Multimodal behavioral cues analysis of the sense of presence and co-presence during a social interaction with a virtual patient

Magalie Ochs, Jérémie Bousquet, Jean-Marie Pergandi, Philippe Blache

frontiers

Multimodal behavioral cues analysis of the sense of presence and co-presence during a social interaction with a virtual patient

Magalie Ochs [1], **Jérémie Bousquet** [1,3], **Jean-Marie Pergandi** [2] **and Philippe Blache** [3]

[1] *Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France*
[2] *Aix Marseille Univ, CNRS, ISM, CRVM, Marseille, France*
[3] *Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France*

Correspondence*:
Magalie Ochs
magalie.ochs@lis-lab.fr

ABSTRACT

A key challenge when studying human-agent interaction, is the evaluation of user's experience. In virtual reality, this question is addressed through the study of the sense of *presence* and *co-presence*, generally assessed thanks to well-grounded subjective post-experience questionnaires. In this article, we aim at correlating objective multimodal cues produced by users to their subjective sense of presence and co-presence. Our study is based on a human-agent interaction corpus collected in task-oriented context: a virtual environment aiming at training doctors to break bad news to a patient played by a virtual agent. Based on a corpus study, we have used machine learning approaches to explore the possibility of automatically predicting the sense of presence and co-presence of the user thanks to specific multimodal behavioral cues. The performance of random forests models demonstrates the capacity to automatically and accurately predict the level of presence. It also shows the relevance of a multimodal model, based on verbal and non-verbal behavioral cues as objective measures of presence.

Keywords: Multimodal social signals, Sense of Presence, Virtual Reality, Conversational agent, Virtual Patient

1 INTRODUCTION

A key challenge when studying human-agent interaction, is the evaluation of user's experience. Most of existing methods relies on subjective evaluations based on questionnaires filled by the users after their interaction with the virtual agent Grassini and Laumann (2020); **?**); Bailenson et al. (2005); Witmer and Singer (1998); Usoh et al. (2000). Such questionnaires assess the user's perception of the virtual agent of the task, of the virtual environment, her global satisfaction, engagement, etc.

In the virtual reality domain, user's experience is usually evaluated through the measure of the *sense of presence* (the feeling of being present in the virtual environment), which can be correlated with the level of *immersion* in a virtual environment. In the literature, two types of immersion are distinguished: (1) *technological* and *physical* immersion Cadoz (1994) rendered possible by the device (for example a 360 degrees view); and (2) psychological immersion Slater et al. (1996) which is independent from the device (a book, projecting us in a virtual world, can provoke a psychological immersion without any technological and physical immersion). *Sense of presence* corresponds to this second type of immersion (close to the concept of *flow* Csikszentmihalyi (2014) making the user losing the notion of time and space). A second notion, called *sense of co-presence* (also commonly designated as *social presence*) is introduced when the virtual environment is populated by virtual agent or avatars. Co-presence corresponds to "the sense of

being and acting with others in a virtual space" Slater et al. (2006)[1].

The sense of (co-)presence is particularly important in the context of user's *training in virtual reality environments*. In this article, we particularly focus of a specific application domain: a virtual reality platform to train doctors with virtual patients. The goal of this platform is to develop doctors' social skills for their interaction with patients. Such skills are of deep importance. For instance, the way doctors deliver bad news related to damage associated with care has a significant impact on the therapeutic process: disease evolution, adherence with treatment recommendations, litigation possibilities, among others Andrade et al. (2010). In order to facilitate doctor's training, we have developed a virtual patient able to interact naturally in a multimodal way with doctors simulating breaking bad news to the patient (for more details on the platform see Ochs et al. (2017). In this paper, we investigate the multimodal behavior cues of (co-)presence of users training to break bad news to a virtual patient.

The problem in the evaluation of presence and co-presence with questionnaires, in spite of their interest, is the subjectivity of the approach (consisting in asking users to self-report their feeling). Previous works have tried to find *objective measures* by hypothesizing that different levels of the sense of presence and co-presence may be connected with different verbal and non-verbal user's behaviors Laarni et al. (2015); Ijsselsteijn (2002). However, only few behavioral cues have been investigated. We propose in this work to take into account a large range of modalities (both verbal and non-verbal) by involving the notion *engagement* (considered as a form of involvement) in the description of the sense of (co-)presence. This idea relies on several observations. First, as shown in Schroeder (2002), the sense of presence and co-presence can be correlated with the level of *immersion*. In such case, the greater the immersion, the higher the feeling of (co)presence. Second, the notion of *involvement* also plays an important role besides immersion Witmer and Singer (1998): the sense of presence increases when participants become more involved in the virtual environment.

Starting with this hypothesis of multimodal behavioral cues of (co-)presence, we investigate the possibility to automatically predict the sense of (co-)presence based on user's multimodal behavior during an interaction with a virtual agent. In this perspective, we have collected a corpus of human-agent interaction in a virtual reality environment. This has been done thanks to specific tools automatically acquiring verbal and non-verbal user's productions. Moreover, we have collected questionnaires indicating the user's sense of presence and co-presence after the interaction. In order to be independent from the environment, our experimental setup involves different virtual reality displays - known to generate different degrees of immersion. Based on machine learning techniques, we have learned a model to correlate verbal and non-verbal cues to different levels of presence and co-presence. The accuracy of the model shows that certain verbal and non-verbal cues of the user's behavior can be used to predict her level of presence and co-presence, based on objective behavioral measures.

The paper is organized as follows. In the next section, we present the theoretical background and related works on the notion of presence and co-presence. In Section 3, we introduce the human-virtual patient interaction corpus collected with different virtual reality displays. Section 4 is dedicated to the pre-processing of the collected data in order to automatically extract relevant verbal and non-verbal behavioral cues that may be used to predict the sense of presence. In Section 5, we present the model learned on the human-virtual patient interaction corpus, with the extracted verbal and non-verbal behavioral cues exploited as features and, the levels of presence and co-presence clustered to classes to predict. We conclude and discuss perspectives Section 6.

2 THE SENSE OF PRESENCE AND CO-PRESENCE

2.1 Definition of the sense of presence

Our definition of presence relies on the notion of immersion, that can be defined in two different ways. First, it can be considered in terms of psychological state as the perception of *being in*, *to be surrounded by* Witmer and Singer (1998). In this case, immersion includes the insulation from the physical environment,

[1] Note that no consensus exists on the notion of co-presence. A detailed discussion on the different definitions can be found in Bailenson et al. (2005).

the perception of a feeling of being *included* in the virtual environment, the natural state of the interaction, a perception of control and the perception of movement in a virtual environment. A second type of definition considers immersion in technological terms, immersion being correlated to technology Bystrom et al. (1999); Draper et al. (1998); Slater and Wilbur (1997). We adopt in our work the first perspective Witmer and Singer (1998).

Several parameters involved in the definition of the sense of presence are described in the literature: (1) *the ease of interaction*: interaction correlates with the sense of presence felt in the virtual environment Billinghurst and Weghorst (1995); (2) *the user control*: the sense of presence increases with the sense of control Witmer and Singer (1998); (3) *the realism of the image*: the more realistic virtual environment is, the more the sense of presence is strong Witmer and Singer (1998); (4) the *duration of the exhibition*: prolonged exposure beyond 15 minutes with the virtual environment does not give the best result for the sense of presence with HMD (*Head Mounted Display*) and there is even a negative correlation between the prolonged exposure in the virtual environment and the sense of presence Witmer and Singer (1998); (5) the *social presence and social presence factors*: the social presence of other individuals (real or avatars), and the ability to interact with these individuals increases the sense of presence Heeter (1992); (6) the *the quality of the virtual environment*: quality, realism, the ability of the environment to be fluid, to create interaction are key factors in the sense of presence of the user Hendrix and Barfield (1996). Two other factors are more particularly related to the individual perception, and contextual and psychological factors that should be taken into account during the evaluation of presence Mestre (2015). In the next section, we introduce the different questionnaires available to measure these factors.

2.2 Questionnaires of presence and co-presence

Several questionnaires have been proposed in order to assess the sense of presence (see Grassini and Laumann (2020) for a review). Four of them are "canonical", they have been used in many different works and are statistically significant: the canonical presence test of Witmer and Singer Witmer and Singer (1998), the ITC-SOPI canonical test Lessiter et al. (2001) that evaluates the psychological immersion, the Slater-Usoh-Steed (SUS) questionnaire to evaluate the spatial presence, and the canonical test IGroup Presence Questionnaire (IPQ) Schubert et al. (2001). We used the last one in our work to evaluate the training system.. This test focuses on three variables dependent on presence factors: spatial presence, involvement in the device, and realism of the device. The test is composed of 14 questions, some of them being taken directly from the Presence Questionnaire Witmer and Singer (1998) and the SUS questionnaire Usoh et al. (2000). In the last version, another variable dependent on the global presence has been added. This test has the advantage to contain few questions (only 14) while including the main presence factors of the other canonical tests.

However, one limit of the IPQ test is the lack of the evaluation of the notion of *co-presence*. Co-presence, also commonly called *social presence*, can be defined as “the sense of being and acting with others in a virtual space” Slater et al. (2006)[2]. In our context, we are interested in evaluating the sense of co-presence of the participants with the virtual agent. In order to evaluate the co-presence, we have used the test proposed in Bailenson et al. (2005) that measures social presence through the following variables: the *perceived co-presence*, the *embarrassment* to measure the social influence of the agent, and the *likability* of the virtual representation. In Bailenson et al. (2005), the authors have shown that this self-report questionnaire is effective “to measure how people perceive an embodied agent”.

2.3 Behavioral measure presence

In order to quantify the sense of presence or co-presence based on reliable parameters, several works tried to identify objectives measures. As highlighted in Slater et al. (1998), we can distinguish “subjective presence” from “behavioral presence”; subjective presence being measured trough presence questionnaire and the behavioral presence corresponding to bodily responds. Three types of objective measures of presence can be distinguished : behavioral (e.g. attention), performance-based (e.g. user’s performance in task realization) and physiological (e.g. brain activity, heart rate) Ijsselsteijn (2002). In this paper, we focus on behavioral measures of presence.

[2] Note that no consensus exists on the notion of co-presence. A detailed discussion on the different definitions can be found in Bailenson et al. (2005)

Some works have studied user's behavior considering the way the user performs specific actions related to the task in the virtual environment. For instance, in Usoh et al. (1999), the authors analyze the navigation path of the users moving towards an object and the correlation with the level of presence. Other works have shown a close relation between body movements (for instance their magnitude) and the sense of presence Slater and Steed (2000); Slater et al. (1998). In Bailenson et al. (2004), the authors have compared social presence self-report measures and the interpersonal distances of the user with virtual agents. The results did not reveal significant correlations between these objective and subjective measures.

Concerning the relation between presence and co-presence, the research works have shown that they generally co-vary: a stronger sense of co-presence comes with a stronger sense of presence Schroeder (2002).

Finally, as underlined in Laarni et al. (2015), none of these works have demonstrated strong evidence of behavioral measures of presence. Moreover, most of the works mainly focus on specific actions related to the context of the task. In this paper, we propose to analyze fine-grained objective behavioral measures of presence by studying verbal and non-verbal behavioral cues.

2.4 Presence, involvement and engagement

In our interdisciplinary approach, we aim at connecting empirical and theoretical backgrounds from different domains around the notion of presence and co-presence. Starting from the definition of these notions in the virtual reality domain, we investigate phenomena that can be observed in human-human and human-machine interaction through multimodal behavioral cues.

As described above Schubert et al. (2001), we consider for our study that the notion of presence covers two different aspects: *involvement* and *psychological immersion* (also called *spatial presence* in Witmer and Singer (1998)): "*Involvement is a psychological state experienced as a consequence of focusing one's energy and attention on a coherent set of stimuli or meaningfully related activities and events ... [Psychological] immersion is a psychological state characterized by perceiving oneself to be enveloped by, included in, and interacting with an environment that provides a continuous stream of stimuli and experiences*" (Witmer and Singer (1998) cited in Schubert et al. (2001)). Note that the terms immersion and presence are often considered as synonyms **?**. In our case, we adopt therefore a broader perspective by including the engagement of the participant.

As for co-presence, the questionnaire considered in the study includes a self-report marker that should reflect the feeling of being with another social entity in the virtual environment, as well as the liking of the virtual agent and the willingness to perform embarrassing acts in front of the virtual agent Bailenson et al. (2005).

Identifying objective cues of the notion of presence remains a difficult task because of the abstract level of definition of this notion. The different questionnaires presented above are based on very high-level notions, that can hardly connect with observable features during an interaction with a virtual agent. We propose in this paper to bridge the gap between presence and observable features by posing an hypothesis: *the senses presence and co-presence are correlated with involvement/engagement*. This hypothesis relies on the idea that the interaction, in particular in a task-oriented context, is more natural, variable and rich when presence and co-presence are high (and vice-versa). Moreover, in a virtual environment, no engagement can be observed without a high level of (co-)presence. If this hypothesis is true, it should be the case that a correlation can be observed between the level of (co-)presence and that of engagement. Concretely, involvement/engagement being possibly assessed based on different objective cues, we propose to use these same features in order to predict the level of (co-)presence.

In the domain of human-machine interaction, and more particularly in the context of interaction with virtual agents or robots, different definitions of engagement have been proposed Glas and Pelachaud (2015). For instance, as described in Glas and Pelachaud (2015), *face engagement* characterized by the "*maintaining of a single focus of cognitive and visual attention*" of the user and the artificial entity during a joint activity, the face engagement being reflected by eye-contact, gaze and facial gestures to interact with each other Le Maitre and Chetouani (2013). A common definition of engagement in human-machine interaction is the one proposed by Sidner and Dzikovska Sidner and Dzikovska (2002) that consider the engagement as a process "*by which two (or more) participants establish, maintain and end their perceived connection*". Some authors have defined engagement as a specific mental state of the participant that has the

goal to be and interact with the other Poggi (2007). Some definition link directly the notion of engagement to the notion of interest and attention Yu et al. (2004) or involvement Bickmore et al. (2010). As pointed in Bickmore et al. (2010), the notion of engagement in a short term interaction, is also tightly related to the notion of "rapport" Gratch et al. (2007) characterizing by positive emotions, mutual attentiveness, and coordination Tickle-Degnen and Rosenthal (1990) and the notion of "flow" Csikszentmihalyi (2014).

2.5 Multimodal cues of presence, involvement, and engagement

Involvement in face-to-face conversations is classically measured by nonverbal cues such as gaze or body orientation. However, more indicators of engagement have been identified in collaborative activities, concerning verbal aspects (e.g. prosody, questioning, comments, explanations, etc.) as well as gestures and facial expressions Helme and Clarke (2001).

As for verbal indicators, several works have addressed the question of the type of lexical, syntactic and semantic aspects that can be related with engagement/involvement. In this perspective, different features has been identified: number of intensifiers vs. qualifier words, number of personal vs. impersonal pronouns, number of definite vs. indefinite articles: these ratios increases as a speaker becomes more cognitively involved Camden and Verba (1986); Nguyen and Fussell (2016). At a higher level, the complexity of the syntactic structure also enters into consideration: the richness of the structure is correlated with the level of engagement of the speaker and how it affects the perceived credibility of a message Tolochko and Boomgaarden (2018): when speakers feel engaged, they speak more, using richer and more variable constructions. This information (that we call in our model syntactic complexity) corresponds to the number of clauses in the utterance which can be approximated with the type of their constituents. Typically, a clause is usually built around a verb. The number of verbs (and also other types of constituents such as conjunctions) can then give an approximation of the number of clauses and then the richness of the syntactic structure Brown et al. (2008); Biber et al. (2016). The technique simply consists in counting the amount of such categories, connected to the realization of different clauses. We complement this approximation with lower-level features also providing indication on the sentence complexity such as the number of words, of modifiers (giving an indication of the semantic richness) in a sentence. Finally, based on the research works presented above, concerning the verbal behavioral cues, in this article, we consider these different features: *lexical richness, discourse elaboration, semantic richness* and *syntactic complexity*.

Concerning non-verbal cues, several works underlines the relationship between engagement and non-verbal behavioral cues. For instance, in their theory on rapport Tickle-Degnen and Rosenthal (1990), the authors argued that the rapport (engagement) between the participants of an interaction is traduced by the head nods, the smiles, the posture mimicry and the gestures coordination. As highlighted in Sidner and Lee (2007), "engagement behavior" include head nods and gaze during human-robot interaction. In Sanghvi et al. (2011), the authors have shown the importance of the quantity of movements to recognize engagement during a human-machine interaction. In this article, based on the research presented above, concerning the non-verbal cues, we consider the *movements of the head and the body* of *both* participants (the user and the virtual patient).

Finally, we aim at analyzing these different multimodal cues by trying to correlate these cues of engagement to (co-)presence.

3 COLLECTION OF HUMAN-VIRTUAL PATIENT INTERACTIONS IN VIRTUAL REALITY ENVIRONMENTS

In order to analyze the multimodal cues of (co-)presence, we have collected a corpus of human-virtual patient interaction thanks to a virtual reality platform we have developed for training doctors to break bad news Ochs et al. (2017). We present in the following the details of the corpus.

3.1 A virtual reality platform for training to break bad new

The corpus has been collected through different *virtual reality environments*. This platform makes it possible for the user (the doctor) to interact with a virtual patient in natural language. The virtual agent has been endowed with a dialog system and a non-verbal behavior model based on a human-human corpus analysis of real interactions with standardized patients Ochs et al. (2017).

The environment has been designed to simulate a real recovery room where breaking bad news are generally performed. Technically, the virtual agent is based on the VIB platform Pelachaud (2009) and integrated in a *Unity* player. Participants were filmed and body motions digitally recorded from the passive reflective markers placed on head (stereo glasses), elbows and wrists. A high-end microphone synchronously recorded the participant's and virtual agent verbal expressions from the Unity player. This environment facilitates the collection of the corpus of human-agent interaction in order to analyze the verbal and non-verbal behavior in different immersive environments.

3.2 Participants

In total, 38 persons (28 males, 10 females) with a mean age of 29 years (SD:10.5) volunteered to participate to the experimentation. 25 participants have been recruited at the University, 13 others are real doctors recruited in a medical institution. These participants had already have an experience in breaking bad news with real patients. The participants were not paid.

3.3 The collect of the human-machine interaction corpus

A specific methodology has been implemented in order to collect the interaction and create this corpus of human-machine interaction.

3.3.0.1 Procedure

When participants arrived at the laboratory, an experimenter sat them down and presented them the instructions before the interaction. Participants are asked to read the instructions several times as well as before each interaction. The understanding of these instructions was checked by means of an oral questionnaire.

3.3.0.2 Task

Participants were instructed that the role they have to play is a doctor that had just (i.e., immediate post operative period) operated the virtual patient by gastroenterologic endoscopy to remove a polyp in the bowel. During the surgery, a digestive perforation occurred[3]. Participants were accurately instructed about the causes of the problem, the effects (pain), and the proposed remediation (a new surgery, urgently). The participants' task was to announce this medical situation to the virtual patient.

3.3.0.3 Type of immersive devices

In order to collect data with different levels of immersion, we have implemented the virtual patient on different virtual reality displays: PC monitor, virtual reality headset (HMD), and virtual reality room (Figure fig.1). The virtual reality cave is constituted of a 3m deep, 3m wide, and 4m high cubic space with three vertical screens and a horizontal screen (floor). A cluster of graphics machine makes it possible to deliver stereoscopic, wide-field, real-time rendering of 3D environments, including spatial sound. This offers an optimal sensorial immersion of the user.

Figure 1. Participants interacting with the virtual patient with different virtual environment displays (from left to right): virtual reality headset (HMD), virtual reality room (CAVE), and PC monitor.

[3] The scenario has been carefully chosen with the medical partners of the project for several reasons (e.g. the panel of resulting damages, the difficulty of the announcement, its standard characteristics of announce).

The order of presentation of each display modality was counterbalanced withing participants of each group. Each participant has interacted with the systems 3 times with three different displays: PC monitor, virtual reality headset (HMD), and virtual reality room (CAVE). Note that we counterbalanced the order of these of each display in order to avoid an effect of the order on the results. The duration of each interaction is in average 3mn16.

The visualization of the interaction, is done through a 3D video playback player we have developed (Figure 2). This player replays synchronously the animation and verbal expression of the virtual agent as well as the movements (based on the head, elbows and wrists body trackers) and video of the participant.

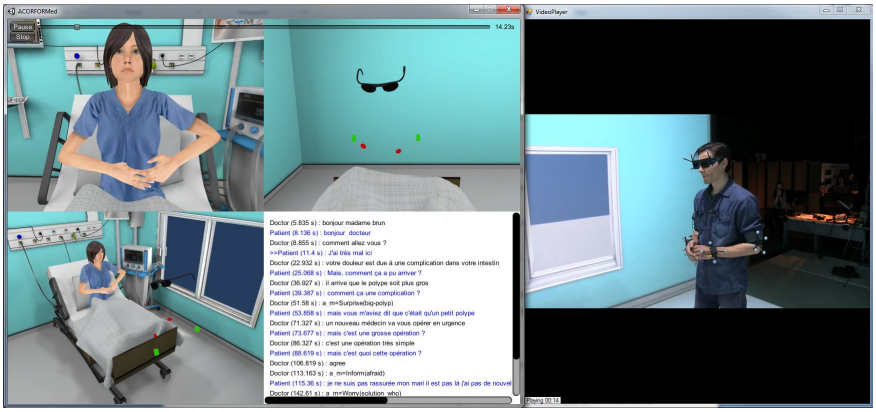


Figure 2. 3D video playback player

3.3.0.4 Subjective assessment of presence

Participants' subjective experience was assessed through two separate post-experience questionnaires (1-5 range) measuring their sense of presence (with the *IGroup Presence Questionnaire*, IPQ Schubert (2003) and their sense of c-presence Bailenson et al. (2005). The questionnaires are described in more details Section 2.2.

To sum up, the corpus contains the following raw data:

- a video of the participant during her interaction with the agent in the three environments: a virtual reality room (CAVE), a virtual reality headset (HMD), and a PC monitor;
- time-series three-dimensional unity coordinates of 5 trackers located on the participant's head, left and right elbows, and left and right wrists during the interaction;
- an audio file from a mic pinned to the participant during the interaction and hence containing only the voice of the participant. The audio file has been transcript from an automatic speech recognition system.

In total, the data contains 114 human-agent interactions. However, due to technical recording problems, some interactions have not be integrated in the corpus. Finally, the corpus is composed of 86 human-agent interactions. In the machine learning point of view, in order to reduce the number of features, we have processed this data to compute relevant verbal and non-verbal behavioral cues. We present these features in the following.

Given the relative small size of data-set, we consider an *early fusion* approach Snoek et al. (2005): data from each unitary modality is processed in order to compute a certain number of features. These features are merely concatenated together to form our data-set that corresponds to a matrix that will fed to learning algorithms. Another advantage of the "early fusion" is that the resulting model will be interpretable with a analysis of the relative importance of the designed features.

4 AUTOMATIC EXTRACTION OF VERBAL AND NON-VERBAL CUES

In order to investigate the users' multimodal behaviors during the interactions with the virtual patient, we have extracted, from the corpus described above, different verbal and non-verbal cues.

4.1 Verbal behavior

Using a specific tool called SPPAS Bigi (2012), a tokenization followed by a phonetization on the transcription file was performed. Participants' verbal expression were assessed by processing the transcript text to recover the following dependant variables. For this sake, the transcript text was then parsed by the Marsatag tool Rauzy et al. (2014), a stochastic parser for written French which has been adapted to account for the specificities of spoken French. Among other outputs, it provides a morpho-syntactic category for each POS token.

4.1.0.1 Features characterizing lexical richness and linguistic complexity.

The user's verbal behavior was firstly assessed by computing the frequency of the part-of-speech (POS) tags. The POS tags were automatically identified using MarsaTag. Nine POS tags were considered: adjective, adverb, auxiliary, conjunction, determiner, noun, preposition, pronoun, verb. Two high-level features characterizing the considered POS tags were measured. The lexical richness was measured as the fraction of adjectives and adverbs out of the total number of tokens as follows: $\frac{nb\_adj+nb\_adv}{\sum tokens}$. The lexical complexity was measured as the fraction of conjunctions, prepositions and pronouns out of the total number of tokens as follows: $\frac{nb\_conj+nb\_prep+nb\_pro}{\sum tokens}$.

4.1.0.2 Length of sentences.

The user's verbal behavior was secondly assessed by computing the length of each sentence, measured as the number of words composing it, being defined from the transcript text by the MarsaTag tool Rauzy et al. (2014).

4.1.0.3 Lengths of inter-pausal units.

The user's verbal behavior was thirdly assessed by computing the length of inter-pausal units (expressed in duration). For this sake, the speech signal was automatically segmented using SPASS Bigi (2012) into Inter-Pausal Units (IPUs), defined as speech blocks surrounded by at least 200 ms silent pauses[4].

4.1.0.4 Answering time.

The user's verbal behavior was also assessed by computing the average answering time expressed in seconds. Considering the interactions as dialogues between two speakers, the answering time corresponds to the period of time between the end of the first speaker speech, and the beginning of the second speaker speech (the speakers could be the doctor or the virtual patient).

4.2 Non-verbal behavior

Following the method proposed in Slater et al. (1998), the body movements considered in this study are the rotation of the arms and the head. More precisely, for each interaction, we first compute difference between each successive rotation angle[5] (difference between rotation angle on one of the three axis at time t and the same at $t - \delta t$, δt being time interval used to record data), around the X, Y and Z axis (pitch, yaw and roll respectively). We perform this for the head, the left and right wrists, and the left and right elbows.

We then compute the averages and standard deviations for each of these 5 body parts, and for each of the 3 axis, to obtain 2×15 values. The values related to the 4 body parts (left and right, wrists and elbows) are then averaged, so we have mean and standard deviation for head and for upper limbs, for the 3 axis

[4] For French language, lowering this 200 ms threshold would lead to many more errors due to the confusion of pause with the closure part of unvoiced consonants, or with constrictives produced with a very low energy.

[5] Using rotations is coherent with the behaviour of our virtual patient, which, lying in bed, does not move much, but sometimes rotates its head or arms.

(12 values). We then average over the 3 axis, and gather the features of the upper limbs, to obtain finally 4 features representing the averages and standard deviations of the rotation of the head and of the arms

The verbal and non-verbal features are computed for the user as well as for the virtual patient.

4.3 Interactional cues

Besides the behavioral cues, we have considered specific features related to the interaction that may provide cues on the level of (co-)presence: the total duration of the interaction and the expertise of the participant (expert in the case of a doctor and non-expert otherwise).

To summarize, each user-virtual patient interaction is characterized by the following features:

- *total duration of the interaction* represented by one continuous value in seconds;
- *expertise of the participant* represented by a binary categorical variable representing whether the participant is an expert (doctor) or a non-expert;
- *rotations of the head and arms* represented by 4 continuous variables (mean of the rotation of the head, standard deviation of the rotation of the head, mean of the rotation of the arms, and standard deviation of the rotation of the arms);
- *average sentence length in terms of number of words* characterized by a continuous variable;
- *average length of Inter-Pausal Units in seconds* represented by a continuous variable;
- *lexical richness* represented by a continuous variable,
- *linguistic complexity* represented by a continuous variable,
- *answering time* represented by one value.

Considering the segmentation of the interaction and the behavior of both participants (user and virtual agent), the collected data is represented by a matrix of 86 lines (one per interaction) and 20 columns (one per feature, considering the verbal and non-verbal cues of the user and of the virtual agent).

In the next section, the matrix is used to learn a model to automatically predict the sense of presence and co-presence of the participants.Note that a statistical analysis of the effects of the virtual reality displays and of the type of the participant (doctors versus novices) on the behavior displayed and on the sense of presence and co-presence is described in details in Ochs et al. (2018). In this paper, we focus on the automatic prediction of the sense of presence and co-presence by considering the type of participant and their verbal and non-verbal behavior as key features. The goal of the work presented in this article is not to predict the different interaction modes (PC monitor, virtual reality headset, or virtual reality room), but the levels of presence and co-presence. We have shown in Ochs et al. (2018) that the three interaction modes imply different levels of presence and co-presence.

5 AUTOMATIC PREDICTION OF THE SENSE OF PRESENCE BASED ON MULTIMODAL CUES

Our goal is to predict users' sense of *presence* and *co-presence* based on objectives measures. In our context, we consider *two classification problems* making it possible to predict

1. the level of the sense of presence ;
2. the level of the sense of co-presence.

The same features, described in the previous section, are used to learn both models. For each interaction, the sense of presence and co-presence have been assessed through two questionnaires. The resulting values are integers from 1 to 5. Our objective is to experiment tasks of prediction of sense of presence on one side, and of co-presence on another side, using selected machine learning algorithms. Practically, we compared three machine learning techniques: *Naives Bayes*, *Support Vector Machine* and *Random Forest*. These methods, among the best classifiers Fernández-Delgado et al. (2014), have the advantage, compared with other statistical models such as RNN, to handle high-dimensional data with a high generalization power Strobl et al. (2008); Forman and Cohen (2004); Salperwyck and Lemaire (2011). They are also well suited for handling small datasets.

5.1 Classifiers' training and test procedure

Figure fig:double-cv illustrates the process, based on a double cross-validation.

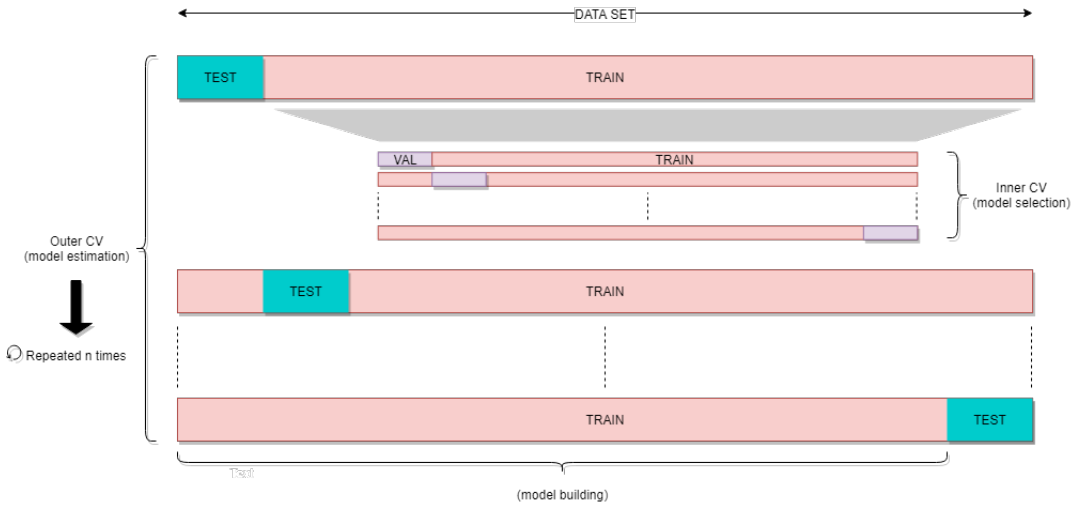


Figure 3. Double cross-validation

The data-set is split into training and test data, each subset created with respect to frequencies of classes to account for class imbalance. We use 10% of data-set as test data. The best hyper-parameters for concerned machine learning algorithm are searched through k-folds cross-validation (with $k = 5$[6]) on the training data subset ('validation' metrics are computed at this stage, in order to estimate and select the best hyper-parameters combination). The classifier configured with the best hyper-parameters is then fitted to the 90% of training data subset, and used as predictor on the 10% test set initially left aside, which has never been 'seen' by the classifier, to obtain 'train' and 'test' metrics. Given the size of the data-set, we may expect a high variance on test scores obtained with this strategy. In order to estimate the variance, we iterate the process on multiple runs (on several random splits of 90% train and 10% test). This outer 10-folds cross-validation is repeated 20 times.

Concerning the random forest algorithm, in order to minimize the generalization error to avoid over-fitting Breiman (2001), we have evaluated beforehand the optimal number of decision trees on the prediction task by considering the performance of the classifiers and the out-of-bag (OOB) estimated accuracy expected to provide a relevant cue on generalization performances of the RF. Based on the results, we used 150 trees (few improvements is observed with a larger number of trees).

As commonly used, we have computed three measures to evaluate the quality of prediction of a model: precision, recall and F1 Score. Note that we compute the weighted metrics to consider the number of instances of each class (i.e. the score of each class is weighted by the number of samples from that class).

In order to estimate the performances of the different classifiers, we compute scores from a classifier returning random predictions, to establish a baseline. We consider three different strategies: `uniform` (generates predictions uniformly at random), `stratified` (generates predictions with respect to the training set's class distribution) and `most frequent` (always predicts the most frequent class in the training set). For each fold of outer cross-validation, random classifier is fitted on the training set and used to generate predictions on the test set, for each strategy. The random classifier final scores are the averages of the scores from the strategy leading to the highest performances.

[6] We consider a small k for this cross-validation to reduce risk of over-fitting as recommended in Baumann and Baumann (2014)

5.2 Identification of the best classifier with the best granularity level of presence and co-presence

The first question to approach the prediction task as a binary or multi-classes problem is the number of classes. In other words, we had to define the level of granularity of presence and co-presence that we can predict. Indeed, the level of presence and co-presence rated by the subjects and associated to each interaction are integers between 0 and 5. Consequently, we can either consider that each value constitute a class (5 classes to predict) or to cluster close values (as for instance the 0 and 1 level to represent a low class of presence of co-presence, 3 for a medium class, and the 4 and 5 to represent high value of presence and co-presence). We explore different clustering algorithms for this discretization task in order to identify the best clusters leading to the best prediction. Discretization parameters are the *number of classes*, between 2 (binary classification) and 5, and the *discretization strategy*: using `kmeans`, values are clustered in order to create as many clusters as the desired number of classes, with `quantile` (all intervals contain the same number of points), and with `uniform` (all intervals have same width). The distribution of the scores of presence and co-presence on the data-set is illustrated Figure fig:discretized-values.

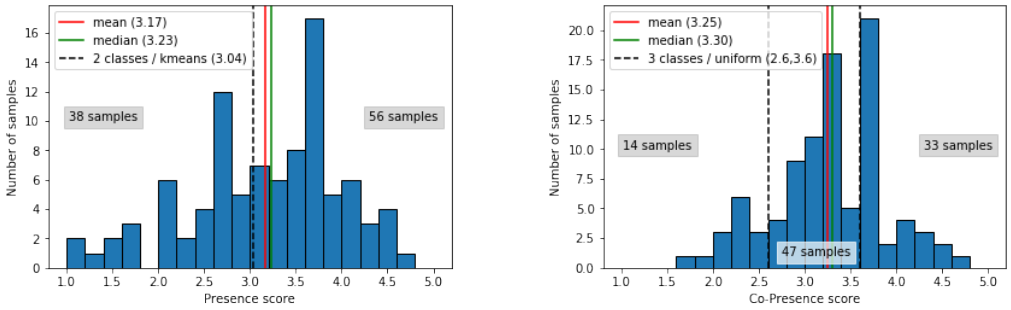


Figure 4. Distribution of the scores of presence and co-presence in the data-set

Our objective is to then limit our experiments to the best found classifier, and to the best discretization. The results show that the best classifiers is the random forests (compared to Naïves Bayes and SVM) both for the prediction of presence and for co-presence. We illustrate the test scores of this classifier Figure fig:discretisationRF. The error bars in the graphics represent the 95% confidence intervals for each measured score. The scores obtained with the random classifier are displayed in transparent gray on the figures.

The best results for presence are obtained with a discretization in 2 classes with the k-means strategy, and for co-presence into 3 classes with uniform strategy. Note that to identify the best discretization, we have compared the results of the random classifier to the results of random forest to optimize the scores of the random forest but also the gap with the scores of the random classifier. The selected discretizations for the score of presence and co-presence are illustrated Figure fig:discretized-values with the vertical dotted lines.

The performance measures, considering all the features described above, reveal an accurate capacity of the model to predict the sense of presence of the user based on multimodal cues with a macro F1-measure closed to 0.8. However, the co-presence seems more difficult to predict with scores closed to 0.5. This lower performance for the co-presence may be explained by the multi-classes classification task (3 classes to predict) whereas the presence is a binary class classification task (2 classes to predict). Note, however, that the scores of co-presence is significantly higher than the baseline (in gray on the figures).

Given the obtained results, we cluster the scores of presence into two classes: *low* or *high* sense of presence; and the scores of presence in three classes: *low*, *medium* or *high* sense of co-presence (as illustrated Figure fig:discretized-values).

5.3 Exploring over-sampling methods to face small data-set

Given the size of the data-set, we have explored different over-sampling methods to increase the amount of data. The over-sampling methods generate new samples of the minority class(es) based on the existing

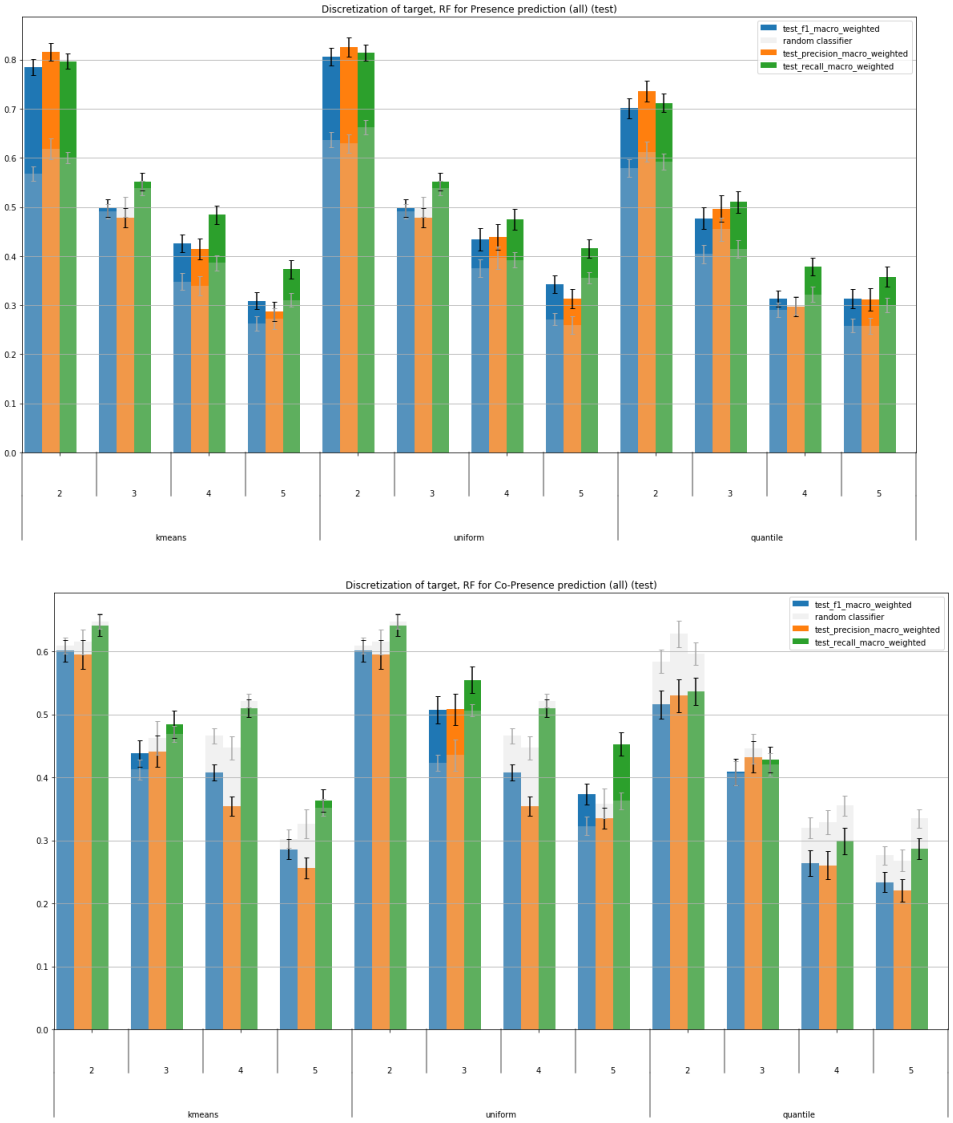


Figure 5. Test scores of the random forest considering different discretization strategies

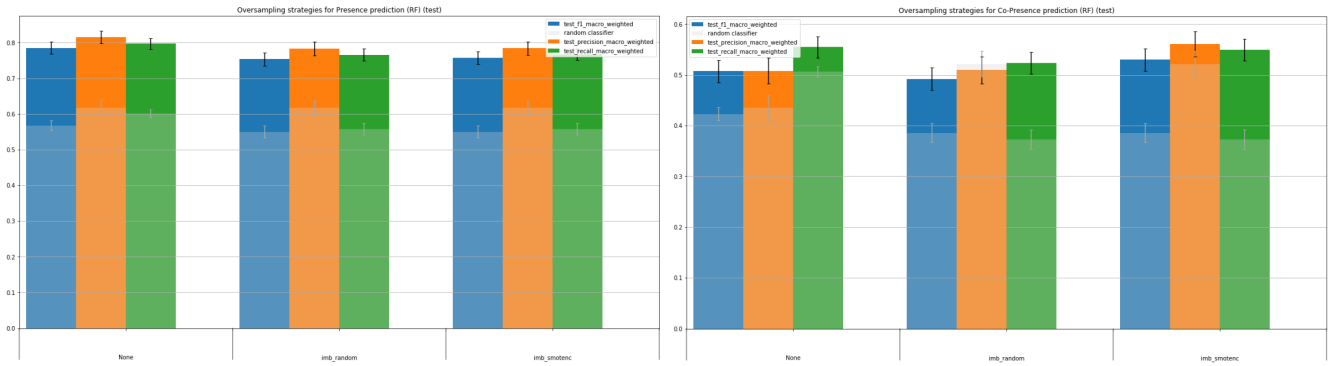
data-set, in order to remove class imbalance. Our goal is to explore whether such methods improve the classifier's performances. We compare two different over-sampling methods:

- *random over-sampling* : samples randomly chosen from the minority class(es) are duplicated;
- SMOTE[7] : new samples are generated by interpolation from a sample randomly chosen from minority class(es) and another sample close to it (randomly selected from k-nearest-neighbors with $k = 3$). Distance of this new sample from existing ones is also random. We use variant SMOTE-NC[8] as it handles categorical variables (as it is not possible to interpolate them, the algorithm chooses most frequent category among nearest neighbours).

The results (illustrated Figure fig:oversampling) show that over-sampling our data-set with these techniques has no influence on the prediction of sense of presence. However, for the prediction of co-presence, SMOTE improves the F_1 score. Consequently, we apply SMOTE for the co-presence classification task.

[7] Synthetic Minority Over-sampling Technique, we use the imbalanced-learn implementation https://imbalanced-learn.readthedocs.io/en/stable/api.html

[8] SMOTE for Nominal and Continuous



(a) presence (c) co-presence

Figure 6. Test scores of the random forest classifier considering different over-sampling strategies or none.

5.4 Verbal and non-verbal behavioral cues importance to predict level of presence and co-presence

In this section, we analyze the importance of the behavioral cues to predict presence and co-presence. The models were configured with respect to findings from the preliminary studies presented above (hyper-parameters search spaces, discretization parameters for presence and co-presence). We consider the Random Forest classifier and the random classifiers as baseline. We focus on test scores which are the best estimation of the generalization capabilities of the models.

In order to analyze the importance of each modality, we consider three sets of features (the features are described in details Section 4)[9]:

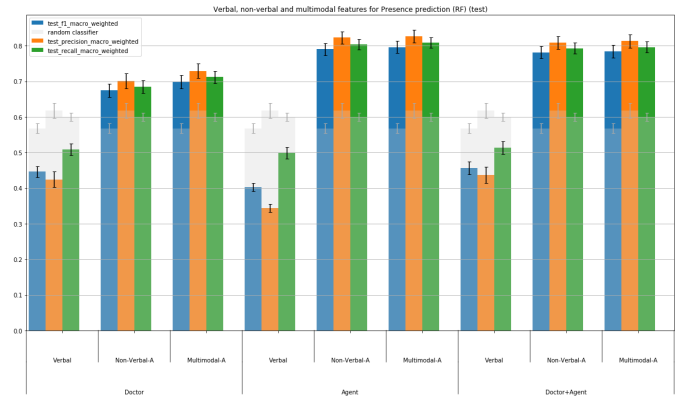
1. *only verbal features*: average sentence length in terms of number of words, average length of *Inter-Pausal Units* in seconds, lexical richness, linguistic complexity, and average answering times;
2. *only non-verbal features*: averages and standard deviations of the rotations of head and arms movements
3. *multimodal features*: the verbal and non-verbal features.

The results are reported Figure fig:modes-test. We consider separately the virtual patient's behavior (condition "Agent") and the user's behavior (condition "Doctor"). In the condition "Doctor+Agent", we consider the behavioral cues of both the virtual patient and the user.

Considering only the "*doctor+agent*" condition (in which both user's and virtual patient's behaviors are considered), the results show the importance of multimodality to predict presence and co-presence. More precisely, taking into consideration the verbal features alone, the scores are not better than a random classification. With the multimodal features, the model can predict with a good score the level of presence of the participant. The scores for co-presence are lower that for presence, which confirms the difficulty to predict the sense of co-presence (that may be explained by the multi-classes classification task compared to the binary classification task for the presence). Note that the non-verbal features provide similar scores as for multimodal features for the prediction of presence and slightly lower score for co-presence. This results show the importance of the non-verbal behavioral cues in the prediction of (co-)presence.

We have compared the importance of the behavior of each participant to the interaction to predict (co-)presence: the *user* (noticed "doctor" on the Figure fig:modes-test) and the *virtual patient* (noticed "Agent" on the Figure fig:modes-test). The results show the importance of the user's behavior for the prediction of presence. Considering only the behavior of the virtual patient or both of them do not lead to better results. Concerning co-presence, it appears that the behavior of the user and the virtual patient have to be considered, the condition "doctor+agent" leading to the best results.

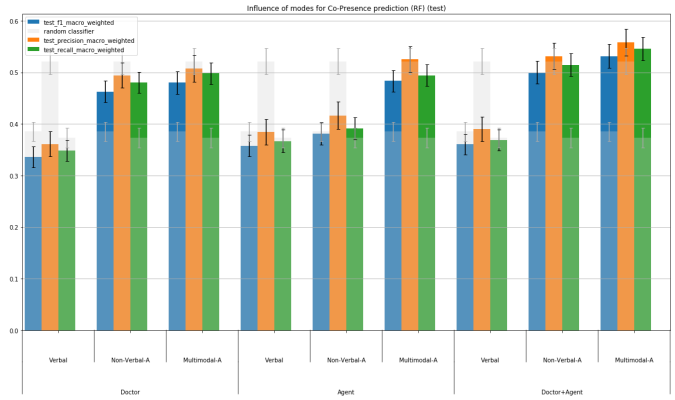
[9] Note that in these groups of features there are no features considered as neither verbal nor non-verbal, like duration of interaction or expertise of participant



(7e) presence

Subject	Mode	f1	precision	recall
Agent	Verbal	$0.45_{\pm 0.02}$	$0.42_{\pm 0.02}$	$0.51_{\pm 0.02}$
	Non-Verbal	$0.67_{\pm 0.02}$	$0.70_{\pm 0.02}$	$0.69_{\pm 0.02}$
	Multimodal	$0.70_{\pm 0.02}$	$0.73_{\pm 0.02}$	$0.71_{\pm 0.02}$
Doctor	Verbal	$0.40_{\pm 0.01}$	$0.34_{\pm 0.01}$	$0.50_{\pm 0.02}$
	Non Verbal	$0.79_{\pm 0.02}$	$0.82_{\pm 0.02}$	$0.80_{\pm 0.01}$
	Mulimodal	$0.80_{\pm 0.02}$	$0.83_{\pm 0.02}$	$0.81_{\pm 0.02}$
Doctor+Agent	Verbal	$0.46_{\pm 0.02}$	$0.44_{\pm 0.02}$	$0.51_{\pm 0.02}$
	Non Verbal	$0.78_{\pm 0.02}$	$0.81_{\pm 0.02}$	$0.79_{\pm 0.02}$
	Mulimodal	$0.78_{\pm 0.02}$	$0.81_{\pm 0.02}$	$0.80_{\pm 0.02}$
Random classifier		$0.57_{\pm 0.02}$	$0.62_{\pm 0.02}$	$0.6_{\pm 0.02}$

(7f) presence



(7g) co-presence

Subject	Mode	f1	precision	recall
Agent	Verbal	$0.34_{\pm 0.02}$	$0.36_{\pm 0.02}$	$0.35_{\pm 0.02}$
	Non-Verbal	$0.46_{\pm 0.02}$	$0.49_{\pm 0.02}$	$0.48_{\pm 0.02}$
	Mulimodal	$0.48_{\pm 0.02}$	$0.51_{\pm 0.03}$	$0.50_{\pm 0.02}$
Doctor	Verbal	$0.36_{\pm 0.02}$	$0.38_{\pm 0.02}$	$0.37_{\pm 0.02}$
	Non-Verbal	$0.38_{\pm 0.02}$	$0.42_{\pm 0.03}$	$0.39_{\pm 0.02}$
	Mulimodal	$0.48_{\pm 0.02}$	$0.53_{\pm 0.02}$	$0.49_{\pm 0.02}$
Doctor+Agent	Verbal	$0.36_{\pm 0.02}$	$0.39_{\pm 0.02}$	$0.37_{\pm 0.02}$
	Non-verbal	$0.50_{\pm 0.02}$	$0.53_{\pm 0.03}$	$0.51_{\pm 0.02}$
	Mulimodal	$0.53_{\pm 0.02}$	$0.56_{\pm 0.03}$	$0.55_{\pm 0.02}$
Random classifier		$0.39_{\pm 0.02}$	$0.52_{\pm 0.03}$	$0.37_{\pm 0.02}$

(7h) co-presence

Figure 7. Test scores of the random forest classifier with different sets of features to analyze the importance of multimodality and the importance of the behavior of each participant of the interaction to predict presence and co-presence.

6 CONCLUSION AND PERSPECTIVES

In this article, we have explored different machine learning methods to analyze the behavioral cues reflecting the sense of presence and co-presence of a user interacting with a virtual patient to break bad news. The proposed method implements an automatic prediction of the sense of presence and co-presence of users based on objective multimodal behavioral measures. Several machine learning techniques have been compared to identify the best parameters to predict the sense of (co-)presence.

Specific verbal and non-verbal behavioral cues have been computed. We have defined high-level features to characterize the user's multimodal behavior. These features describe in particular head and arms movements as well as the lexical richness and linguistic complexity of the verbal behavior. Thanks to a machine learning approach, these features have been correlated to the sense of presence and co-presence assessed with specific subjective questionnaires. The performance measures of the learned models show the accurate predictive capacity of the models. More precisely, we can predict automatically and accurately the sense of presence. The results show that the random forest algorithm, with discretization of the scores of presence in two classes, enables to automatically predict accurately the sense of presence of the user. These results show the interest (and the originality) of the proposed features set - verbal, non-verbal and interactional - for this prediction task. These features can be considered as *objective cues* of the sense of presence of the user during a social interaction with a virtual patient. The prediction of co-presence appears as more difficult to predict. Several elements can be highlighted to explain this results. First, in the co-presence task, a discretization in three classes have been considered. This multi-classes classification problem is more difficult than the binary one considered for presence. Second, these results may reveal that the set of features considered in this article may be not totally adequate for predicting the sense of

co-presence, other features should be considered to improve the prediction. Third, some works highlight the fact that presence and co-presence post-questionnaire experiences may be not sufficient to assess user's sense of presence and co-presence Slater (2004); Bailenson et al. (2004). As in Bailenson et al. (2004), the lack of correlation between behavioral parameters - that have been shown to be cues of engagement in the human-human or human-machine interaction - and the self-report measures may be explained by the inadequacy of the questionnaire to catch certain phenomena. Then, some behavioral cues may be viewed as complementary measures to assess the interaction in virtual environment instead of objective measures replacing self-report questionnaires.

ACKNOWLEDGEMENT

This work has been funded by the French National Research Agency project ACORFORMED (ANR-14-CE24-0034-02) and supported by grants ANR-16-CONV-0002 (ILCB) and ANR-11-IDEX-0001-02 (A*MIDEX), STIC-AMSUD Program for the "Empatia" Project.

REFERENCES

Andrade, A., Bagri, A., Zaw, K., Roos, B., and Ruiz, J. (2010). Avatar-mediated training in the delivery of bad news in a virtual world. *Journal of palliative medicine* 13, 1415–1419

Bailenson, J. N., Aharoni, E., Beall, A. C., Guadagno, R. E., Dimov, A., and Blascovich, J. (2004). Comparing behavioral and self-report measures of embodied agents' social presence in immersive virtual environments. In *Proceedings of the 7th Annual International Workshop on PRESENCE*. 1864–1105

Bailenson, J. N., Swinth, K., Hoyt, C., Persky, S., Dimov, A., and Blascovich, J. (2005). The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence: Teleoperators and Virtual Environments* 14, 379–393

Baumann, D. and Baumann, K. (2014). Reliable estimation of prediction errors for qsar models under model uncertainty using double cross-validation. *Journal of cheminformatics* 6, 47. doi:10.1186/s13321-014-0047-1

Biber, D., Gray, B., and Staples, S. (2016). Contrasting the grammatical complexities of conversation and academic writing: Implications for eap writing development and teaching. *Language in Focus Journal* 2

Bickmore, T., Schulman, D., and Yin, L. (2010). Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence* 24, 648–666

Bigi, B. (2012). Sppas: a tool for the phonetic segmentations of speech. In *The eighth international conference on Language Resources and Evaluation*. 1748–1755

Billinghurst, M. and Weghorst, S. (1995). The use of sketch maps to measure cognitive maps of virtual environments. In *Virtual Reality Annual International Symposium, 1995. Proceedings.* (IEEE), 40–47

Breiman, L. (2001). Random forests. *Machine learning* 45, 5–32

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., and Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods* 40, 540–545

Bystrom, K.-E., Barfield, W., and Hendrix, C. (1999). A conceptual model of the sense of presence in virtual environments. *Presence: Teleoperators and Virtual Environments* 8, 241–244

Cadoz, C. (1994). *Les réalités virtuelles* (Flammarion)

Camden, C. and Verba, S. (1986). Communication and consciousness: Applications in marketing. *Speech Communication*

Csikszentmihalyi, M. (2014). Toward a psychology of optimal experience. In *Flow and the foundations of positive psychology* (Springer). 209–226

Draper, J. V., Kaber, D. B., and Usher, J. M. (1998). Telepresence. *Human factors* 40, 354–375

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research* 15, 3133–3181

Forman, G. and Cohen, I. (2004). Learning from little: Comparison of classifiers given little training. In *European Conference on Principles of Data Mining and Knowledge Discovery* (Springer), 161–172

Glas, N. and Pelachaud, C. (2015). Definitions of engagement in human-agent interaction. In *International Workshop on Engagment in Human Computer Interaction (ENHANCE)*. 944–949

Grassini, S. and Laumann, K. (2020). Questionnaire measures and physiological correlates of presence: A systematic review. *Frontiers in Psychology* 11. doi:10.3389/fpsyg.2020.00349

Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. (2007). Creating rapport with virtual agents. In *International Workshop on Intelligent Virtual Agents* (Springer), 125–138

Heeter, C. (1992). Being there: The subjective experience of presence. *Presence: Teleoperators & Virtual Environments* 1, 262–271

Helme, S. and Clarke, D. (2001). Identifying cognitive engagement in the mathematics classrooms. *Mathematics Educational Journal* 13(2)

Hendrix, C. and Barfield, W. (1996). Presence within virtual environments as a function of visual display parameters. *Presence: Teleoperators & Virtual Environments* 5, 274–289

Ijsselsteijn, W. A. (2002). Elements of a multi-level theory of presence: Phenomenology, mental processing and neural correlates. *Proceedings of PRESENCE* 2002, 245–259

Laarni, J., Ravaja, N., Saari, T., Böcking, S., Hartmann, T., and Schramm, H. (2015). Ways to measure spatial presence: Review and future directions. In *Immersed in Media* (Springer). 139–185

Le Maitre, J. and Chetouani, M. (2013). Self-talk discrimination in human–robot interaction situations for supporting social awareness. *International Journal of Social Robotics* 5, 277–289

Lessiter, J., Freeman, J., Keogh, E., and Davidoff, J. (2001). A cross-media presence questionnaire: The itc-sense of presence inventory. *Presence: Teleoperators and virtual environments* 10, 282–297

Mestre, D. R. (2015). On the usefulness of the concept of presence in virtual reality applications. In *IS&T/SPIE Electronic Imaging*. 93920J–93920J

Nguyen, D. T. and Fussell, S. R. (2016). Effects of conversational involvement cues on understanding and emotions in instant messaging conversations. *Journal of Language and Social Psychology* 35, 28–55

Ochs, M., Mestre, D., de Montcheuil, G., Pergandi, J.-M., Saubesty, J., Lombardo, E., et al. (2018). Training doctors' social skills to break bad news: Evaluation of the impact of virtual environment displays on the sense of presence. *Journal on Multimodal User Interfaces (JMUI)*

Ochs, M., Montcheuil, G., Pergandi, J.-M., Saubesty, J., Donval, B., Pelachaud, C., et al. (2017). An architecture of virtual patient simulation platform to train doctor to break bad news. In *International Conference on Computer Animation and Social Agents (CASA)*

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

Poggi, I. (2007). *Mind, hands, face and body: a goal and belief view of multimodal communication* (Weidler)

Rauzy, S., Montcheuil, G., and Blache, P. (2014). Marsatag, a tagger for french written texts and speech transcriptions. In *Proceedings of Second Asian Pacific Corpus linguistics Conference*. 220

Salperwyck, C. and Lemaire, V. (2011). Impact de la taille de l'ensemble d'apprentissage : une étude empirique. In *Workshop 'CIDN : Clustering incrémental et méthodes de détection de nouveauté', workshop joint to the conference 'Extraction et Gestion des Connaissances (EGC), Brest'*

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., and Paiva, A. (2011). Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th international conference on Human-robot interaction* (ACM), 305–312

Schroeder, R. (2002). Copresence and interaction in virtual environments: An overview of the range of issues. In *Presence 2002: Fifth international workshop*. 274–295

Schubert, T., Friedmann, F., and Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence: Teleoperators and virtual environments* 10, 266–281

Schubert, T. W. (2003). The sense of presence in virtual environments: A three-component scale measuring spatial presence, involvement, and realness. *Zeitschrift für Medienpsychologie* 15, 69–71

Sidner, C. and Lee, C. (2007). Attentional gestures in dialogues between people and robots. *Engineering approaches to conversational informatics. Wiley and Sons*

Sidner, C. L. and Dzikovska, M. (2002). Human-robot interaction: Engagement between humans and robots for hosting activities. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces* (IEEE), 123–128

Slater, M. (2004). How colorful was your day? why questionnaires cannot assess presence in virtual environments. *Presence: Teleoperators & Virtual Environments* 13, 484–493

Slater, M., Linakis, V., Usoh, M., Kooper, R., and Street, G. (1996). Immersion, presence, and performance in virtual environments: An experiment with tri-dimensional chess. In *ACM virtual reality software and technology (VRST)* (ACM Press New York, NY), vol. 163, 72

611 Slater, M., McCarthy, J., and Maringelli, F. (1998). The influence of body movement on subjective presence
612 in virtual environments. *Human Factors* 40, 469–477
613 Slater, M., Sadagic, A., Usoh, M., and Schroeder, R. (2006). Small-group behavior in a virtual and real
614 environment: A comparative study. *Small-Group Behavior* 9
615 Slater, M. and Steed, A. (2000). A virtual presence counter. *Presence: Teleoperators & Virtual*
616 *Environments* 9, 413–434
617 Slater, M. and Wilbur, S. (1997). A framework for immersive virtual environments (five): Speculations
618 on the role of presence in virtual environments. *Presence: Teleoperators and virtual environments* 6,
619 603–616
620 Snoek, C. G. M., Worring, M., and Smeulders, A. W. M. (2005). Early versus late fusion in semantic video
621 analysis. In *ACM Multimedia*
622 Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable
623 importance for random forests. *BMC bioinformatics* 9, 1
624 Tickle-Degnen, L. and Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates.
625 *Psychological inquiry* 1, 285–293
626 Tolochko, P. and Boomgaarden, H. G. (2018). Analysis of linguistic complexity in professional and citizen
627 media. *Journalism Studies* 19, 1786–1803. doi:10.1080/1461670X.2017.1305285
628 Usoh, M., Arthur, K., Whitton, M. C., Bastos, R., Steed, A., Slater, M., et al. (1999). Walking, walking-
629 in-place, flying, in virtual environments. In *Proceedings of the 26th annual conference on Computer*
630 *graphics and interactive techniques* (ACM Press/Addison-Wesley Publishing Co.), 359–364
631 Usoh, M., Catena, E., Arman, S., and Slater, M. (2000). Using presence questionnaires in reality. *Presence:*
632 *Teleoperators and Virtual Environments* 9, 497–503
633 Witmer, B. G. and Singer, M. J. (1998). Measuring presence in virtual environments: A presence
634 questionnaire. *Presence: Teleoperators and virtual environments* 7, 225–240
635 Yu, C., Aoki, P. M., and Woodruff, A. (2004). Detecting user engagement in everyday conversations. *arXiv*
636 *preprint cs/0410027*