Multimodal cues of the sense of presence and co-presence in human-virtual agent interaction

Jérémie Bousquet
Aix-Marseille Unversité
jeremie.bousquet@gmail.com

Magalie Ochs
LIS R2I
magalie.ochs@lis-lab.fr
Philippe Blache
Laboratoire Parole et Langage
philippe.blache@lpl-aix.fr

Abstract

In this study we explore a corpus of human-agent interactions collected in a task-oriented context : a virtual environment aiming at training doctors to break bad news to a patient, the latter being a virtual agent. Values, indicators of the sense of "presence" and of "co-presence", were assigned to each interaction thanks to subjective post-experience questionnaires filled by the human playing the role of doctor. In this article we aim at exploring the possibility of predicting user's sense of presence and co-presence, thanks to multi-modal cues captured from human-virtual agent interactions. For this classification task we considered two learning algorithms, Random Forests and Support Vector Machines (SVM), and we provide some figures on their efficiency on this corpus. Moreover, this study presents some results regarding importance of the different modalities and selection among features, the interest of considering not only human, but also agent's behavior and cues (in particular for co-presence prediction), verbal and non-verbal cues importance, or the effect of time segmentation of the discourse into three phases (greetings, main, closure), compared to considering an interaction as a whole.

1 Introduction

In Human-Agent Interaction domain, one key aspect is the evaluation of the user's experience. Presence is the feeling of being present in the virtual environment, and can be considered as psychological immersion in virtual environment. Co-presence, also commonly called social presence, can be defined as "the sense of being and acting with others in a virtual space". A way to measure sense of presence and co-presence is to use subjective questionnaires, filled by participants. Another approach would be to identify objective measures of presence, in our case in trying to correlate directly subjective measure of presence reported through questionnaires and behavioral cues of the users. In this article, we propose an approach correlating subjective measures of presence and user's multi-modal behavior cues using machine learning techniques. Sense of presence and co-presence are important concerns when trying to improve quality of human-agent interactions.

This article is rooted in Ochs et al. (2018), based on same corpus consisting of data captured during human-agent interactions through different sensors and means (sound and video, transcriptions, movements), topic of interactions being a doctor breaking bad news to a patient. Scores for sense of presence and co-presence are also part of this corpus thanks to subjective questionnaires filled by participants.

In this article we will experiment intuition that collected behavioral cues can effectively be used to predict the sense of presence and co-presence. We will also study the importance of splitting the interactions in phases (greetings, main, closure), and if verbal and non-verbal cues should be related to these phases for our classification task. Finally our work will compare the relative importance of verbal and non-verbal cues, for predicting the sense of presence and co-presence.

Amount of data samples collected for this corpus is fairly low (about a hundred), and both the subjective questionnaires and the conversation topic are very specific to this corpus of data, as such no other publicly available data-set could be used to directly contribute to our prediction task. Some elements will be presented regarding the efficiency and reliability of the automatic learning algorithms used for the prediction task, in this constrained context.

2 State of the art

Original subjective scoring of sense of presence and of co-presence was based on the *IGroup Presence Questionnaire*, IPQ (Schubert, 2003)) and (Bailenson et al., 2005) for *co-presence*.

Previous research works have shown a correlation between body movements and presence (Slater and Steed, 2000; **?**). We propose an approximation for the identification of such non-verbal cues based on the *entropy* of participants' movements in the virtual environment. Entropy is a common measure in virtual reality to assess the movements of the participants (Maïano et al., 2011). To obtain the entropy of the curve defined by the movement of each tracker on the participant, we followed method described in (Dodson et al., 2013).

3 Method and data

3.1 Presence and co-presence

In order to evaluate participant's experience, each participant had to fill a questionnaire on their subjective experience to measure their sense of *presence* (with the *IGroup Presence Questionnaire*, IPQ (Schubert, 2003)) and their sense of *copresence* (Bailenson et al., 2005) (Section 2.1). Each questionnaire provides a score ranging from 1 to 5.

Based on these scores of the two questionnaires, we consider three levels of presence and co-presence : a *low level* ($score \in [1, 2.5]$), an *average level* $score \in]2.5, 3.5[$ and a *high level* of presence and co-presence ($score \in [3.5, 5]$), and our machine learning task will be a classification task.

3.2 Corpus

The corpus has been collected in the context of a project aiming at developing a *virtual reality environment* to train doctors to *break bad news to a virtual patient* (for details on the project, see Ochs et al. (2018)). In order to collect data with different levels of immersion, the virtual patient, with which the doctors can interact in natural language, has been implemented on different virtual reality displays (see Figure 1). The human participants could be real doctors or novices.

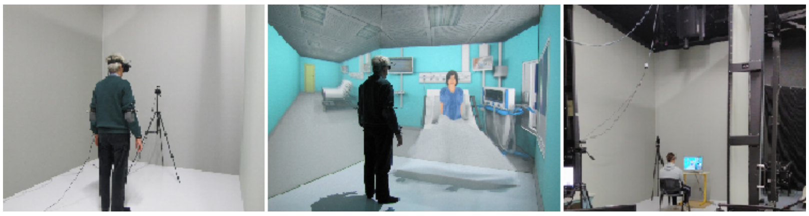

FIGURE 1 – Interactions with different virtual environment displays : virtual reality headset, virtual reality room, PC.

Different equipments allow to record the following data for each interaction :

— a video of the participant during her/his interaction with the agent ;

— time-series three-dimensional Unity coordinates of 5 trackers located on the participant's head, left and right elbows, and left and right wrists (also available for the virtual agent) ;

— separate audio files with the voice of the participant, and of agent, along with transcriptions from an Automatic Speech Recognition (ASR) system.

In total, the data contains 114 human-agent interactions. However, due to technical recording problems, some interactions have not been integrated in the corpus. After filtering the corpus is finally composed of 75 human-agent interactions. Table 1 shows how the levels of presence, and of co-presence (the classes for our prediction task), are distributed in terms of number of samples - showing imbalanced data with fewer data with a "low" sense of presence.

Class	Description	Presence count	Co-presence count
1	Low	14	7
2	Medium	29	36
3	High	35	35

TABLE 1 – Distribution of classes among samples for presence and co-presence

3.3 Features engineering (early fusion)

From this "raw" corpus, we extract a few high-level features that will be used for machine learning. As such we implement this multi-modal learning task using "early fusion" : highly dimensional data of different modes (here, text, sound, body parts locations ...) are first converted to high-level features through independent mechanisms, then concatenated to form the data-set used for training. We'll see below how those high-level features are computed from the raw corpus, as well as how this integrates with the concept of phases.

Extraction of the various high-level features listed hereafter, is fully automatized thanks to a set of Python scripts, and the use of the tools Marsatag (Rauzy et al., 2014) and SPPAS (Bigi, 2012) for the verbal features. As a result of script execution, a matrix with high-level features as columns, and samples as rows, is generated. To conduct the experiments wished for this study, the script had to be executed four times to generate high-level features for doctor or for agent, and for both, with and without phases split. Due to some data not correctly treated by automatic process (because of low quality of sound or other sources of errors), and to the fact that from the four sets of data are only the kept the ones in common (based on candidate ID and virtual environment), the corpus is reduced to 78 samples.

Segmentation of the interaction into three phases. The interaction between the participants and the virtual patient can be split into 3 phases : the beginning, the central part, and the conclusion. We make the hypothesis that verbal and nonverbal behaviors may differ depending on the phases of the interaction. We defined the size of each phase relatively to the total duration of the interaction. As a first approximation, we defined the duration of each phase as follows : 15% of the total conversation for the introduction, 70% for the main part, and 15% for the conclusion. A script makes it possible to automatically compute the high-level features (described below) with different phases segmentation (specifying split percentages), or without any phase segmentation.

Lexical richness and linguistic complexity. Our hypothesis is that the level of presence and co-presence have consequences on the elaboration level of the discourse : the higher these feelings, the more elaborated the discourse. At the verbal level, such elaboration is revealed by different phenomena in particular the sue of more modifiers (that we call *lexical richness*) and the use of more embedded constructions (called here *linguistic complexity*). An approximation of these high-level features is based on the distribution of the part-of-speech tags, for each participant and each phase of the interaction. Using a specific tool called SPPAS (Bigi, 2012), we performed a tokenization followed by a phonetization on the transcription file. The part-of-speech (POS) tags were automatically identified using MarsaTag (Rauzy et al., 2014). MarsaTag is 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 token. We consider 9 parts-of-speech tags : adjective, adverb, auxiliary, conjunction, determiner, noun, preposition, pronoun, verb.

Based on these POS tags, we computed two high level features for each phase :

Lexical richness :

$$ratio1 = \frac{\text{nb adjectives+nb adverbs}}{\text{nb tokens}}$$

Linguistic complexity :

$$ratio2 = \frac{\text{nb conjunctions+nb prepositions+nb pronouns}}{\text{nb tokens}}$$

Length of sentences We compute the average length of sentences (defined as syntactically homogeneous sequences and automatically identified with Marsatag) in each phase of the interaction for each participant. The length corresponds to the number of words of a sentence. Those features are prefixed "Avg_SentenceLength" in figures.

Body movements. Following the method described in (Dodson et al., 2013), we have computed the upper-bound on the Shannon entropy of curves of each plane (x, y and z) and each tracked point (head, left wrist, right wrist, left elbow, and right elbow). The different entropy values are averaged to obtain two features : the average movements of the head, and the average movement of the arms. Those features are prefixed "HeadEntropy" and "Avg_HandEntropy" in figures.

Sets of features In addition to the features extracted as above, we consider the *duration of interaction* and the *expertise of participant* (ie, doctor or novice).

Figure 2 presents the sets of features defined for our experiments (white cells), with feature sets used for learning as rows, and features as columns.

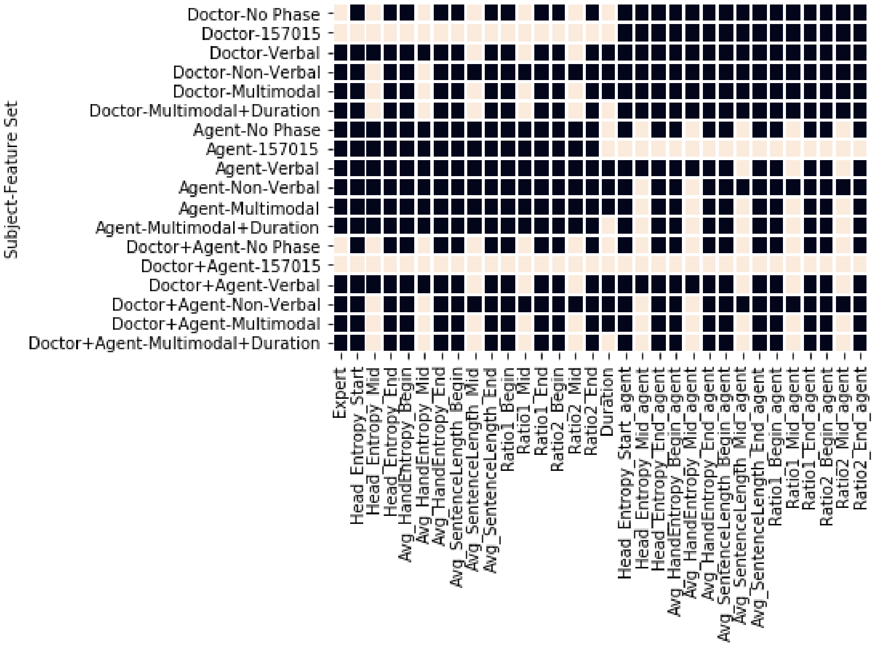

FIGURE 2 – Definition of features sets.

3.4 Method

For each data-set combination, and feature sub-set, and for each classifier (except where indicated), we apply the same methodology to obtain and compute scores for predicting sense of presence, and sense of co-presence (which are two different prediction tasks).

Class imbalance We saw that we have a problem of class imbalance. To solve imbalance we first upsample data by duplicating samples from minority class(es) in order to have same number of samples for each class. Downsampling is clearly not an option here given limited number of samples in total.

Hyper-parameters Specifically for Random Forests classifiers, we first want to choose an optimal value for number of decision trees to be grown. We know from Breiman (2001) that increasing the number of trees does not increase risk of over-fitting, so to find a good balance we plot a validation curve with the Random Forest classifier : we compute 10-fold cross validation scores for a range of number of trees from 1 to 1000 (sklearn's n_estimators). The result is pictured in figure 3. As a conservative approach we used value 300 for number of trees in further experiments.

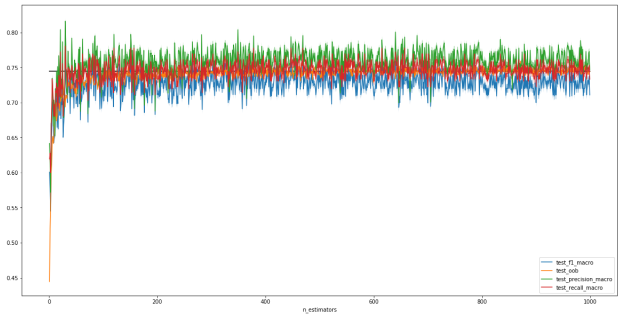

FIGURE 3 – Validation curve for Random Forest classifier (doctor features, no phase, presence prediction).

Note : here and after, specifically for Random Forests we also compute oob (out-of-bags) accuracy score. As described in Breiman (2001), oob can be considered a good measure of generalization of the ensemble (oob accuracy is computed based on samples left out when growing decision trees. Each tree being trained on bootstrap sample with replacement of the original data).

Then, we search for best hyper-parameters maximizing f1 score by doing 10-folds cross-validation on the hyper-parameters ranges in tables 2 and 3 [1]. Data from high-level features is provided as it is for Random Forests classifier and is centered and scaled for SVM.

Parameter	**Explored values**
n_estimators	300
max_features	[None, sqrt, log2]
max_depth	[None, 2, 5, 10]
min_samples_leaf	[1, 5, 10]

TABLE 2 – Hyper-parameters spaces considered for grid searching for Random Forests classifier.

Parameter	**Explored values**
gamma	[10^{-3}, 10^{-2}, 10^{-1}, 1, 10]
C	[10^{-2}, 10^{-1}, 1, 10]
kernel	[linear, rbf, sigmoid]

TABLE 3 – Hyper-parameters spaces considered for grid searching for SVM classifier.

Interestingly, hyper-parameters search was during a first run executed on a range of values for n_estimators (number of trees), instead of using the same value of 300 for all cases. Found hyper parameters showed then great variations (while computed scores did not). After setting a sufficiently large number of trees, best hyper-parameters found were very close from each other and showed little variations. It is also interesting to notice that kernel that showed best results for SVM is systematically gaussian (rbf). It is an indication that our classes are probably not linearly separable. Distinct sets of best hyper-parameters found are reproduced for the phases versus no phase experiments only in Annex B.

Prediction scores To compute prediction scores, we trained each classifiers with the best parameters found in previous step. To minimize variance, we repeated ten times, ten-folds cross validations, the final scores being the averages of the hundred scores computed. Given the low number of samples available, we did not kept aside part of the data to form a real test set, hence had to rely on cross validation scores. The scores computed are precision, recall, and f1 (and oob for Random Forests classifier only), macro average and per class.

[1] Meaning of parameters values as of sklearn's RandomForestClassifier and SVC documentation.

Note : it may be considered useless to perform 10-fold cross validation for Random Forests as we could have merely relied on oob (as it is a validation score), still the same method was applied on both classifiers, and the same scores were computed for comparable results between classifiers.

4 Experiences and results

4.1 Phases versus no phase

Our first experiment was to predict sense of presence, then of co-presence, using features computed on the complete interactions (no phasing), based on features relative to the subject/participant alone, or the virtual agent alone, or both. We also computed scores in the same context but with features generated from interactions splitted in three phases.

Results can be seen in tables 4 and 5.

Metric	Subject	Phases	Classifier	f1	precision	recall	oob
Presence	**Subject**	**No Phase**	**Random Forests**	0.76	0.79	0.77	0.70
			SVM	0.85	0.88	0.86	
		Phases	**Random Forests**	0.73	0.79	0.76	0.72
			SVM	0.82	0.87	0.84	
	Agent	**No Phase**	**Random Forests**	0.71	0.73	0.73	0.68
			SVM	0.78	0.85	0.80	
		Phases	**Random Forests**	0.73	0.77	0.74	0.71
			SVM	0.86	0.91	0.86	
	Subject+Agent	**No Phase**	**Random Forests**	0.73	0.75	0.75	0.69
			SVM	0.77	0.84	0.78	
		Phases	**Random Forests**	0.77	0.80	0.78	0.77
			SVM	0.83	0.87	0.84	

TABLE 4 – Presence prediction scores in phases vs no phase experiment

Metric	Subject	Phases	Classifier	f1	precision	recall	oob
Co-Presence	**Doctor**	**No Phase**	**Random Forests**	0.71	0.75	0.73	0.74
			SVM	0.93	0.94	0.93	
		Phases	**Random Forests**	0.75	0.75	0.76	0.79
			SVM	0.84	0.85	0.84	
	Agent	**No Phase**	**Random Forests**	0.82	0.83	0.83	0.83
			SVM	0.85	0.91	0.86	
		Phases	**Random Forests**	0.75	0.77	0.76	0.77
			SVM	0.84	0.88	0.86	
	Doctor+Agent	**No Phase**	**Random Forests**	0.89	0.91	0.90	0.89
			SVM	0.88	0.93	0.89	
		Phases	**Random Forests**	0.85	0.86	0.85	0.84
			SVM	0.87	0.93	0.89	

TABLE 5 – Co-presence prediction scores in phases vs no phase experiment

Obtained scores demonstrate far better results than we would obtain at random for three classes, we can conclude that we effectively could predict sense of presence and co-presence from multimodal cues extracted from our human-agent interactions corpus.

Regarding split of interactions in phases, it is more difficult to conclude : in some cases it seems beneficial, in others not. If we consider only oob scores, phasing positively impacts presence prediction, and negatively impacts co-presence prediction.

SVMs scores are consistently higher than Random Forest'.

4.2 Verbal, non-verbal, multimodal features

Our second experiment was to define subsets of features corresponding to either verbal cues, non-verbal cues, or both (see also figure 2).

Results can be seen in tables 6 and 7.

Metric	Subject	Mode	Classifier	f1	precision	recall	oob
Presence	**Doctor**	**Verbal**	**Random Forests**	0.68	0.74	0.71	0.73
			SVM	0.78	0.82	0.79	
		Non-Verbal	**Random Forests**	0.73	0.77	0.76	0.73
			SVM	0.78	0.78	0.81	
		Multimodal	**Random Forests**	0.71	0.71	0.74	0.70
			SVM	0.82	0.88	0.82	
	Agent	**Verbal**	**Random Forests**	0.64	0.64	0.66	0.63
			SVM	0.81	0.84	0.81	
		Non-Verbal	**Random Forests**	0.64	0.67	0.67	0.70
			SVM	0.70	0.75	0.72	
		Multimodal	**Random Forests**	0.73	0.74	0.76	0.71
			SVM	0.85	0.91	0.86	
	Doctor+Agent	**Verbal**	**Random Forests**	0.64	0.65	0.66	0.67
			SVM	0.86	0.91	0.86	
		Non-Verbal	**Random Forests**	0.63	0.67	0.66	0.68
			SVM	0.74	0.77	0.76	
		Multimodal	**Random Forests**	0.73	0.74	0.76	0.75
			SVM	0.82	0.91	0.84	

TABLE 6 – Presence prediction scores in verbal vs non-verbal experiment

Results when including both verbal and non-verbal are better which was intuitively expected (except for doctor features alone when predicting sense of presence). We also can see that non-verbal features (ie, entropies of movements curves) lead to better scores than verbal features.

4.3 Learning curves

Considering the limited number of data samples available, even if we performed cross-validation, it is worth considering if we suspect over-fitting of our classifiers on our data-sets.

To gain some view on this aspect, we decided to compute and plot the learning curves for all testing cases. To do so, we trained the classifiers on the same sets of best hyper-parameters found, using ten-folds cross validation, this time using an increasing number of data samples (after upsampling was performed). The range of counts of data samples used was between 35 and the total available. Scoring function used was negative log loss (the closer to zero the better).

We show in figures 5 and 4 the results for SVMs and Random Forests in the phase vs no phase experiments (results for the verbal vs non-verbal experiments are available in Annex A). Bold lines correspond to mean values, and areas correspond to standard error on the mean.

From those figures we see a training loss either low either decreasing in most cases, and a validation loss decreasing with number of samples. From this we can expect that training would benefit from additional samples, and we see no evidence of over-fitting. Of course a real test set (and so sufficient number of data) would allow to more definitively answer this question (interestingly we see very specific and different patterns for the learning curves, if we compare SVMs and Random Forests).

5 Conclusion

We showed that it was possible to predict sense of presence and co-presence, using multi-modal cues extracted from a corpus of human-virtual agent interactions. We noticed that non-verbal features (upper

Metric	Subject	Mode	Classifier	f1	precision	recall	oob
Co-Presence	**Doctor**	**Verbal**	**Random Forests**	0.71	0.74	0.74	0.75
			SVM	0.87	0.91	0.88	
		Non-Verbal	**Random Forests**	0.80	0.82	0.81	0.79
			SVM	0.72	0.76	0.73	
		Multimodal	**Random Forests**	0.82	0.83	0.84	0.80
			SVM	0.85	0.91	0.86	
	Agent	**Verbal**	**Random Forests**	0.71	0.74	0.73	0.74
			SVM	0.83	0.87	0.84	
		Non-Verbal	**Random Forests**	0.80	0.83	0.81	0.78
			SVM	0.71	0.75	0.73	
		Multimodal	**Random Forests**	0.82	0.86	0.83	0.80
			SVM	0.86	0.90	0.86	
	Doctor+Agent	**Verbal**	**Random Forests**	0.73	0.75	0.74	0.70
			SVM	0.85	0.89	0.86	
		Non-Verbal	**Random Forests**	0.81	0.85	0.81	0.78
			SVM	0.79	0.82	0.79	
		Multimodal	**Random Forests**	0.84	0.88	0.85	0.81
			SVM	0.87	0.92	0.88	

TABLE 7 – Co-presence prediction scores in verbal vs non-verbal experiment

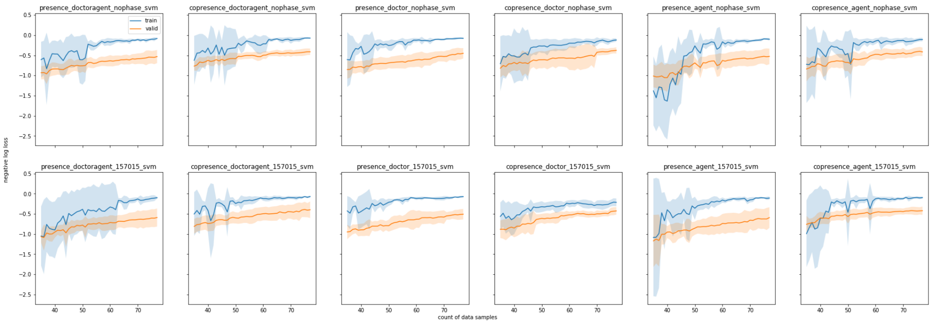

FIGURE 4 – Learning curves for SVM classifiers in phase vs no phase experiments.

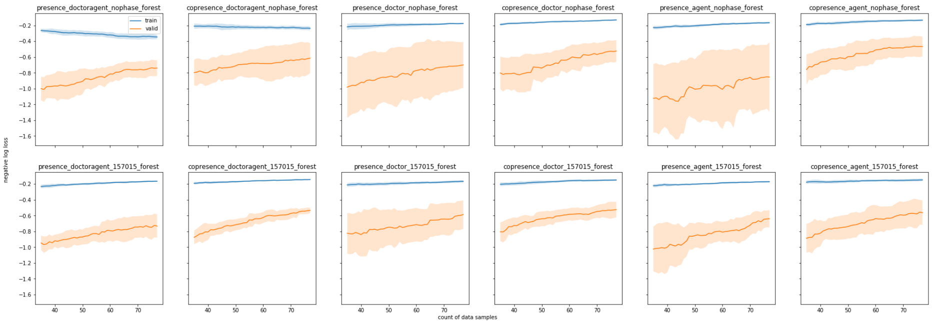

FIGURE 5 – Learning curves for Random Forests classifiers in phase vs no phase experiments.

bound of Shannon entropy of curves) led to more accurate predictions. We also brought some elements on the quality of learning of machine learning algorithms used (Random Forests and SVM).

To go further, it could be interesting to consider other classifiers also recommended when number of samples is limited (for example, Naive Bayes). It is worth noticing that this study makes limited use of the time series available in original corpus (through the concept of phases), the interest and feasibility of alternatives to early fusion approach and to classical classifiers used here could be studied. Of course, obtaining more data samples would benefit to confirm the results of this study, maybe by trying to fix or improve quality of badly captured samples from the original corpus, which might be a tedious process, but computed learning curves confirmed that any additional sample may be useful to the learning task.

Also, some experiments in order to gain insight on features importances at the feature level could be conducted. Features importances is noticeably generated natively for Random Forests classifiers (with sklearn), another method would have to be experimented for SVM if possible.

References

Jeremy N Bailenson, Kim Swinth, Crystal Hoyt, Susan Persky, Alex Dimov, and Jim Blascovich. 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(4) :379–393.

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

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1) :5–32.

Michael Maurice Dodson, Michel Mendes France, and Michel Mendes. 2013. On the entropy of curves. *Journal of Integer Sequences*, 16(2) :3.

Christophe Maïano, Pierre Therme, and Daniel Mestre. 2011. Affective, anxiety and behavioral effects of an aversive stimulation during a simulated navigation task within a virtual environment : A pilot study. *Computers in Human Behavior*, 27(1) :169–175.

Magalie Ochs, Sameer Jain, and Philippe Blache. 2018. Toward an Automatic Prediction of the Sense of Presence in Virtual Reality Environment. In *6th International Conference on Human-Agent Interaction (HAI-2018)*, Southampton, United Kingdom, December.

Stéphane Rauzy, Grégoire Montcheuil, and Philippe Blache. 2014. Marsatag, a tagger for french written texts and speech transcriptions. In *Proceedings of Second Asian Pacific Corpus linguistics Conference*, page 220.

Thomas W Schubert. 2003. The sense of presence in virtual environments : A three-component scale measuring spatial presence, involvement, and realness. *Zeitschrift für Medienpsychologie*, 15(2) :69–71.

Mel Slater and Anthony Steed. 2000. A virtual presence counter. *Presence : Teleoperators & Virtual Environments*, 9(5) :413–434.

Annexe A. Learning curves

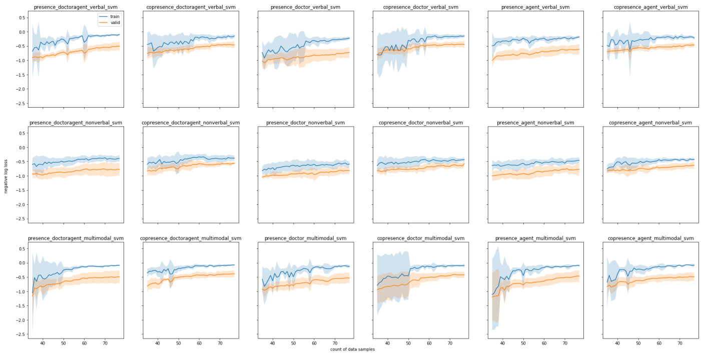

FIGURE 6 – Learning curves for SVM classifiers in verbal vs non verbal experiments.

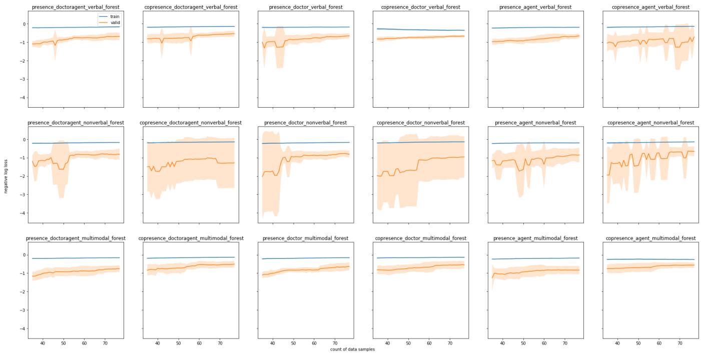

FIGURE 7 – Learning curves for Random Forests classifiers in verbal vs non verbal experiments.

Annexe B. Hyper-parameters

Classifier	**Best Params**
Random Forests	{’max_features’ : None, ’n_estimators’ : 300, ’max_depth’ : None, ’min_samples_leaf’ : 1}
	{’max_features’ : None, ’n_estimators’ : 300, ’max_depth’ : 10, ’min_samples_leaf’ : 1}
	{’max_features’ : ’sqrt’, ’n_estimators’ : 300, ’max_depth’ : 5, ’min_samples_leaf’ : 1}
	{’max_features’ : ’sqrt’, ’n_estimators’ : 300, ’max_depth’ : None, ’min_samples_leaf’ : 1}
	{’max_features’ : ’sqrt’, ’n_estimators’ : 300, ’max_depth’ : 10, ’min_samples_leaf’ : 1}
SVM	{’svc__gamma’ : 10.0, ’svc__kernel’ : ’rbf’, ’svc__C’ : 1.0}
	{’svc__gamma’ : 1.0, ’svc__kernel’ : ’rbf’, ’svc__C’ : 1.0}
	{’svc__gamma’ : 1.0, ’svc__kernel’ : ’rbf’, ’svc__C’ : 10.0}
	{’svc__gamma’ : 0.1, ’svc__kernel’ : ’rbf’, ’svc__C’ : 10.0}

TABLE 8 – Distinct best parameters from gridsearch for phases vs no phase experiments