support from the French government, managed by the French National Agency for Research (ANR), through the Brain and Language Research Institute funding (LABEX ANR-11-LABX-0036) and the convergence Institute for Language, Communication and the brain (ILCB ANR-16-CONV-0002), and the French Fund for Medical Research. The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille University - A*MIDEX, a French Investissements dAvenir programme (project PhysSocial).

7. Bibliographical References

Bach, D. R. and Friston, K. J. (2013). Model-based analysis of skin conductance responses: Towards causal models in psychophysiology. *Psychophysiology*, 50(1):15–22.

Benedek, M. and Kaernbach, C. (2010). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology*, 47(4):647–658.

Bevacqua, E., Prepin, K., Niewiadomski, R., de Sevin, E., and Pelachaud, C. (2010). Greta: Towards an interactive conversational virtual companion. *Artificial Companions in Society: perspectives on the Present and Future*, pages 143–156.

Bigi, B. and Hirst, D. (2012). Speech phonetization alignment and syllabification (sppas): a tool for the automatic analysis of speech prosody. In *Speech Prosody*, pages 1–4.

Chaminade, T. (2017). An experimental approach to study the physiology of natural social interactions. *Interaction studies*, 18(2):254–275. Fq6kd Times Cited:0 Cited References Count:31.

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. Academic Press.

Latinus, M. and Belin, P. (2011). Human voice perception. *Current Biology*, 21(4):R143–R145.

Ochs, M., Libermann, N., Boidin, A., and Chaminade, T. (2017). Do you speak to a human or a virtual agent? automatic analysis of user's social cues during mediated communication. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI), Glasgow, UK*, page 9 pages.

Pelachaud, C. (2015). Greta: an interactive expressive embodied conversational agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 5–5. International Foundation for Autonomous Agents and Multiagent Systems.

Roth, W. T., Dawson, M. E., and Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, 49:1017–1034.

Wolfe, F. H., Auzias, G., Deruelle, C., and Chaminade, T. (2015). Focal atrophy of the hypothalamus associated with third ventricle enlargement in autism spectrum disorder. *NeuroReport*, 26(17):1017–1022.

Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE.

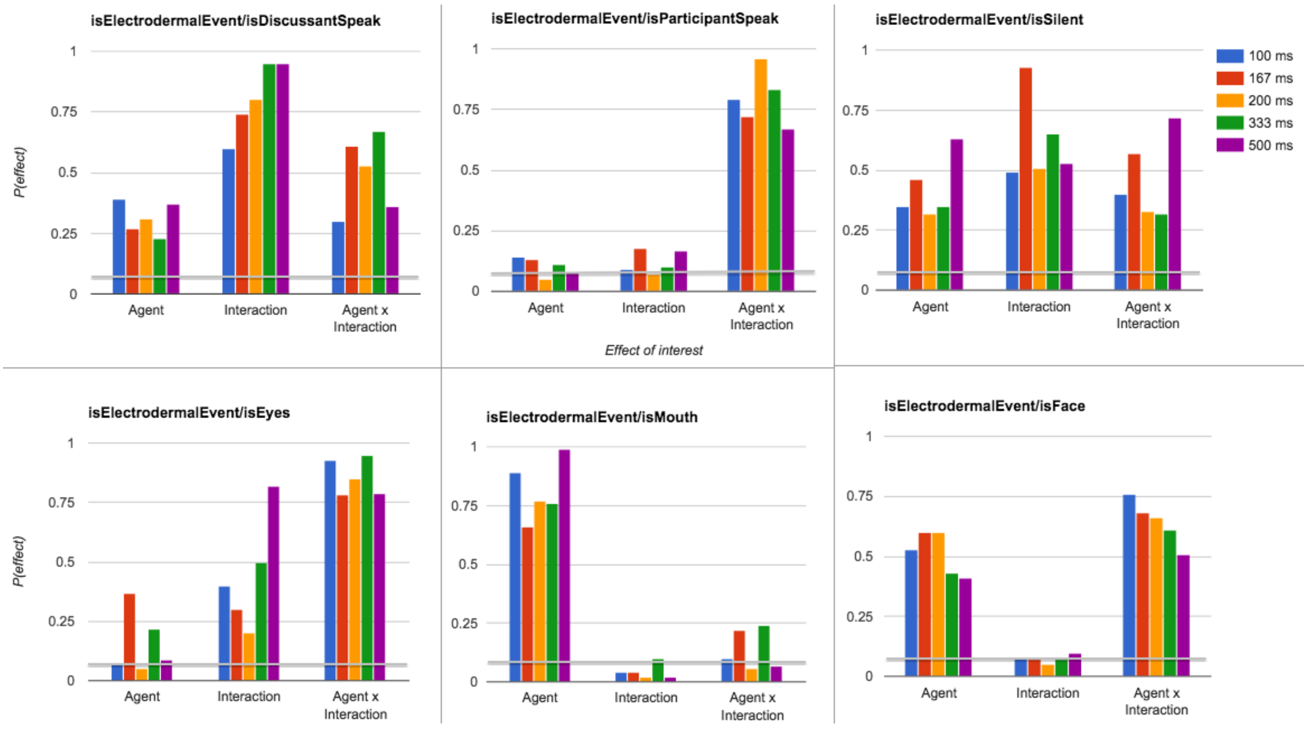

Figure 2: Outputs of ANOVAs on the effect of experimental factors Agent and Interaction on posterior probabilities of obtaining a physiological response co-occuring with a particular behaviour. In abscissa are the effects of interest (main effect ”nature of the Agent (human vs artificial)”, ”nature of the Interaction (live vs video)”, and interaction between the two factors) and in ordinate the probability (grey line: p=0.05). Colours represent the five time windows used to define IPUs.

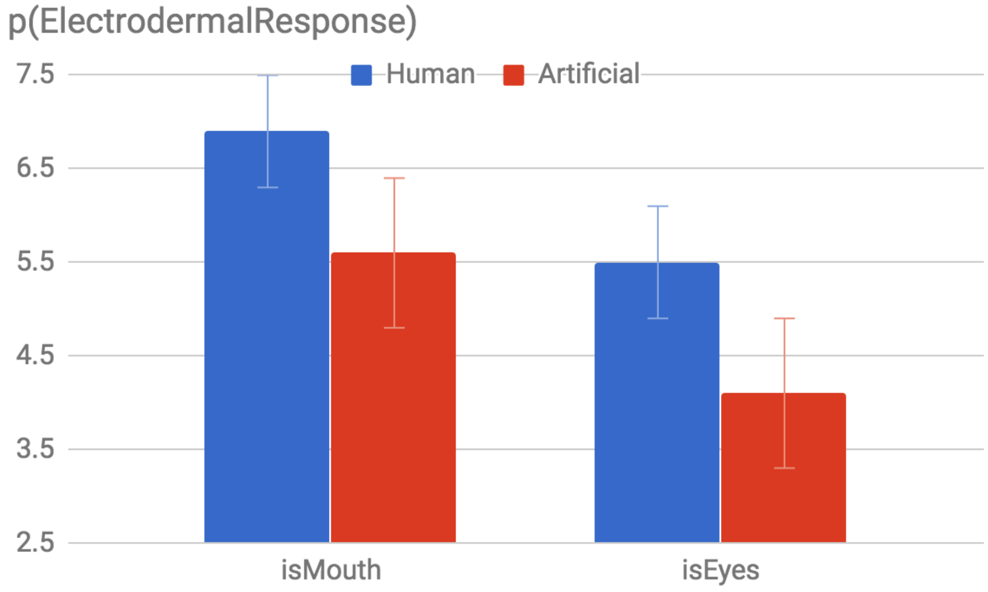

Figure 3: Posterior probability in per cent) of observing a physiological event given a behavior (participant looking at the mouth or looking at the eyes) as a function of the Agent.