Do You Speak to a Human or a Virtual Agent? Automatic Analysis of User's Social Cues during Mediated Communication

Magalie Ochs
Laboratoire des Sciences de l'Information et des Systèmes, LSIS, UMR7296, Aix-Marseille Université, CNRS, ENSAM, Université de Toulon, 13397 Marseille, France

Nathan Libermann
Laboratoire des Sciences de l'Information et des Systèmes, LSIS, UMR7296, Aix-Marseille Université, CNRS, ENSAM, Université de Toulon, 13397 Marseille, France

Axel Boidin
Picxel Company, Marseille, France

Thierry Chaminade
Institut de Neurosciences de la Timone, UMR 728 Aix-Marseille Université, CNRS, ENSAM, Université de Toulon, 13397 Marseille, France

ABSTRACT

While several research works have shown that virtual agents are able to generate natural and social behaviors from users, few of them have compared these social reactions to those expressed during a human-human mediated communication. In this paper, we propose to explore the social cues expressed by a user during a mediated communication either with an embodied conversational agent or with another human. For this purpose, we have exploited a machine learning method to identify the facial and head social cues characteristics in each interaction type and to construct a model to automatically determine if the user is interacting with a virtual agent or another human. The results show that, in fact, the users do not express the same facial and head movements during a communication with a virtual agent or another user. Based on these results, we propose to use such a machine learning model to automatically measure the social capability of a virtual agent to generate a social behavior in the user comparable to a human-human interaction. The resulting model can detect automatically if the user is communicating with a virtual or real interlocutor, looking only at the user's face and head during one second.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**;

KEYWORDS

Embodied Conversational Agent; Social signals; human-machine mediated communication

ACM Reference Format:
Magalie Ochs, Nathan Libermann, Axel Boidin, and Thierry Chaminade. 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'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3136755.3136807

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3136807

1 INTRODUCTION

Recently, virtual characters are increasingly used for communication with users. Performing different roles, they could play, for instance, the role of virtual assistant in commercial website (e.g. *Ines* [1] or *Yoko*[2]) or the role of virtual guide in public spaces, such as in Museum [14]. When computers are used in these roles, they are often embodied by animated cartoon or human-like virtual characters, called *Embodied Conversational Agents* (ECA) [7]. This enables a more natural style of communication for the human and allows the computer to avail of both verbal and non-verbal behavior channels of communication. Several studies have demonstrated the acceptance and the efficiency of such agents [9, 16]; indeed, the *persona effect* reveals that the presence of an ECA improves the experience of an interaction for the user (for instance [17, 24]).

When people interact with such virtual agents, several research works tend to show that users react naturally and socially as they would do with another person [16, 27]. This user's social behavior, triggered automatically and unconsciously [20], is characterized by different social cues such as smiles [15] or head movements [3]. However, the question that remains unclear is *how the social cues expressed by the users in front of a virtual agent differ from those expressed during a same interaction with a human?*

As stated in [31], *social cues* (also called *behavioral cues*) correspond to *observable* cues (mainly non-verbal) conveying information on "*feelings, mental state, personality and other traits of people*" and also determining "*the nature and quality of the social relationship with others*". Consequently, the comparison between the social cues expressed by a user interacting with a virtual agent or expressed during a similar interaction with a human may enable us to compare the social experience of the user depending on the nature of her interlocutor (virtual or real). If the user expresses the same social cues, we can suppose that the quality and the nature of the social relationship is equivalent. In this article, we aim at exploiting the social cues expressed by the user during a human-machine interaction to measure the quality of an interaction compared to a human-human interaction. For this purpose, we have explored the differences in the social cues expressed by the user during a

[1] https://www.nespresso.com/fr/fr/service-customer-care
[2] http://www.toshiba.fr/support/laptops/

mediated communication either with an embodied conversational agent or with another human. The objective is to identify whether an interaction in the same context induces different social reactions in a user if her interlocutor is embodied by a virtual agent or is another human. For this purpose, a corpus-based analysis using machine learning method has been performed. The audio-visual corpus is composed of mediated communication of users interacting either with another human or an embodied conversational agent in a same context of dialog. The social cues of the users are extracted automatically. In the research reported in this paper, we have focused on the social cues related to the user's faces and head, social cues particularly relevant of the user's social attitudes [6, 30]. These social cues expressed by the users during either a human-human or a human-virtual agent mediated communication are compared using machine learning algorithms. The methodology consists in considering a problem of classification to highlight which social cues differ from one type of interaction to another. Such a methodology has the advantage to enable us to (1) identify the importance of the social cues in these two types of interactions and (2) to develop a computational model to automatically detect whether a user is talking to another human or a virtual agent in a mediated communication, by only using information from her face and head[3].

The paper is organized as follows. In the next section, we present related works in the domain of Social Signal Processing. In Section 3, we introduce the corpus exploited in this research work. Section 4 is dedicated to the analysis of the head and face social cues that are relevant to distinguish the interaction of a user with another human or a virtual agent. In Section 5, we more particularly analyze the minimum time necessary for the classifier to determine if a user is talking to a human or a virtual agent by looking at her face and head. We discuss the results in Section 6.

2 RELATED WORKS

In the emerging domain of Social Signal Processing (SSP) [30], different research axes may be identified [31]:

(1) the *modeling axis* aims at studying the social cues used for the expressions of social functions such as intentional stance or empathy ;
(2) in the *analysis axis*, researchers focus on the development of computational models for the automatic detection of social cues and the inference of related social functions (e.g. detection of suspect behavior based on the non-verbal cues expressed by people [23];
(3) the *synthesis axis* aims at creating interactive systems (e.g. ECA or humanoid robots) able to express social cues eliciting social behavior from the users (e.g. the expressions of social stances through smiles [22]).

The work presented in this article is at the frontier between the (1) modeling axis and the (2) analysis one. Indeed, we aim at (1) identifying the differences between the social cues expressed by the user during a mediated communication depending on whether her interlocutor is human or virtual and (2) developing a computational model to automatically recognize if the user talks to a virtual agent or a human.

The non-verbal cues involved in the expressions of social functions cover different modalities [12]: *face and head behavior* (e.g. facial expressions, head movements), *vocal behavior* (e.g. prosody, interruptions), *the appearance* and *space and environment*. In this article, given the context of the study that is a mediated communication through Skype (Section 3), as a first step, we focus on the *face and head behavior* of the user. Facial expressions and head movements are indeed well-known to reflect one's emotions and stances (e.g. agreement).

Several studies have been conducted to analyze the social cues expressed by users interacting with virtual characters. Most of existing studies have used the user's social cues to measure the impact of virtual character's behavior. For instance, in [4], the gaze behavior of users has been compared depending on the gaze behavior of the virtual interlocutor, or the effects of virtual character's smiles have been measured through the user's smiling behavior [15]. In these studies, the effects of different behaviors of a same virtual character are evaluated in terms of user's expressed social cues. In the present study, we aim at measuring the effects of different interlocutors (virtual or human) in the user's expressed social cues. Nowadays, as far as we know, no study has compared the social cues expressed by a user depending on whether she is interacting with a virtual character or with a human.

Machine learning techniques are commonly used in SSP domain to develop computational models to automatically recognize social functions based on social cues by exploiting different learning methods: for instance to recognize emotions using deep learning method [13], to identify social roles in meetings based on a HMM [29] or to predict the personnality of the participants [1], to determine the level of engagement of the users during interaction based on k-nearest neighbors algorithm [19], or to predict the rapport of a user with a virtual agent [8, 33]. As far as we know, currently, no model has learned to automatically recognize if the user is discussing with a virtual agent or another human. In this article, we have explored classification algorithms to construct such a model (Section 4).

In order to better understand the social cues implied in the expression of social functions, some researchers have also proposed to use a specific machine learning technique called the *feature selection* algorithms. Compared to the classical statistical tests (e.g. One-way Anova and post-hoc Tukey HSD tests), the feature selection algorithms have the advantage to identify the relevant social cues characterizing a phenomena. For instance, [11] used Random Forest method to identify the most relevant body cues characterizing emotions in a database of daily actions performed with different emotions. Using a similar approach, in this article, we explore the random forest method to compare the social cues expressed by a user in different conditions of interactions (Section 4).

In the next section, we present in more details the corpus that we have exploited to analyze the social cues expressed by the user depending on her interlocutor (virtual or real) (Section 3) and to construct a model to automatically detect if the user is speaking to a human or a virtual agent (Section 4).

[3]In the presented work, we have not evaluated the capacity of individuals to classify the interaction. Indeed, our objective was to evaluate the capacity of a computer to do this task. Moreover, the proposed machine learning method has enabled us to identify more precisely the relevant characteristics that differ in the two types of interaction, a task that could be done, in fact, by humans but more time-consuming and difficult to validate in this case.

3 THE CORPUS

3.1 Experimental Recording of the Corpus

The corpus is composed of data that was recorded as part of an experiment that compares behavioral and physiological responses when a participant has a natural social interaction with a human or an embodied conversational agent[4]. Experimental procedures used for data collection adhered to the Declaration of Helsinki. Data was acquired in a proof-of-concept experiment assessing the feasibility of the non-interventional procedure with few participants prior to requiring a formal ethical approval, that is now being submitted for consideration by the local ethics committee. Note that participants signed informed consent.

In a nutshell, pairs of naive participants are tested together in an experimental setup using videoconferencing to support the discussion. A cover story provides credible, but spurious, explanations to many aspects of the experimental set-up. Participants are made to believe they participate in a marketing experiment to validate an incoming advertising campaign. The cover story is fundamental in this experiment. It provides a common goal for the two interacting agents as well as a topic for the discussion. The fact that videoconferencing is used, a requirement in order to record the face from the front and to present the artificial conversational agent, is presented as a necessity to control precisely the time the two participants discuss together. This time pressure - one minute per trial - is important to avoid wavering. For the sake of keeping the instructions natural, the experimenter presented, apparently informally, the goal and setting of the experiment to the pair of naive participants, who did not know each other, upon arrival. It took 15 to 30 minutes to provide all required information, depending on the questions participants asked during the presentation. Because only a female voice was available for the artificial agent, only women were included to avoid mixing the gender of the two agents discussing. The corpus comprises 11 female students recruited by word-of-mouth (mean age 22.7 years, standard deviation 6.4 years).

An embodied conversational agent presented as autonomous is used to compare physiological and behavioral responses to interactions with a fellow human. In order to fulfill its function, a simple Wizard of Oz (WoZ) procedure controls the ECA, so that unbeknown to the participant, a human controls the ECA directly. One of the experimenter controlled the agent's behavior in all conditions and for all participants. The exchanges were brief (one minute) and centered on the issue introduced in the cover story (the neuromarketing experiment). Finally, behaviours, including speech production and upper body movements, were programmed a priori to the experiment, the WoZ being limited to selecting the most adapted behaviour given the ongoing conversation.

The embodied conversational agent used to interact with the participants was the Greta system [25]. GRETA is an experimental platform specifically dedicated to investigate verbal and nonverbal aspects of human-machine interactions and is particularly relevant for the current project as it is able to reproduce human emotional states and generic behavioral feedbacks [21]. A voice synthesizer from company CereProc was used to generate speech [2]. Each participant interact with another participant or the ECA three times.

Upon arrival to the laboratory, one participant was located in the recording room while the second went in the discussant room. In the Participant room, the recorded participant sat comfortably on a chair in front of a computer screen topped by the webcam used for the Skype discussion. Two cameras forming the recording part of the Facelab eyetracker used for gaze tracking were located under the screen and connected to a computer dedicated to gaze tracking. The left hand of the participant was fit with a photoplethysmograph sensor on the thumb to record blood pulse and the two electrodes of the electrodermal activity sensor on the index and middle fingers. A computer was connected to the screen and the webcam (bidirectional dotted arrow) and the participantâĂŹs headphones. Headphones were used so that the speech from both participants were acquired separately. The installation of the discussant room consisted in another computer connected to the discussant screen, webcam and headphones running skype.

Each trial consisted in viewing an image (related to the cover story) for 10 seconds, followed by 3 to 5 seconds of black screen, and then one minute during which the participant talks with the discussant, depending on the experimental condition (human or ECA). In each trial, we recorded the participants' screen and microphone as well as the corresponding video and audio data from the discussant (Figure 1). Participants were informed about the cover story and the actual objective of the experiment, as well as the recording of video and audio data for analysis, during the debriefing session following the experiment.

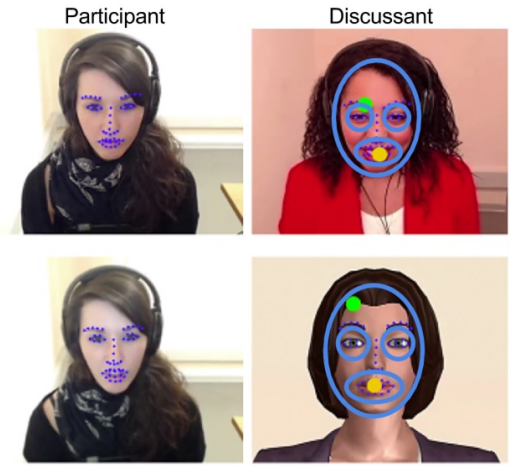

Figure 1: Combination of face and gaze tracking on one frame. Blue dots represent features tracked by the face tracking program. Circles indicate regions of interest on the Discussant based on face tracking, the green dot the direction of the ParticipantâĂŹs head and yellow dot the direction of ParticipantâĂŹs gaze.

A screen recording software is used to record the video and audio of the conversation. The video data is analyzed to extract facial features for each frame. A face recognition algorithm (Facial Feature

[4]Note that corpus including human-human and human-agent interactions already exist (such as the SEMAINE database [18]). We have created our own database to collect also physiological data and to analyze in a second step the physiological reactions of users in front of a virtual agent or another human.

Detection & Tracking; [32]) is run frame by frame to identify the face present in the image. Screen x and y voxel coordinates of 49 key-points on the face are recorded as well as the position and rotation of the face mask in relation to the screen normal vector.

In total, the collected corpus is composed of 6 videos per participant (each participant interacting three times with another participant and with the ECA to discuss three different images). We had 11 participants. However, 2 videos were set aside because of recording problems. So, in total, the corpus is composed of 64 videos (32 for human-machine interaction and 32 for human-human interaction).

3.2 Pre-processing of the Data

In this article, we focus in the head and face movements of the participants interacting either with an ECA or another human. The face tracking returns for each image the coordinate x,y of 49 points of the face, resulting in 98 points per image. Then, each video is transformed into a 1800*98 matrix characterizing the data resulting from the face tracking for each image of the video (30 images per seconds for video of 1 minute, being 1800 images in total).

Based on the face tracking results, we have used Picxel's software face2market[5] to compute the rotation angles of the head as well as the intensity activation of the muscle of the face. The software is able to detect different facial expressions and to measure their intensity from videos. Based on the Facial Action Coding System (FACS) proposed in [10], the software computes its own AUs. We kept the name of *Action Unit* but they differ from those of [10]), in particular in the way they are computed. In the FACS, each AU is associated to a value corresponding to the tension of the muscles. From the characteristic points detected by the face tracking results, Face2Market first computes a distance from the neutral position of the face (0 if neutral, 100 if maximum). In our study, the neutral position is considered to be the first frame of the video. Then, the distance from the neutral position allows us to determine if the different AUs are activated as defined in [10]. Moreover, Face2Market computes for each image of the video the head movements following 3 rotation axes : vertical for the head nodes, lateral for the head shakes and horizontal for the tilts. We used these rotation angles to correct the distances according to a 3D face modeling and compare them to the neutral face. Finally, we choose to focus our study on 5 AUs : AU1 : Inner Brow Raiser, AU2 : Outer Brow Raiser, AU5 : Upper Lead Raiser, AU20 : Lip Stretcher, AU25 : Mouth Opening. The choice of these AUs has been done since their combination is sufficient to roughly describe the basic facial expressions as smile, anger and surprise. Three of the five AUs occurring in the left and right side of the face (AU1, AU2 and AU5), they are considered as 2 different values, and allow us to confirm the activation of each AU and to compute the intensity of the movement.

Finally, in total, we consider 11 variables: 8 variables for the facial movements (given that 3 AUs have two values since they are symmetric) and 3 variables for the head movements. Each video of 1 minute is then described by a matrix of 1800*11.

To exploit the corpus with machine learning method, each video has been labeled as human-human interaction or human-machine interaction, the two classes that we consider in learning.

[5]http://www.picxel.fr/index.php/en/

In order to verify the stability of the learned model, we have constructed three learning corpora from the initial corpus of 64 videos. In this initial corpus, each participant has interacted 3 times with a human and 3 times with the virtual agent. For each participant, we have 6 videos of 1 minute (3 human-human and 3 human-machine). We had 11 women participants. To construct the 3 new learning corpora, for each of them, we selected randomly for each participant one of the human-human video and one of the human-machine video. They are used for the test set. The other videos are used for the training set.

Finally, we have 3 different corpora but with the same proportion and an equal number of elements for each class[6]. In a nutshell, from the initial corpus, we have constructed three corpora (called corpus1, corpus2, and corpus3 in the following sections). Each of them is composed of the same 64 videos but with different training and test sets. The training set of each corpus counts 42 videos and the test set counts 22 videos. The two classes (human-human and human-machine) are represented equally in each corpus. Each video of one minute is represented by a matrix of 1800*11 characterizing the activation of actions units (AU1, AU2, AU5, AU20, and AU25) of the user's face and the head position (lateral, horizontal and vertical angle rotations of the head).

4 SOCIAL CUES CHARACTERIZING HUMAN-HUMAN VERSUS HUMAN-MACHINE INTERACTION

In Human and Social Sciences, several research works deal with the identification of social cues associated to social functions (e.g. the smiles effects on impression formation [26]). In this article, we do not directly study a particular social function but aim at analyzing the difference in expressed social cues depending on the interlocutor (virtual or real). For this purpose, in order to identify the relevant social cues characterizing the differences between a human-machine versus a human-human mediated communication, we have explored a machine learning method called the feature selection algorithm. These algorithms have the advantage of identifying the most relevant features characterizing a class.
In our context, we consider two classes:

(1) human-human interaction: the user interacts with another human through skype;
(2) human-machine interaction: the user interacts with the ECA through video (as in a skype communication).

By using a feature selection method, our objective is to identify the user's expressed social cues that differ between these two classes. In other words, this method enables us to determine the most relevant social cues characterizing a human-machine interaction compared to a human-human one. In this work, we focus on the facial and head movements (Section 3.2).

Different feature selection algorithms exist. Here, we have used the *Random Forest approach* [5] implemented in the Scikit-learn open-source tool[7]. The Random Forest is a popular machine learning method used in several domains (bioinformatics, computer vision, image processing, etc.). The Random forest method has the

[6]Note that given the composition of the corpus, the classification models are not trained and tested to be robust for new subjects.
[7]http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

particular advantage, compared to other statistical models as for instance RNN, to measure the relevance score for each feature but also to handle high-dimensional data with a high generalization power [28]. Note that our objective in the presented study was not only to obtain good performance in classification but also to highlight relevant features that differ depending on the interaction. It's why we have chosen to use the Random Forest approach. Such a method has already been used in a similar task to identify relevant social cues in body movements [11]. We have used the random forest algorithm with 100 decision tree.

As commonly used, we have computed four measures to evaluate the quality of prediction of a model : the accuracy (or classification accuracy) represents the number of correct predictions from all predictions made; the precision a measure of classifier exactness; the recall a measure of classifiers completeness, and the F1 Score the balance between the precision and the recall.

4.1 Automatic Analysis of the User's Muscles of the Face and Head Positions

In order to measure the importance of each modality, as a first step, we have considered only the facial expressions of the user (i.e. the activation of muscles of the face, Section 4.1.1) and we have then added the head position (i.e. the rotation of the head, Section 4.1.2).

4.1.1 The facial expressions. In order to analyze the *facial expressions* differing in human-human and human-machine mediated communication, we have applied the random forest classifier on the 3 corpora of the videos, each video being described by the muscle activation values of the face for each image frame (i.e. a matrix of 1800*8) . The results are reported in Table 1 for each of the 3 corpora. The performances of the resulting classifier are far from satisfying

Table 1: Performance for the classification of the human-machine (H-M) or human-human (H-H) interaction based on the activation values of actions units of the user's face.

	Accura.	Precision		Recall		F1	
		H-H	H-M	H-H	H-M	H-H	H-M
Corpus1	41%	42%	40%	45%	36%	43%	38%
Corpus2	40%	38%	43%	27%	55%	32%	48%
Corpus3	59%	58%	60%	64%	55%	61%	57%
Average	46%	46%	47%	45%	48%	45%	47%

with an average accuracy under 50%. In other words, these results tend to show that the facial muscle activations is not relevant to distinguish an interaction with a virtual character or a human. It seems that the users tend to express similar facial expressions during a interaction with a virtual character and with a human in the context considered in this study[8]. In the next sections, we analyze the importance of other features considering in addition the head position.

[8]Note that in the presented study, we have voluntary not considered the "sequentiality" of data but the "instantaneity" of the information to try to identify if specific facial expressions ("screenshots" without considering their dynamic) differ depending on the type of the interaction.

4.1.2 The facial expressions and head positions. The head positions characterized in the video by the angle values on the 3 axes (vertical, horizontal, and lateral) have been used with the facial muscle activation to try to distinguish a user's interaction with either a virtual character or a human. In the considered corpus, each video is described by a matrix of 1800*11 to represent both the intensity of the AU and the rotation angles of the head on each frame. The results are reported in Table 2.

Table 2: Performance for the classification of the human-machine (H-M) or human-human (H-H) interaction based on the activation values of actions units of the user's face and on her head rotation values.

	Accur.	Precision		Recall		F1	
		H-H	H-M	H-H	H-M	H-H	H-M
Corpus1	68%	64%	75%	82%	55%	72%	63%
Corpus2	68%	70%	67%	64%	73%	73%	70%
Corpus3	59%	58%	60%	64%	55%	61%	57%
Average	65%	64%	67%	70%	61%	66%	63%

Compared to the performance considering only the facial expression (Section 4.1.1), the results show that the head positions improve the classification with an average accuracy of 65%. The performance remains low, revealing that the facial expressions and head positions are not the best cues to detect if a user is interacting with a virtual character or a human.

4.2 Look at the Dynamic of the User's Face to Determine Automatically the Interlocutor Type

4.2.1 Dynamic of the head and face movements. In order to go beyond the facial and head position, we have analyzed the importance of the dynamic of the face and the head to distinguish human-human versus human-machine interaction. For this purpose, we have computed the time derivative (by subtracting the value at the time T+1 to the value at the time T for each value of AUs and head rotation).

The dynamic of the face and the head are then represented by the time derivative of each action unit and of the three angles of the head rotation. A null value for a given action unit means that the corresponding muscle of the face did not move. In the same way, a null time derivative for a given angle of the head means that the head did not move on the corresponding axis. On the contrary, the more the value of the time derivative is high, the more the movement of the muscle of the face or the rotation of the head was pronounced.

In this section, we report the results considering only the derivatives as features (in the next section 4.2.2 we consider the previous features and the derivatives). The results of the random forest algorithm on the data only characterized by the dynamic of the head and face are reported in Table 3.

The dynamic of the facial muscles and of the head enable us to obtain an accuracy for classification in average around 83%. It means that the movements of the face and of the head of a user seem to vary consequently depending on the type of her interlocutor

Table 3: Performance for the classification of the human-machine (H-M) or human-human (H-H) interaction based on the time derivative values of the actions units of the user's face and the time derivative of the user's head positions.

	Accur.	Precision		Recall		F1	
		H-H	H-M	H-H	H-M	H-H	H-M
Corpus1	77%	80%	75%	73%	82%	76%	78%
Corpus2	86%	83%	90%	91%	82%	87%	86%
Corpus3	86%	83%	90%	91%	82%	87%	86%
Average	83%	82%	85%	85%	82%	83%	83%

(virtual or human). Compared to previous results (Section 4.1), these results show that the dynamic of the head and face muscles is a better feature to distinguish a human-human or human-machine interaction than the user's facial expression and head position. In the next section, we consider all these features together for machine learning.

4.2.2 Head and face movements and positions. In order to consider both the facial expressions and head positions and their dynamic, we have computed a matrix for each video gathering the values of actions unites and head positions and their time derivatives. The results of the random forest algorithm on these data is reported in Table 4.

Table 4: Performance for the classification of the human-machine (H-M) or human-human (H-H) interaction based on the values of the action units of the face, the position of the head and the time derivative of these face and head values.

	Accur.	Precision		Recall		F1	
		H-H	H-M	H-H	H-M	H-H	H-M
Corpus1	90%	91%	91%	91%	91%	91%	91%
Corpus2	95%	92%	100%	100%	91%	96%	95%
Corpus3	90%	91%	91%	91%	91%	91%	91%
Average	92%	91%	94%	94%	91%	92%	92%

The accuracy of the classification considering the face and head positions and their dynamics are in average of 92%. The results are stable on the different corpora varying between 90% and 95%. These results reveal the relevant social cues that distinguish a human-machine interaction from a human-human one. Based on these features, we can automatically detect if the user is speaking to a virtual agent or a human. In other words, looking at the user face and head: the muscle activation, the position and movements is sufficient to identify her interlocutor type (virtual or real) in the current interaction setup.

In the next section, we present a fine-grained analysis of the importance of the facial and head movements and positions to characterize the human-machine interaction compared to the human-human one.

4.2.3 The social cues characterizing human-machine interaction. The random forest algorithm provides a relevance measure for each point of the matrix. These scores represent the importance of the features for the classification. The score is in [0,1] with the sum of each line and column equals to 1. In our context, this score represents the importance of each feature at each frame of the video to distinguish a human-human interaction to a human-virtual agent one. In order to obtain the relevance of the features for the entire video (on the 1800 frames), we have summed the values on the columns. The importance of the features obtained for each corpus are reported in Figure 2.

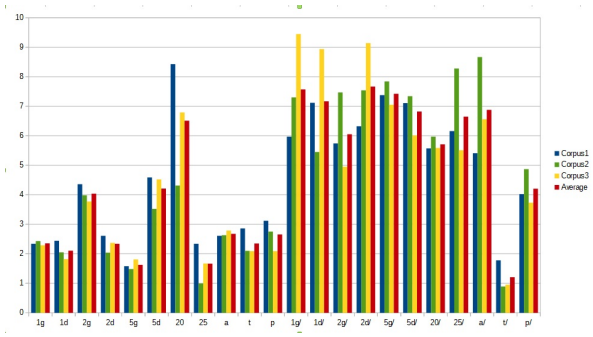

Figure 2: Relevance measures, resulting from the feature selection algorithm. The ordinate represents the average percentage of importance for each feature on the 3 corpora. The features are represented in abscissa: *1g* for AU1 left, *1d* for AU1 right, *2g* for AU2 left, *2d* for AU2 right, *5g* for AU5 left, *5d* for AU5 right, *20* for AU20, *25* for AU25, *a*, *t* and *p* represent the angles of the head rotation. *1g/*, 1d/, ..., *a/*, *t/*, *p/* represents the times derivatives of the features.

The Figure 2 highlights the importance of some features in particular. For instance, the lip stretcher (AU20), seems particularly discriminative to distinguish an interaction with another human or an agent. The dynamic of the eyebrows (e.g. 1g/, 1d/, 2g/, 5d/), of the lip part (AU25) that may characterize the speech activity, and the head nod (a/) seem to have a certain importance for the classification compared to the other features. In other words, the lips and eyebrow behavior as well as the speech activity of the user seem to represent relevant social cues that differ from an interaction with a virtual agent to an interaction with another human.

Some of the features most important to discriminate between H-H and H-M interactions are in line with informal observations we could make on the interactions videos. For instance, it seems that the internal dynamics of the usersâĂŹ face, in terms of lips movements for example, and of the head, nodes for example, are more accentuated during a mediated interaction with another human that during an interaction with the virtual agent. More generally, it seems that participants act more, including speaking more, when interacting with a human than an artificial agent. These increased non-verbal behaviors may traduce an engagement of the user more pronounced in a human-human mediated interaction: the head and eyebrows movements could be interpreted as expressions of simple communicative cues as well as encouragement for listener response (backchannels). Note that these different features may be easily automatically detected during an interaction, for instance

by exploiting the PicxelâĂŹs software used to analyze the corpus (Section 3.2).

5 A QUICK GLANCE TO DETERMINE AUTOMATICALLY THE INTERLOCUTOR TYPE

5.1 Optimized Time Interval for Classification

In the previous section, we have considered the entire video to determine automatically if the user is talking to a human or an ECA. In this section, we propose to determine the smallest segment of video necessary for an accurate classification. In other words, we investigate the video duration necessary for a computational model to determine automatically the user's interlocutor type in a mediated communication. For this purpose, we have segmented the videos of the corpora with different durations: from 2 ms (1 frame of the video) to 50 seconds (1500 frames of the video). Figure 3 illustrates the accuracy of the random forest classifier for the corpora with video samples of different durations. The considered corpora included the data on the head and face position and movements. The video samples are selected randomly. To avoid an effect of the influence of the position of the sample in the video on the results, we have performed 10 runs for each duration and corpus and report the average results in Figure 3. The results show that the

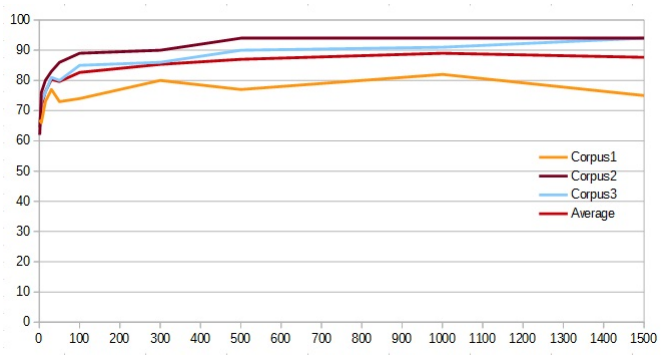

Figure 3: Classification accuracy with video samples of different durations. The ordinate represents the classification accuracy depending of the duration of the video described in abscissa. The average curve corresponds to the average accuracy of the three corpora. The time on the abscissa axis is indicated in number of frames (1 frame is 2 ms).

entire video is not necessary for an accurate classification. Indeed, between 500 frames (around 17 seconds) and 1000 frames (around 33 seconds), we obtain satisfying results similar to those obtained for the entire video. With a sample in this interval of duration, the classifier may determine if the user is implied in a human-machine or a human-human mediated interaction with an average accuracy between 87% and 89%. The average accuracy decreases if the video samples are longer. The performance of the classification does not increase with the duration of the video samples.

5.2 Oversampled Based on Optimized Time Interval

The results reported in the previous section show that the entire video of 1 minute is not necessary to predict automatically the user's interlocutor type. Indeed, shorter video samples provide satisfying performance. Given the results described in the previous section, we consider samples of 30 seconds.

To consider shorter video samples has several advantages. First, we reduce the size of the matrix representing each sample. Another major advantage is to increase significantly the size of the learning corpus. Indeed, for each video of 1 minute, we can extract a large set of different video samples of 30 seconds. We have constructed different training and test sets of video samples of 30 seconds varying the number of samples extracted from each video. We have studied the optimized number of video samples of 30 seconds needed per video for an accurate classification. The results are reported in Figure 4. Let be X the number of samples extracted from one video. We have computed the accuracy of the feature section algorithm considering the head positions and facial expressions and their dynamics with X varying from 1 to 1000. For instance, if $X = 2$, the training set is composed of 2*42 video samples of 30 seconds and the test set is composed of 2*22 video samples of 30 seconds. The results show

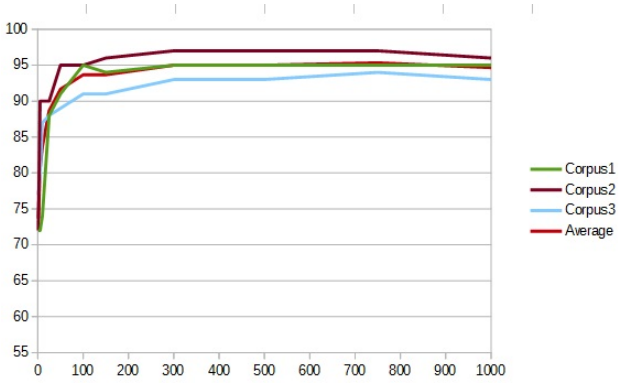

Figure 4: Classification accuracy with training and test sets of different sizes. The value on the abscissa axis represents the number of samples extracted from one video (X). The average curve corresponds to the average accuracy of the three corpora.

that the accuracy is maximum for X around 300, *i.e.* with a training set composed of 300*42=12600 video samples of 30 seconds, with an accuracy of 95%. In other words, only 30 seconds of video is necessary for the algorithm to determine if the user is conversing with a human or a virtual character by looking at his face and head movements and positions.

In order to test whether the increase of size of the training set may enable us to consider shorter video samples, we have computed the classification accuracy for video samples of 1 seconds considering a training set with X=300 (the number of samples extracted from one video), *i.e.* a training set of 12600 video samples of 1 second. The results are reported in Table 5.

The results highlight that the increase of the size of the corpus enables the computational model to distinguish more rapidly

Table 5: Performance for the classification of the human-machine (H-M) or human-human (H-H) interaction based on the values of the action units of the face, the position of the head and the time derivative of these face and head values, considering video samples of 1 second and a training set of 12600 video samples.

	Accur.	Precision		Recall		F1	
		H-H	H-M	H-H	H-M	H-H	H-M
Corpus1	95%	94%	96%	96%	94%	95%	95%
Corpus2	97%	96%	98%	98%	96%	97%	97%
Corpus3	92%	93%	91%	91%	94%	92%	92%
Average	94%	94%	95%	95%	94%	94%	94%

a human-human-versus a human-machine interaction, in only 1 second with an accuracy around 94%.

6 DISCUSSION

In this article, we have proposed to apply machine learning techniques to analyze the differences between the user's social cues expressed during a mediated communication either with a virtual agent or another human. We have more particularly exploited a feature selection algorithm, called the random forest, on features characterizing the facial expressions of a user (described by specific muscle activations, *i.e.* action units), the dynamic of the facial expressions (*i.e.* the activity of specific muscles of the face: the activity of action units), the head position (vertically, horizontally, and laterally) and movements (head nodes, head tilts, and shakes) during a recorded mediated communication of 1 minute with a human or an agent. The feature selection algorithm has been exploited (1) to analyze the most relevant face and head social cues that distinguish a human-human versus a human-machine interaction, and (2) to construct a computational model to automatically determine if a user is communicating with a virtual agent or another human.

The results concerning the point (1) shows that, in fact, the user does not express the same facial and head social cues during a communication with a virtual agent or another user. However, it seems that it is not specifically the facial expressions that differ for a user interacting with another human or a virtual agent, but rather the dynamic of the facial expressions: the frequent changes of the eyebrows and lips positions. These social cues may characterize smiling behavior, speech activity and certain emotions or cognitive states (such as surprise, uncertainty, anger). Indeed, the speech activity seems to not be the same in the two conditions: the users seem to talk more in front of another human than in front of a virtual agent. In our model, this information is conveyed through the dynamic of the face, and in particular through the dynamic of the action unit related to the lips. A computation of the duration of the speech of the participants in the two conditions will enable us to easily confirm this hypothesis. More fine-grained analysis considering other action units and using other statistical techniques could be explored to identify more particularly the emotions and cognitive states experienced during each condition.

In the point (2), using the same algorithm, considering the task as a classification problem, we have constructed a computational model that can detect automatically if the user is communicating with a virtual agent or another human, looking at only her face and head during 1 second, with an accuracy around 94%. This result shows that in fact the virtual agent does not generate the same social cues as another human during a mediated communication. We can determine very quickly mainly based on her lips, eyebrows, and head movements, if the user is talking to another human or an agent. However, in the presented study, we have not considered the temporal position of the video samples of one second in the entire video of one minute. In a next step, a distinction between the samples situated at the beginning of the interaction, at the middle or at the end will enable us to identify from which moment during the interaction we can automatically detect if the user is interacting with another human or with a virtual agent.

If we consider the human-human interaction as the natural and social one, our research results mean that the virtual agent is not able to generate natural and social behavior on behalf of the human user as another human can do. This result is not totally surprising since virtual agents are far from comparable to human in terms of believability in its non-verbal behavior. However, the proposed model may also be used to evaluate a virtual agent: the more difficult is to construct an accurate machine learning model to distinguish a human-human interaction from a human-machine one, the more the virtual agent is successful in its capacity to generate social behavior in human. In other words, in this paper, we have explored the possibility to develop model to evaluate the social competence (in terms of the capacity to generate social behavior in human) of virtual entity by looking at only the user's social reactions. Of course, the presented research has several limits, in particular given the specific context of communication (conversational context, participants, appearances of the virtual agent, etc.). Others experiments with different contexts have to be conducted to validate the generality of the findings. Moreover, other features considering the multi-modal aspects of the communication (as for instance the user's gaze and verbal behavior and her physiological responses, data contained in the existing corpus) as well as other machine learning techniques should be explored.

ACKNOWLEDGMENTS

Research supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and ANR-11-IDEX-0001-02 (A*MIDEX)

REFERENCES

[1] Aran, O. and Gatica-Perez, D. (2013). One of a kind: inferring personality impressions in meetings. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (ACM), 11–18
[2] Aylett, M. P. and Pidcock, C. J. (2007). The cerevoice characterful speech synthesiser sdk. In *IVA*. 413–414
[3] Bailenson, J. N. and Yee, N. (2005). Digital chameleons automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science* 16, 814–819
[4] Beall, A., Bailenson, J., Loomis, J., Blascovich, J., and Rex, C. (2003). Non-zero-sum mutual gaze in collaborative virtual environments. In *Proceedings of HCI international*
[5] Breiman, L. (2001). Random forests. *Machine learning* 45, 5–32
[6] Carney, D. R., Hall, J. A., , and LeBeau, L. (2005). Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior* 29, 105–123
[7] Cassell, J. (2000). More than just another pretty face: Embodied conversational interface agents. *Communications of the ACM* 43, 70–78
[8] Cerekovic, A., Aran, O., and Gatica-Perez, D. (2016). Rapport with virtual agents: What do human social cues and personality explain?
[9] Dehn, D. M. and van Mulken, S. (2000). The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies*

52, 1–22

[10] Ekman, P., Friesen, W. V., and Hager, J. C. (2002). *The facial action coding system* (Weidenfeld and Nicolson)

[11] Fourati, N. and Pelachaud, C. (2015). Relevant body cues for the classification of emotional body expression in daily actions. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on* (IEEE), 267–273

[12] Hecht, M. L., DeVito, J. A., and Guerrero, L. K. (1999). Perspectives on nonverbal communication: Codes, functions, and contexts. In *The nonverbal communication reader* (Waveland Press Lone Grove,, IL). 3–18

[13] Kim, Y., Lee, H., and Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (IEEE), 3687–3691

[14] Kopp, S., Gesellensetter, L., Krämer, N. C., and Wachsmuth, I. (2005). A conversational agent as museum guide–design and evaluation of a real-world application. In *International Workshop on Intelligent Virtual Agents* (Springer), 329–343

[15] Krämer, N., Kopp, S., Becker-Asano, C., and Sommer, N. (2013). Smile and the world will smile with youâĂŤthe effects of a virtual agentâĂŸs smile on usersâĂŹ evaluation and behavior. *International Journal of Human-Computer Studies* 71, 335–349

[16] Krämer, N. C. (2008). Social effects of virtual assistants. a review of empirical results with regard to communication. In *Proceedings of the international conference on Intelligent Virtual Agents (IVA)* (Berlin, Heidelberg: Springer-Verlag), 507–508

[17] Mayer, R. E. and DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied* 18, 239

[18] McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3, 5–17

[19] Mower, E., Feil-Seifer, D. J., Mataric, M. J., and Narayanan, S. (2007). Investigating implicit cues for user state estimation in human-robot interaction using physiological measurements. In *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on* (IEEE), 1125–1130

[20] Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 81–103

[21] Ochs, M., Niewiadmoski, R., and Pelachaud, C. (2010). How a virtual agent should smile? morphological and dynamic characteristics of virtual agent's smiles. In *Proceedings of the international conference on Intelligent Virtual Agents (IVA)* (Springer Berlin Heidelberg), 427–440

[22] Ochs, M., Niewiadomski, R., Brunet, P., and Pelachaud, C. (2012). Smiling virtual agent in social context. *Cognitive Processing, Special Issue on"Social Agents"* 13, 519–532

[23] Oliver, N. M., Rosario, B., and Pentland, A. P. (2000). A bayesian computer vision system for modeling human interactions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22, 831–843

[24] Pardo, D., Mencia, B. L., Trapote, Á. H., and Hernández, L. (2009). Non-verbal communication strategies to improve robustness in dialogue systems: a comparative study. *Journal on Multimodal User Interfaces* 3, 285–297

[25] Pelachaud, C. (2009). Studies on gesture expressivity for a virtual agent. *Speech Communication* 51, 630–639

[26] Rashotte, L. S. (2002). What does that smile mean? the meaning of nonverbal behaviors in social interaction. *Social Psychology Quarterly* , 92–102

[27] Reeves, B. and Nass, C. (1996). *How people treat computers, television, and new media like real people and places* (CSLI Publications and Cambridge university press Cambridge, UK)

[28] Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics* 9, 1

[29] Vinciarelli, A. (2007). Speakers Role Recognition in Multiparty Audio Recordings Using Social Network Analysis and Duration Distribution Modeling. *IEEE Transactions on Multimedia* 9, 1215–1226. doi:10.1109/TMM.2007.902882

[30] Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27, 1743–1759

[31] Vinciarelli, A. and Pentland, A. S. (2015). New social signals in a new interaction world: the next frontier for social signal processing. *IEEE Systems, Man, and Cybernetics Magazine* 1, 10–17

[32] Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 532–539

[33] Yu, Z., Gerritsen, D., Ogan, A., Black, A. W., and Cassell, J. (2013). Automatic prediction of friendship via multi-model dyadic features. In *Proceedings of SIGDIAL*. 51–60