ORIGINAL RESEARCH
published: 15 April 2022
doi: 10.3389/fcomp.2022.746804

Multimodal Behavioral Cues Analysis of the Sense of Presence and Social Presence During a Social Interaction With a Virtual Patient

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

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

User's experience evaluation is a key challenge when studying human-agent interaction. Besides user's satisfaction, this question is addressed in virtual reality through the sense of *presence* and *social presence*, generally assessed thanks to subjective post-experience questionnaires. We propose in this article a novel approach making it possible to evaluate automatically these notions by correlating objective multimodal cues produced by users to their subjective sense of presence and social presence. This study is based on a multimodal human-agent interaction corpus collected in a 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 applied machine learning approaches to build a model predicting the user's sense of presence and social presence thanks to specific multimodal behavioral cues. We explore different classification algorithms and machine learning techniques (oversampling and clustering) to cope with the dimensionality of the dataset and to optimize the prediction performance. We obtain models to automatically and accurately predict the level of presence and social presence. The results highlight the relevance of a multimodal model, based both on verbal and non-verbal cues as objective measures of (social) presence. The main contribution of the article is two-fold: 1/ proposing the first presence and social prediction presence models offering a way to automatically provide a user's experience evaluation and 2/ showing the importance of multimodal information for describing these notions.

Keywords: multimodal social signals, sense of presence, virtual reality, virtual patient, conversational agent

OPEN ACCESS

Edited by:
Ryo Ishii,
Nippon Telegraph and Telephone,
Japan

Reviewed by:
Dulce Fernandes Mota,
Instituto Superior de Engenharia do Porto (ISEP), Portugal
Hung-Hsuan Huang,
The Univeristy of Fukuchiyama, Japan
Shogo Okada,
Japan Advanced Institute of Science and Technology, Japan

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

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *24 July 2021*
Accepted: *21 February 2022*
Published: *15 April 2022*

Citation:
Ochs M, Bousquet J, Pergandi J-M and Blache P (2022) Multimodal Behavioral Cues Analysis of the Sense of Presence and Social Presence During a Social Interaction With a Virtual Patient.
Front. Comput. Sci. 4:746804.
doi: 10.3389/fcomp.2022.746804

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 (Witmer and Singer, 1998; Usoh et al., 2000; Bailenson et al., 2005; Grassini and Laumann, 2020; Vasconcelos-Raposo et al., 2021). 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). As highlighted (Fromberger et al., 2015), the sense of presence can be defined as a subjective and psychological reaction to immersive environments. In some research works, the terms “immersion” and “presence” are considered as synonyms. Following the recent review on presence proposed in Grassini and Laumann (2020) and the definition given by Slater (1999), we distinguish these two concepts: the *level of immersion* being related to the technical parameters of the virtual environment and the *sense of presence* refers to the psychological feelings resulting from the immersion. The sense of presence includes the notion of *social presence* (also called *co-presence*), defined as “*the extent to which other entities presented in the virtual environment ‘are there’ from the user point of view*” (Slater et al., 2006) and *self-presence* considered as “*the sense of the user being able to perceive him/herself as part of the virtual environment*” (Lee, 2004).[1]

Evaluating the sense of (social) presence in virtual environment is of deep importance. Several studies have shown the relationship between the sense of presence and human performance during a task (Baus and Bouchard, 2017). More specifically, the sense of (social) presence is particularly relevant in the context of user’s training in virtual reality environments (Stevens and Kincaid, 2015). In this article, we particularly focus on a specific application domain: a virtual reality platform for training doctors interacting with virtual patients. The goal of this platform is to develop doctors’ social skills when breaking news to patients. Such skills are essential for doctors. 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, etc. (Andrade et al., 2010). In order to improve and facilitate doctors’ 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 article, we investigate more particularly the multimodal behavior cues of (social) presence of users training to break bad news to a virtual patient.

The problem in the evaluation of presence and social presence with questionnaires, in spite of their interest, is the subjectivity of the approach (consisting in asking users to self-report their feelings). Previous works have tried to find *objective measures* by hypothesizing that different levels of the sense of presence and social presence may be connected with different verbal and non-verbal user’s behaviors (Ijsselsteijn, 2002; Laarni et al., 2015). 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 of *engagement*[2] in the description of the sense of (social) presence. This idea relies on several observations. First, as shown in Schroeder (2002), the sense of presence and social presence can be correlated with the level of *immersion*. In such case, the greater the immersion, the higher the feeling of (social) presence. Second, the notion of *engagement* 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 (social) presence, we investigate the possibility to automatically predict the sense of (social) 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 users’ productions (verbal and non-verbal). Moreover, we have collected questionnaires indicating the users’ sense of presence and social presence after the interactions. In order to be independent from the environment, our experimental setup involves different virtual reality displays—known to generate different degrees of immersion (PC, virtual reality headset, CAVE). Based on machine learning techniques, we have learned models to correlate verbal and non-verbal cues to different levels of presence and social presence. The accuracy of the models shows that certain verbal and non-verbal cues of the user’s behavior can be used to predict her level of presence and social presence, based on objective behavioral measures.

The article is organized as follows. In the next section, we present the theoretical background and related works on the notion of presence and social 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 models 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 social presence clustered to classes to predict. We conclude and discuss perspectives Section 6.

2. THE SENSE OF PRESENCE AND SOCIAL PRESENCE

2.1. Definition of the Sense of (Social) Presence

The concept of presence is central in virtual reality domain reflecting the quality of a given virtual environment. This concept has been widely studied by different authors leading to different definitions (for a detailed review see Skarbez et al., 2017). This being said, *presence* is commonly defined as the feeling of “*being there*” in a virtual place. In Witmer and Singer (1998), the researchers define presence as “*the subjective experience of being in one place or environment, even when one is physically situated in another*”.

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

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

[2] We consider in this article engagement as a synonym of involvement.

presence is strong (Witmer and Singer, 1998); (4) the *duration of the exhibition*: prolonged exposure beyond 15 min 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.

In this article, besides the *presence*, we are particularly interested in the notions of *social presence*, the task of the users in our context implying an interaction with a humanoid virtual character. As for the notion of presence, different definitions of social presence have been proposed (Skarbez et al., 2017). In this research, we consider the definition social presence (called social presence illusion) proposed in Skarbez et al. (2017) "*the sense of being together with another or others*". The authors distinguish the co-presence from the *social presence* (called social presence illusion) defined as "*the moment-by-moment awareness of the copresence of another sentient being accompanied by a sense of engagement with them*". Given our context of a human-virtual agent interaction, we focus in particular on the *social presence illusion*.

2.2. Questionnaires of (Social) Presence

Several questionnaires have been proposed in order to assess the sense of presence (see Skarbez et al., 2017 or Grassini and Laumann, 2020 for complete surveys). Some of them, considered as "canonical", have been used in many different works: the canonical presence test of 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, the Lombard's questionnaire including six dimensions of presence (Lombard et al., 2000), the Reality Judgment and Presence Questionnaire (RJPQ) proposed in Baños et al. (2000), and the canonical test IGroup Presence Questionnaire (IPQ) (Schubert et al., 2001). We used the last one in our work to evaluate the users' presence. This test focuses on three variables dependent on presence factors: spatial presence, engagement 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 *social presence*. In our context, we are interested in evaluating the sense of social presence of the participants with the virtual agent. In order to evaluate the social presence, we have used the test proposed in Bailenson et al. (2005) that measures social presence through the following variables: the *perceived social 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*". The questionnaire is 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).

2.3. Behavioral Measures of (Social) Presence

Several works have explored different *objectives measures* to evaluate the sense of presence or social presence. In this perspective, as mentioned in Slater et al. (1998), we propose to distinguish "*subjective*" from "*behavioral*" presence, subjective presence being measured by means of questionnaires while behavioral presence corresponds to bodily responses. Three types of objective measures have been proposed: behavioral (e.g., attention, gestures), performance-based (e.g., user's performance in task realization) and physiological (e.g., brain activity, heart rate) (Ijsselsteijn, 2002; Grassini and Laumann, 2020). In this article, we focus more specifically on behavioral measures.

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 toward an object and the correlation with the level of presence. Other works have shown a close relation between body movements (for instance their amplitude) and the sense of presence (Slater et al., 1998; Slater and Steed, 2000). In Bailenson et al. (2004), the authors have compared social presence self-report measures and the interpersonal distances of the user with virtual agents (results did not reveal significant correlations between these objective and subjective measures).

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

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

2.4. (Social) Presence, Engagement, and Psychological Immersion

In our interdisciplinary approach, we aim at connecting empirical and theoretical backgrounds from different domains around the notion of presence and social 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: *engagement* and *psychological immersion* [also called *spatial presence* in Witmer and Singer (1998)]: "*engagement 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 term psychological immersion in this definition refers to the sense of presence (see Section 2.1). Following Witmer and Singer (1998), we adopt in our work a broader perspective of presence including the engagement of the participant.

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 article to bridge the gap between presence and observable features by posing an hypothesis: *the senses presence and social presence are correlated with engagement*. This hypothesis relies on the definition of social presence as proposed above (Biocca et al., 2001; Schubert et al., 2001) in the context of an interaction. Moreover, in a virtual environment, no engagement can be observed without a high level of social presence. Consequently, a correlation should be observed between the level of (social) presence and that of engagement. Engagement being possibly assessed based on different objective cues, we propose to use these same features in order to predict the level of (social) 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 (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). 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 and Social Presence

The notions of presence ans social presence are not directly evaluated in the literature on the basis of objective multimodal behavioral cues. Considering the notion of engagement as relevant for participating in such an evaluation, we propose in this section a brief overview of the different features that can be used for the description of these notions.

Concerning verbal cues, several works have addressed the question of the type of lexical, syntactic and semantic aspects that can be related with engagement. 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 (social) presence.

For this purpose, we have first collected a corpus of human-virtual agent interaction described in the next section, then extracted automatically behavioral cues from the corpus (section 4) to finally applied machine learning methods (Section 5). This approach has been already explored in several research works to analyze other aspects of the user experience. For instance, in Wei et al. (2021), the authors have proposed a model to automatically and accurately predict the user's satisfaction based on multimodal features, and in Foster et al. (2017) to measure the level of engagement of a user in interaction with a robot. However, as far as we know, the automatic prediction of presence and social presence remains unexplored.

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

In order to analyze the multimodal cues of (social) 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 platform is *semi-autonomous* because some modules of the system are automatic (for example the dialogue generation) where some others are manual. In particular, the speech recognition and the comprehension modules are simulated by a human: the doctor verbal production is interpreted in real time by the operator (always the same person) which selects the adequate input signal to be transmitted to the dialog system. Indeed, these modules may be particularly critical in case of failure and then damage the interaction strongly. They represent moreover the most difficult part of the system to be developed. Replacing the module by the operator comes to a perfect speech recognition and comprehension. This makes it possible to completely control the corresponding parameters and concentrate on the collection of the corpus and the evaluation of the (social) presence. As described in Ochs et al. (2018a), a specific interface has been designed and tested for this purpose to enable the experimenter to select the sentences as close as possible to what has been said. Note that at the difference with a "Wizard of Oz", the experimenter does not select the virtual patient's reaction but only sends to the dialog model the recognized doctor's sentence.

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, 36 persons (26 males, 10 females) with a mean age of 29 years (SD:10.5) volunteered to participate to the experimentation. Twenty-five participants are students with different backgrounds (linguistics, computer science, and psychology) 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. Participant inclusion for this type of task is an issue taking into account the difficulty to recruit experts (doctors with an experience in breaking bad news). We collaborate in this experiment with a hospital, giving us the opportunity to recruit 13 doctors, which is important. Other participants are naive. In order to reduce the differences between experts and naive, we have designed an experimental setup with a very precise and well-documented task and a detailed scenario that all the participants have to strictly follow (Section 3).

3.3. Collection 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.

3.3.1. 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.

3.3.2. 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] The task is to announce this medical situation to the virtual patient. Participants were accurately instructed about the causes of the problem, the effects (pain), and the proposed remediation (a new surgery, urgently). They also received precise instructions about the type of vocabulary (not too technical), the attitude (in particular empathy), and how to guide the dialogue by respecting different phases (opening, describing the situation, delivering the bad news, and the remediation).

[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).

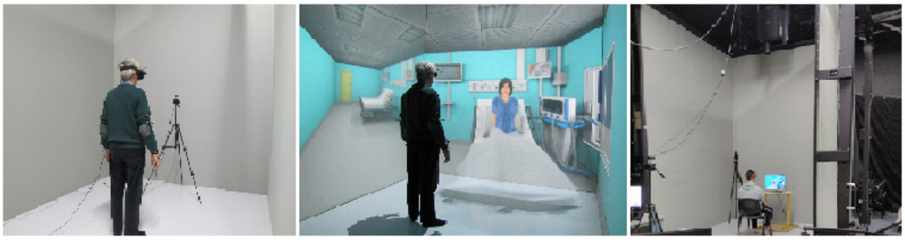
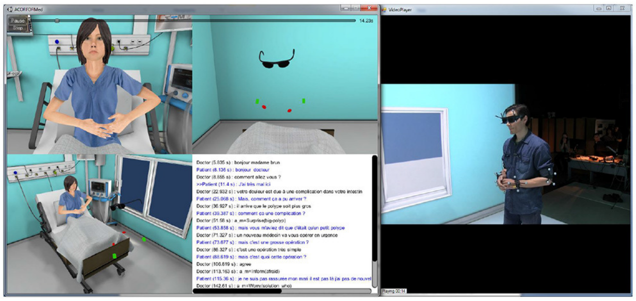


FIGURE 1 | Participants interacting with the virtual patient with different virtual environment displays (from left to right): virtual reality headset (HMD), virtual reality room (CAVE), and PC monitor.



Doctor (5.835 s) : bonjour madame brun
Patient (8.136 s) : bonjour docteur
Doctor (8.855 s) : comment allez vous ?
>>Patient (11.4 s) : J'ai très mal ici
Doctor (22.932 s) : votre douleur est due à une complication dans votre intestin
Patient (25.068 s) : Mais, comment ça a pu arriver ?
Doctor (36.927 s) : il arrive que le polype soit plus gros
Patient (39.387 s) : comment ça une complication ?
Doctor (51.58 s) : a_m=Surprise(big-polyp)
Patient (53.858 s) : mais vous m'aviez dit que c'était qu'un petit polype
Doctor (71.327 s) : un nouveau médecin va vous opérer en urgence
Patient (73.677 s) : mais c'est une grosse opération ?
Doctor (86.327 s) : c'est une opération très simple
Patient (88.619 s) : mais c'est quoi cette opération ?
Doctor (106.819 s) : agree
Doctor (113.163 s) : a_m=Inform(afraid)
Patient (115.36 s) : je ne suis pas rassurée mon mari il est pas là j'ai pas de nouvel
Doctor (142.61 s) : a_m=Worry(solution_who)

FIGURE 2 | 3D video playback player.

3.3.3. 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 1**). The virtual reality cave is constituted of a 3 m deep, 3 m wide, and 4 m 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.

The order of presentation of each display modality was counterbalanced within 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 displays in order to avoid an effect of the order on the results. The duration of each interaction is in average 3 min 16 s.

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.

3.3.4. 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 social 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.

Note that the behavior of the agent (controlled by the operator) depends on the participants' production. In the context of a simulation of a natural interaction, it is necessary to introduce a variability in the agent's responses to adapt to each participant's behavior. As a consequence, these responses are not exactly the same through the different dialogues in order to ensure a believable interaction. We control the variability by defining a very precise scenario that the participants have to follow, by always using the same person to control the virtual agent and by limiting the behavior of the agent to a set of predefined behaviors that the operator can select (same set of behavior for the three conditions).

3.3.5. The Dataset

In total, the data contains 108 human-agent interactions representing a total of 5 h 34 min (each interaction lasting in average 3 min 16 s). The dataset size is in the average for this type of data, taking into account the difficulty of this specific task, recruiting different types of participants (experts and non-experts) using three different devices. Technically, as described in the last section, we need then to apply specific machine learning methods overcoming this limitation.

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 dataset, 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 dataset 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.

Many works have shown the interest of using both verbal and non-verbal features for modeling social skills, interaction behaviors and language phenomena. We follow this methodology by enriching the classical features used in the literature (POS tags, sentence length, nods, gestures, etc.) with higher level information (lexical richness, syntactic complexity) shown to be also relevant for the description of engagement (as described Section 2.5).

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.

4.1.1. 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: $\frac{nb\_conj+nb\_prep+nb\_pro}{\sum tokens}$.

4.1.2. 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).

4.1.3. 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]

4.1.4. 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 (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 (social) 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;

[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 behavior of our virtual patient, which, lying in bed, does not move much, but sometimes rotates its head or arms.

- *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).

The non-verbal behavior of the participants may vary depending on the devices. Even if we do not consider the device as input features, the behavior of the participants are represented through the features. The machine learning task enables us to correlate the behavior of the participant to the sense of (social) presence. Considering all the interactions with the different devices in the same dataset enables us to have different levels of sense of presence (as shown in Ochs et al., 2018b). Moreover, our objective is to have an average predictive model whatever is the device.

In the next section, the matrix is used to learn a model to automatically predict the sense of presence and social 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 vs. novices) on the behavior displayed and on the sense of presence and social presence is described in details in Ochs et al. (2018b). In this article, we focus on the automatic prediction of the sense of presence and social 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 social presence. We have shown in Ochs et al. (2018b) that the three interaction modes imply different levels of presence and social presence.

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

Our goal is to predict users' sense of *presence* and *social 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 social presence.

The same features, described in the previous section, are used to learn both models. For each interaction, the sense of presence and social 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 social 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 (Forman and Cohen, 2004; Strobl et al., 2008; Salperwyck and Lemaire, 2011). Moreover, they are easier to interpret when entering into the comparison of different feature combinations and feature importance. Last, but not least, they are also well-suited for handling small datasets.

In order to evaluate the best approach for the automatic prediction of presence and social presence, we have explored different clustering strategies (Section 5.2) and oversampling methods (Section 5.3). **Figure 3** illustrates the different steps and the prediction approach.

5.1. Classifiers' Training and Test Procedure

The dataset 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 dataset 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 "trai" and "test" metrics. Given the size of the dataset, 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. **Figure 4** illustrates the process, based on a double cross-validation.

Concerning the Random Forest (RF) 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).

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

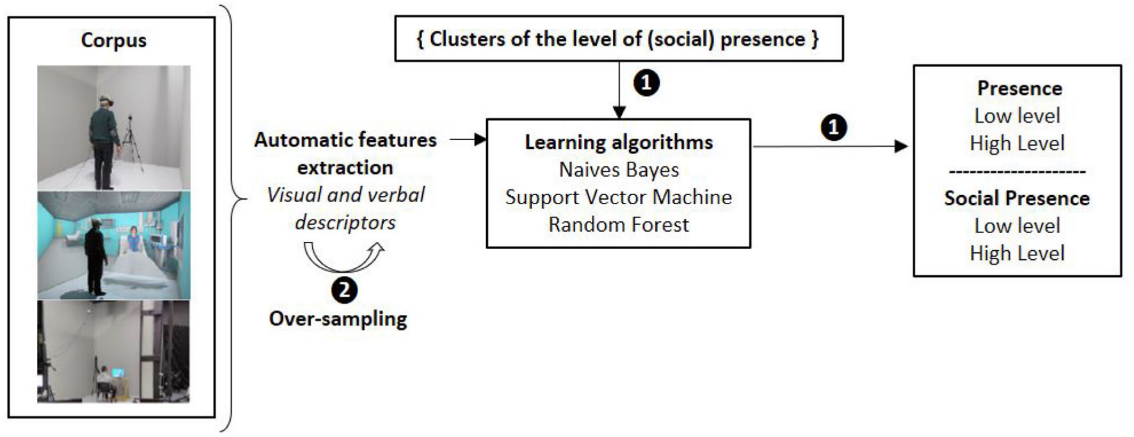


FIGURE 3 | Overview of our multimodal approach to automatically predict presence and social presence.

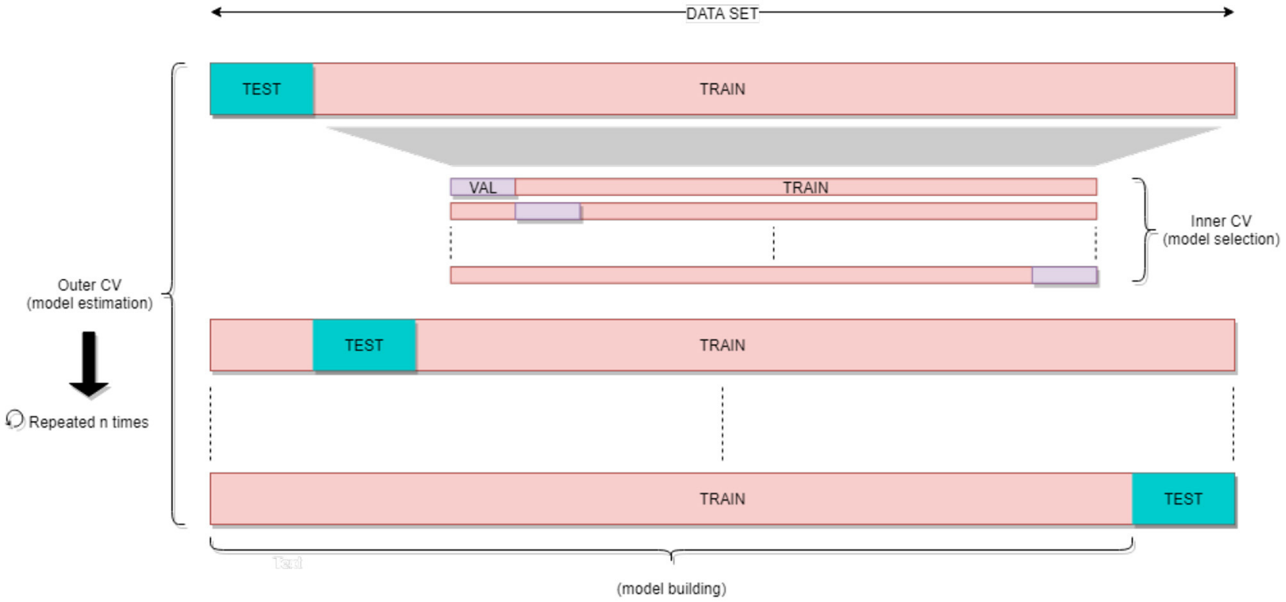


FIGURE 4 | Double cross-validation.

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.

5.2. Identification of the Best Classifier With the Best Granularity Level of Presence and Social 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 social presence that we can predict. Indeed, the level of presence and social 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 social presence, 3 for a medium class, and the 4 and 5 to represent high value of presence

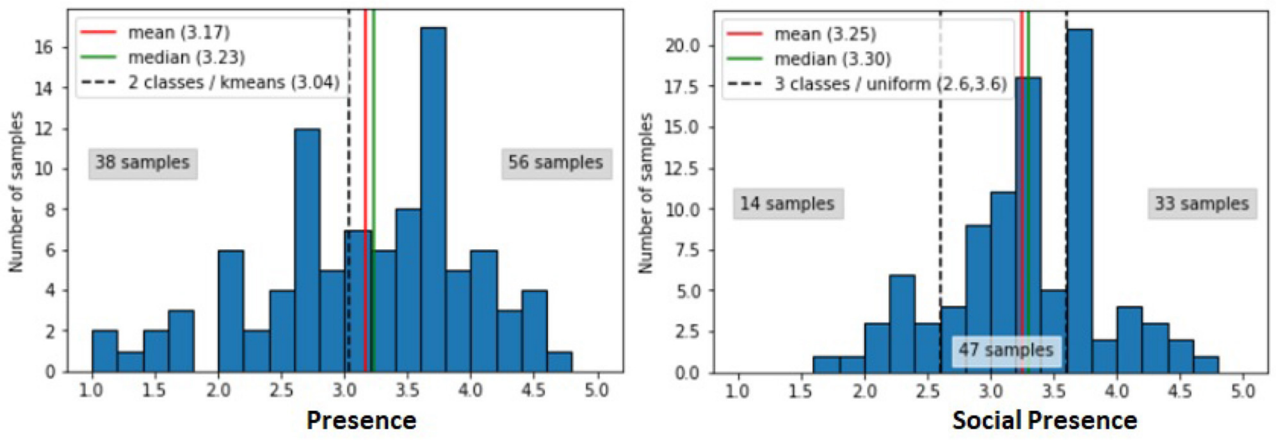


FIGURE 5 | Distribution of the scores of presence and social presence in the dataset.

and social 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 social presence on the dataset is illustrated **Figure 5**.

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 Forest (compared to Naïves Bayes and SVM) both for the prediction of presence and for social presence. We illustrate the test scores of this classifier (**Figure 6**). 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 social 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 social presence are illustrated (**Figure 5**) 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 social presence seems more difficult to predict with scores closed to 0.5. This lower performance for the social 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 social 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 social presence (as illustrated **Figure 5**).

5.3. Exploring Over-sampling Methods to Face Small Dataset

Given the size of the dataset, 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 dataset, 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 neighbors).

The results (illustrated **Figure 7**) show that over-sampling our dataset with these techniques has no influence on the prediction of sense of presence. However, for the prediction of social

[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.

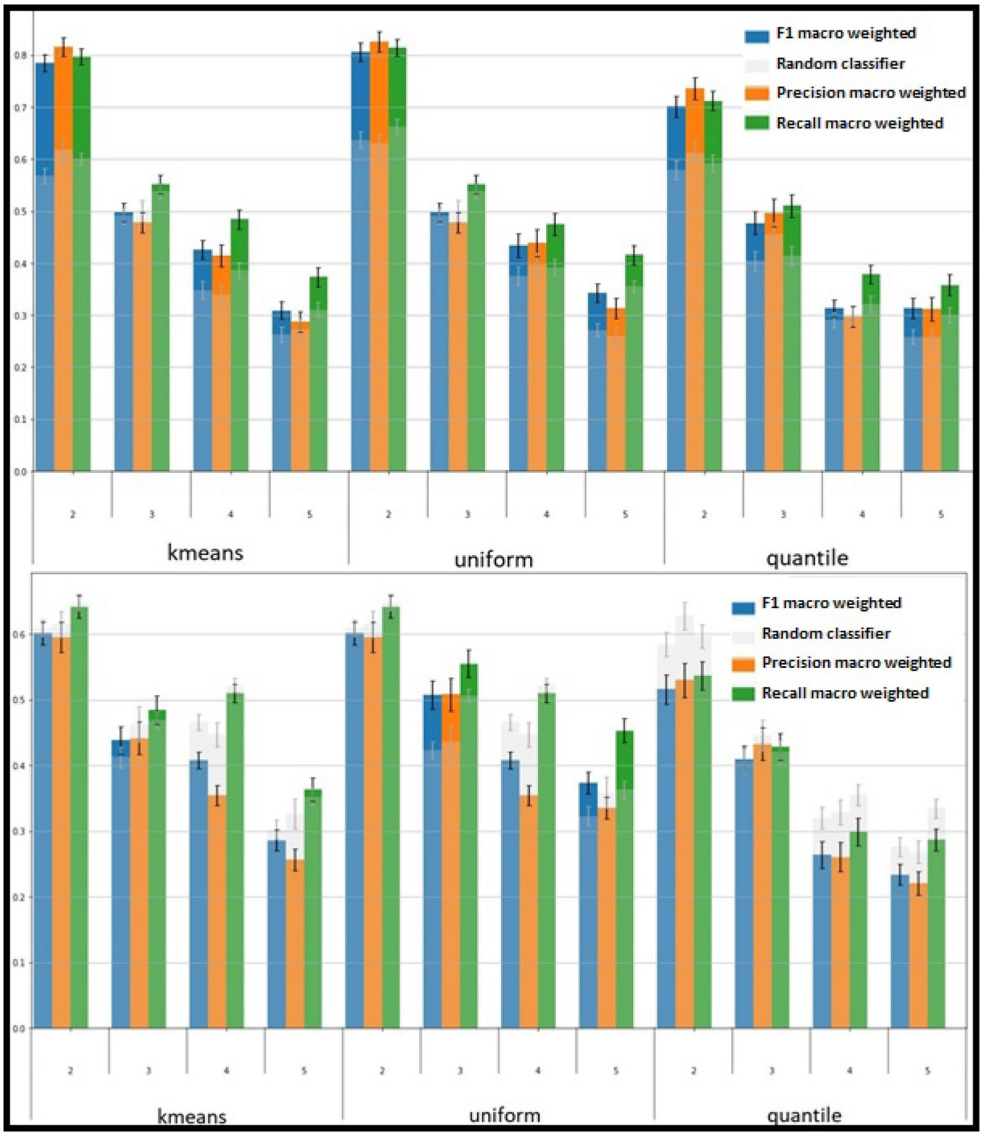


FIGURE 6 | Test scores of the random forest considering different discretization strategies (On top: presence; below: social presence).

presence, SMOTE improves the F_1 score. Consequently, we apply SMOTE for the social presence classification task.

5.4. Verbal and Non-verbal Behavioral Cues Importance

In this section, we analyze the importance of the behavioral cues to predict presence and social presence. The models were configured with respect to findings from the preliminary studies presented above (hyper-parameters search spaces, discretization parameters for presence, and social presence). We consider the Random Forest classifier and use a random classifiers as baseline (no state-of-the-art model being available in the literature). 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. *Verbal features only*: 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. *Non-verbal features only*: averages and standard deviations of the rotations of head and arms movements
3. *Multimodal features*: the verbal and non-verbal features.

[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.

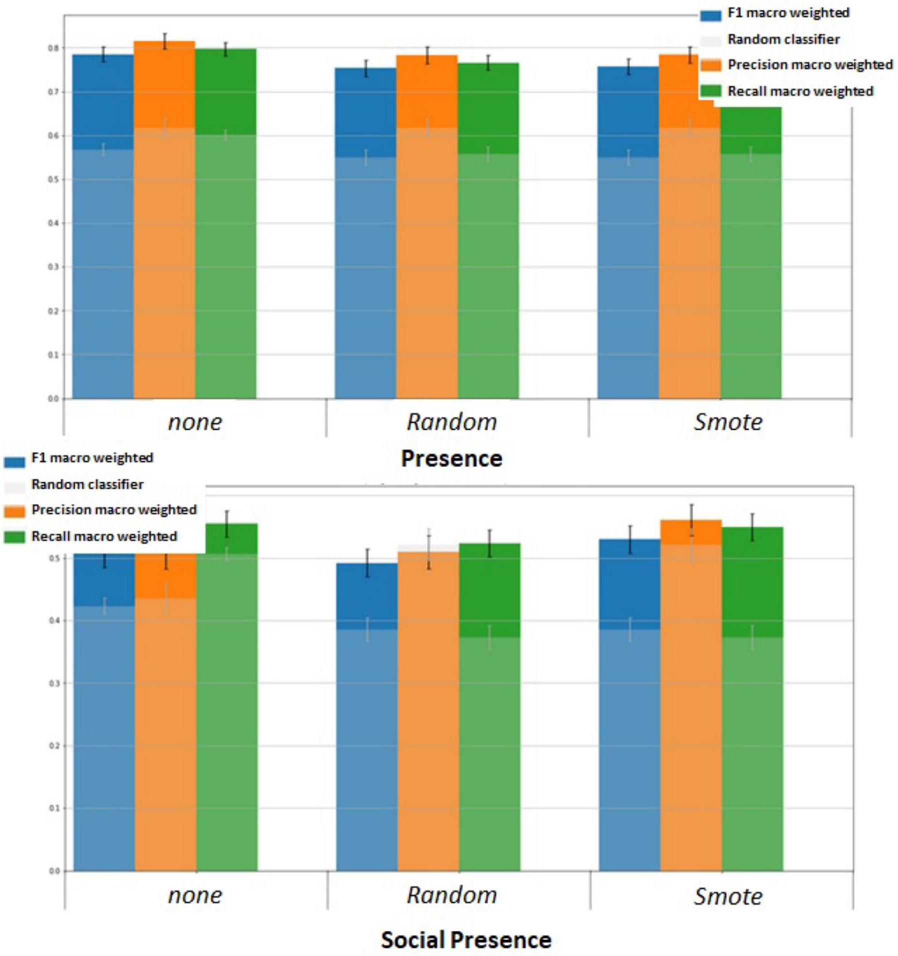


FIGURE 7 | Test scores of the random forest classifier considering different over-sampling strategies or none.

The results are reported **Figures 8–10**. 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. These three different conditions ("*Doctor*", "*Agent*", "*Agent* + *Doctor*") are considered first because the agent's and participant's behaviors differ in each interaction. More importantly, it is necessary to take into consideration, on top of the participant's features, also the agent's features (the agent's behavior having a direct impact on the participant's production). The goal there is then to identify the importance of the agent's behavior and/or doctor's behavior in the evaluation of the sense of (social) presence. We propose therefore, based on the same set of features for agents and participants, to compare the three conditions exposed above.

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 social 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 social presence are lower that for presence, which confirms the difficulty to predict the sense of social 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 social presence. This results show the importance of the non-verbal behavioral cues in the prediction of (social) presence.

We have compared the importance of the behavior of each participant to the interaction to predict (social) presence: the *user* (noted "*Doctor*" in **Figure 10**) and the *virtual patient* (noticed "Agent" in **Figure 10**). 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 social 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.

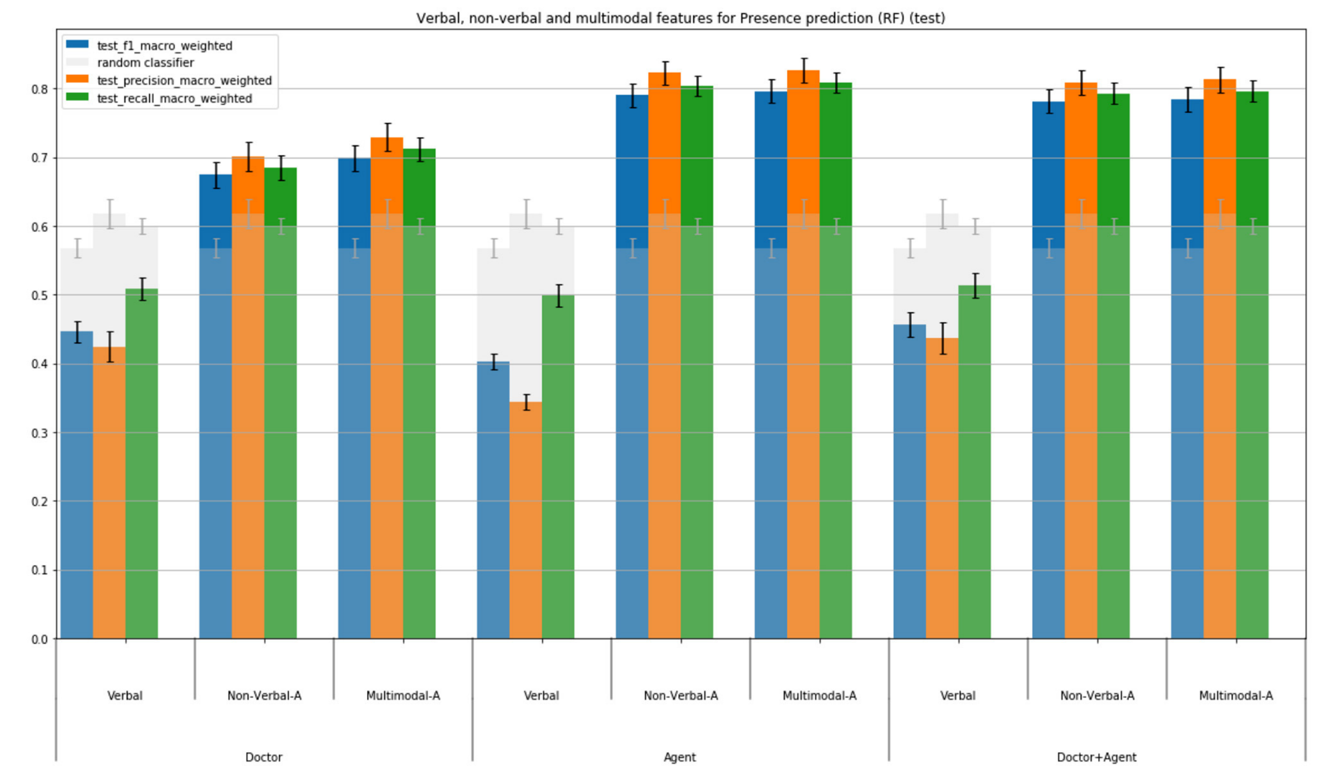


FIGURE 8 | Modality importance for predicting *Presence*.

5.5. Discussion, Limits of the Approach

Assessing automatically user's experience in general and the sense of (social) presence in particular requires first a precise and controlled dataset. Several remarks can be done regarding this aspect. First, the size of our dataset is rather small, and we need to increase it. Even though specific methods makes it possible to artificially balanced the classes (in our case thanks to oversampling), no automatic data generation can be applied. As a consequence, new data acquisition should be done, by taking into consideration some drawbacks. First, we argued in favor of mixing experts and non experts in a task-oriented application. Doing that needs first to acquire a more balanced recruitment, making it possible to have enough data in order to control and compare this condition. In the same perspective, we also need to control the participants' previous experience in virtual reality, which can have a consequence on their self-evaluation.[10]

Increasing the size of the dataset will make it possible to apply more controlled learning techniques, in particular by giving the possibility to create a separate dataset of unseen data. It will also fix the question of inter-participant variability. Note that, still in this variability perspective, it would be interesting and necessary to also acquire data with different tasks, staying in a first stage in the training context.

The learning algorithms used in this research work do not consider the temporal aspect of the interaction. However, the interaction is dynamic and generally organized by phases (e.g., greetings, argumentation, closing). The choice of the learning algorithms was motivated by the relatively short duration of the interactions (in average 3 min 16 s) and their capacity to cope with the specificity of the dataset (high dimensional). In order to take into account the dynamic of the interaction, the different phases could be annotated and used in the learning process.

Another limitation of our experiment concerns the nature of input data. Our goal being to propose an automatic assessment, features have also to be extracted automatically from raw data. Moreover, as explained above, we showed the interest of bringing various features from different modalities into the model. As a consequence, we need tools efficient enough for feature extraction. Moreover, we also need to acquire input data that can be processed by such tools. In our experiment, we only focused on the most directly accessible features. However, it is important to enlarge the feature set toward the acquisition of different information potentially playing a role in (social) presence evaluation such as intonation, rhythm, gaze, postures, etc. Data acquisition should then take into account this perspective by identifying the relevant tools and the specific requirements with respect to the input data.

One important aspect and originality of our dataset is that it relies on different setups (Virtual reality room, virtual reality headset, PC). We think important to keep this variability.

[10] Even if this cannot be generalized, all of our participants gave a positive feedback on the realism of the agent's reactions, whatever the setup.

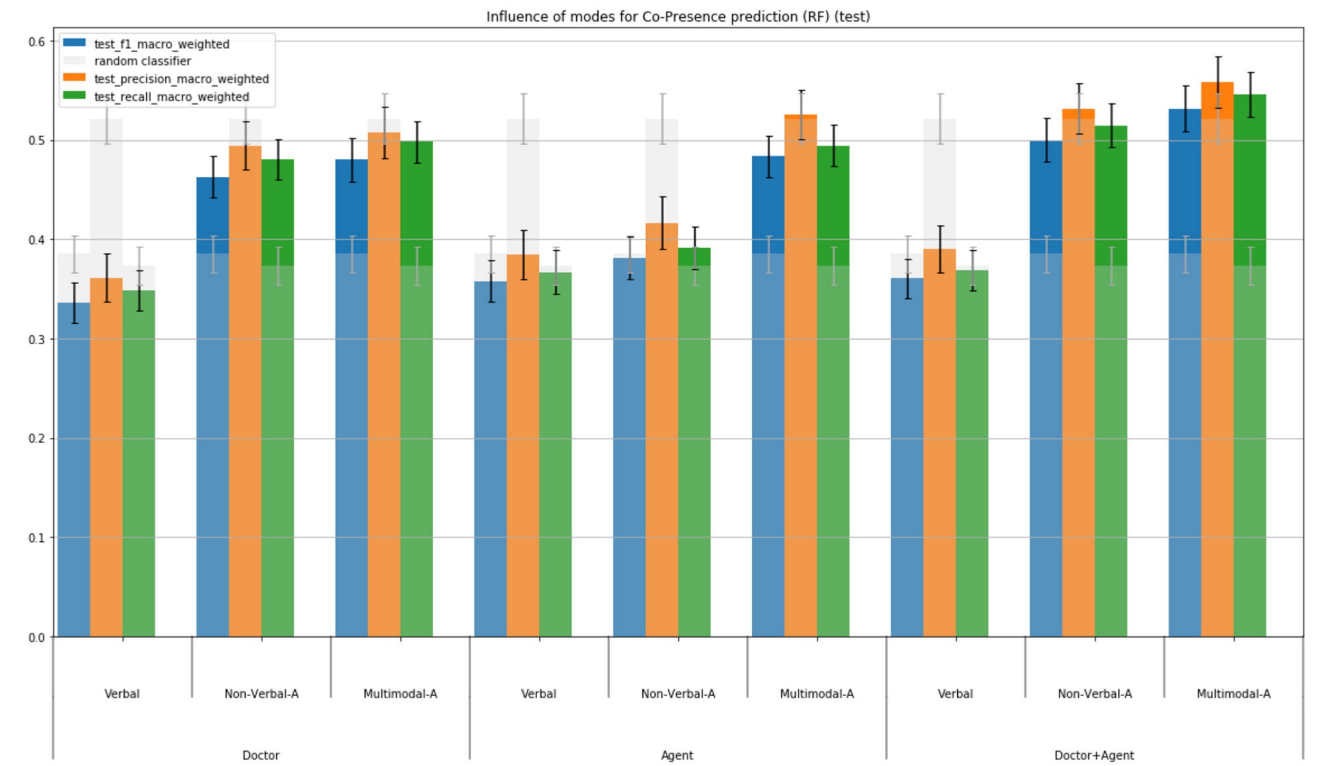
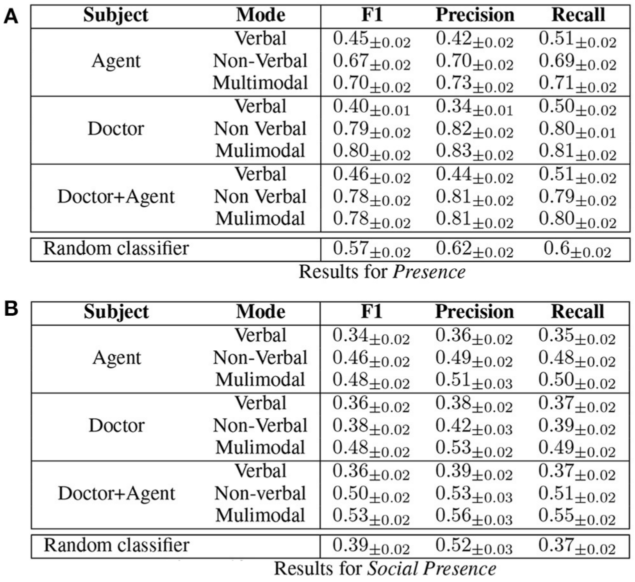


FIGURE 9 | Modality importance for predicting *Social presence*.



A

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		$0.57_{\pm 0.02}$	$0.62_{\pm 0.02}$	$0.6_{\pm 0.02}$

Results for *Presence*

B

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		$0.39_{\pm 0.02}$	$0.52_{\pm 0.03}$	$0.37_{\pm 0.02}$

Results for *Social Presence*

FIGURE 10 | 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 social presence. **(A)** Results for *Presence*. **(B)** Results for *Social Presence*.

However, due to the size of the dataset, it is not possible at this stage to do specific studies per setup in order to compare them. This aspect constitutes the third constraint to deal with for acquiring new data.

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 social 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 social 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 (social) 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 social 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 social presence appears as more difficult to predict. Several elements can be highlighted to explain this results. First, in the social 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 social presence, other features should be considered to improve the prediction. Third, some works highlight the fact that presence and social presence post-questionnaire experiences may be not sufficient to assess user's sense of presence and social presence (Bailenson et al., 2004; Slater, 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.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

MO and JB: machine learning experiments. J-MP: pre-processing of the data. PB: linguistic analysis. All authors contributed to the article and approved the submitted version.

FUNDING

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. *J. Palliat. Med.* 13, 1415–1419. doi: 10.1089/jpm.2010.0108

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* (Valencia), 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 Teleoperat. Virt. Environ.* 14, 379–393. doi: 10.1162/105474605774785235

Baños, R. M., Botella, C., Garcia-Palacios, A., Villa, H., Perpi ná, C., and Alcaniz, M. (2000). Presence and reality judgment in virtual environments: a unitary construct? *Cyberpsychol. Behav*. 3, 327–335. doi: 10.1089/10949310050078760

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

Baus, O., and Bouchard, S. (2017). Exposure to an unpleasant odour increases the sense of presence in virtual reality. *Virtual Real*. 21, 59–74. doi: 10.1007/s10055-016-0299-3

Biber, D., Gray, B., and Staples, S. (2016). Contrasting the grammatical complexities of conversation and academic writing: implications for EAP writing development and teaching. *Lang. Focus J.* 2, 1–18. doi: 10.1515/lifijsal-2016-0001

Bickmore, T., Schulman, D., and Yin, L. (2010). Maintaining engagement in long-term interventions with relational agents. *Appl. Artif. Intell*. 24, 648–666. doi: 10.1080/08839514.2010.492259

Bigi, B. (2012). "SPPAS: a tool for the phonetic segmentations of speech," in *The Eighth International Conference on Language Resources and Evaluation* (Istanbul), 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* (IEEE), 40–47. doi: 10.1109/VRAIS.1995.512478

Biocca, F., Harms, C., and Gregg, J. (2001). "The networked minds measure of social presence: pilot test of the factor structure and concurrent validity," in *4th Annual International Workshop on Presence* (Philadelphia, PA), 1–9.

Breiman, L. (2001). Random forests. *Mach. Learn*. 45, 5–32. doi: 10.1023/A:1010933404324

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. *Behav. Res. Methods* 40, 540–545. doi: 10.3758/BRM.40.2.540

Camden, C., and Verba, S. (1986). Communication and consciousness: Applications in marketing. *Speech Commun*. 50, 64–73. doi: 10.1080/10570318609374213

Csikszentmihalyi, M. (2014). "Toward a psychology of optimal experience," in *Flow and the Foundations of Positive Psychology* (Dordrecht: Springer), 209–226. doi: 10.1007/978-94-017-9088-8_14

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res*. 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*, (Pisa: Springer), 161–172. doi: 10.1007/978-3-540-30116-5_17

Foster, M. E., Gaschler, A., and Giuliani, M. (2017). Automatically classifying user engagement for dynamic multi-party human-robot interaction. *Int. J. Soc. Robot*. 9, 659–674. doi: 10.1007/s12369-017-0414-y

Fromberger, P., Meyer, S., Kempf, C., Jordan, K., and Müller, J. L. (2015). Virtual viewing time: the relationship between presence and sexual interest in androphilic and gynephilic men. *PLoS ONE* 10, e0127156. doi: 10.1371/journal.pone.0127156

Glas, N., and Pelachaud, C. (2015). "Definitions of engagement in human-agent interaction," in *International Workshop on Engagment in Human Computer Interaction (ENHANCE)* (Paris), 944–949. doi: 10.1109/ACII.2015.7344688

Grassini, S., and Laumann, K. (2020). Questionnaire measures and physiological correlates of presence: a systematic review. *Front. Psychol.* 11, 349. 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* (Stockholm: Springer), 125–138. doi: 10.1007/978-3-540-74997-4_12

Heeter, C. (1992). Being there: the subjective experience of presence. *Presence Teleoperat. Virtual Environ.* 1, 262–271. doi: 10.1162/pres.1992.1.2.262

Hendrix, C., and Barfield, W. (1996). Presence within virtual environments as a function of visual display parameters. *Presence Teleoperat. Virtual Environ.* 5, 274–289. doi: 10.1162/pres.1996.5.3.274

Ijsselsteijn, W. A. (2002). "Elements of a multi-level theory of presence: phenomenology, mental processing and neural correlates," in *Proceedings of Presence 2002* (Porto), 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. doi: 10.1007/978-3-319-10190-3_8

Le Maitre, J., and Chetouani, M. (2013). Self-talk discrimination in human-robot interaction situations for supporting social awareness. *Int. J. Soc. Robot.* 5, 277–289. doi: 10.1007/s12369-013-0179-x

Lee, K. M. (2004). Presence, explicated. *Commun. Theory* 14, 27–50. doi: 10.1111/j.1468-2885.2004.tb00302.x

Lessiter, J., Freeman, J., Keogh, E., and Davidoff, J. (2001). A cross-media presence questionnaire: the ITC-sense of presence inventory. *Presence Teleoperat. Virtual Environ.* 10, 282–297. doi: 10.1162/105474601300343612

Lombard, M., Ditton, T. B., Crane, D., Davis, B., Gil-Egui, G., Horvath, K., et al. (2000). "Measuring presence: a literature-based approach to the development of a standardized paper-and-pencil instrument," in *Third International Workshop on Presence* (Delft), 2–4.

Mestre, D. R. (2015). "On the usefulness of the concept of presence in virtual reality applications," in *IS&T/SPIE Electronic Imaging* (San Francisco), 93920J. doi: 10.1117/12.2075798

Nguyen, D. T., and Fussell, S. R. (2016). Effects of conversational involvement cues on understanding and emotions in instant messaging conversations. *J. Lang. Soc. Psychol.* 35, 28–55. doi: 10.1177/0261927X15571538

Ochs, M., Blache, P., de Montcheuil, G. M., Pergandi, J.-M., Saubesty, J., Francon, D., et al. (2018a). "A semi-autonomous system for creating a human-machine interaction corpus in virtual reality: application to the acorformed system for training doctors to break bad news," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki).

Ochs, M., Mestre, D., de Montcheuil, G., Pergandi, J.-M., Saubesty, J., Lombardo, E., et al. (2018b). Training doctors' social skills to break bad news: evaluation of the impact of virtual environment displays on the sense of presence. *J. Multimodal User Interfaces.* 13, 41–51. doi: 10.1007/s12193-018-0289-8

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)* (Paris).

Pelachaud, C. (2009). Studies on gesture expressivity for a virtual agent. *Speech Commun.* 51, 630–639. doi: 10.1016/j.specom.2008.04.009

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* (Taipei), 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* (Lausanne: ACM), 305–312. doi: 10.1145/1957656.1957781

Schroeder, R. (2002). "Copresence and interaction in virtual environments: an overview of the range of issues," in *Presence 2002: Fifth International Workshop* (Porto), 274–295.

Schubert, T., Friedmann, F., and Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence Teleoperat. Virtual Environ.* 10, 266–281. doi: 10.1162/105474601300343603

Schubert, T. W. (2003). The sense of presence in virtual environments: a three-component scale measuring spatial presence, involvement, and realness. *Z. Medienpsychol.* 15, 69–71. doi: 10.1026//1617-6383.15.2.69

Sidner, C., and Lee, C. (2007). *Attentional Gestures in Dialogues Between People and Robots. Engineering Approaches to Conversational Informatics*. Pittsburgh, PA: Wiley and Sons. doi: 10.1002/9780470512470.ch6

Sidner, C. L., and Dzikovska, M. (2002). "Human-robot interaction: engagement between humans and robots for hosting activities," in *Fourth IEEE International Conference on Multimodal Interfaces* (Pittsburgh, PA: IEEE), 123–128. doi: 10.1109/ICMI.2002.1166980

Skarbez, R., Brooks, F. P. Jr, and Whitton, M. C. (2017). A survey of presence and related concepts. *ACM Comput. Surv.* 50, 1–39. doi: 10.1145/3134301

Slater, M. (1999). Measuring presence: a response to the witmer and singer presence questionnaire. *Presence* 8, 560–565. doi: 10.1162/105474699566477

Slater, M. (2004). How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence Teleoperat. Virtual Environ.* 13, 484–493. doi: 10.1162/1054746041944849

Slater, M., McCarthy, J., and Maringelli, F. (1998). The influence of body movement on subjective presence in virtual environments. *Human Fact.* 40, 469–477. doi: 10.1518/001872098779591368

Slater, M., Sadagic, A., Usoh, M., and Schroeder, R. (2006). Small-group behavior in a virtual and real environment: a comparative study. *Small Group Behav.* 9, 37–51. doi: 10.1162/105474600566600

Slater, M., and Steed, A. (2000). A virtual presence counter. *Presence Teleoperat. Virtual Environ.* 9, 413–434. doi: 10.1162/105474600566925

Snoek, C. G. M., Worring, M., and Smeulders, A. W. M. (2005). "Early versus late fusion in semantic video analysis," in *ACM Multimedia* (Singapore). doi: 10.1145/1101149.1101236

Stevens, J., and Kincaid, J. (2015). The relationship between presence and performance in virtual simulation training. *Open J. Modell. Simul.* 3, 41–48. doi: 10.4236/ojmsi.2015.32005

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinform.* 9, 307. doi: 10.1186/1471-2105-9-307

Tickle-Degnen, L., and Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychol. Inq.* 1, 285–293. doi: 10.1207/s15327965pli0104_1

Tolochko, P., and Boomgaarden, H. G. (2018). Analysis of linguistic complexity in professional and citizen media. *J. Stud.* 19, 1786–1803. doi: 10.1080/1461670X.2017.1305285

Usoh, M., Arthur, K., Whitton, M. C., Bastos, R., Steed, A., Slater, M., et al. (1999). "Walking, walking-in-place, flying, in virtual environments," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY: ACM Press/Addison-Wesley Publishing Co.), 359–364. doi: 10.1145/311535.311589

Usoh, M., Catena, E., Arman, S., and Slater, M. (2000). Using presence questionnaires in reality. *Presence Teleoperat. Virtual Environ.* 9, 497–503. doi: 10.1162/105474600566989

Vasconcelos-Raposo, J., Melo, M., Barbosa, L., Teixeira, C., Cabral, L., and Bessa, M. (2021). Assessing presence in virtual environments: adaptation of the psychometric properties of the presence questionnaire to the portuguese populations. *Behav. Inform. Technol.* 40, 1417–1427. doi: 10.1080/0144929X.2020.1754911

Wei, W., Li, S., Okada, S., and Komatani, K. (2021). *Multimodal User Satisfaction Recognition for Non-Task Oriented Dialogue Systems*. New York, NY: Association for Computing Machinery. doi: 10.1145/3462244.3479928

Witmer, B. G., and Singer, M. J. (1998). Measuring presence in virtual environments: a presence questionnaire. *Presence Teleoperat. Virtual Environ.* 7, 225–240. doi: 10.1162/105474698565686

Yu, C., Aoki, P. M., and Woodruff, A. (2004). Detecting user engagement in everyday conversations. *Proceedings of Interspeech 2004*. doi: 10.21437/Interspeech.2004-327

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.