WACAI 2014

Worshop **A**ffects, **C**ompagnons **A**rtificiels, **I**nteraction

Gérard Bailly, Magalie Ochs, Alexandre Pauchet

Rouen, 30 Juin et 1er Juillet 2014

Preface

L'objectif du workshop **WACAI 2014** est de réunir les recherches et développements en cours autour des *Agents Conversationnels Animés (ACA) et des robots sociaux*. Cette rencontre est plus particulièrement centrée sur l'étude, la modélisation, le développement et l'évaluation de l'interaction de systèmes interactifs avec leurs partenaires (humains ou artefactuels). A l'interface entre sciences de l'ingénieur et sciences humaines et sociales, l'ambition de WACAI est d'appréhender l'interaction personne-système dans toute sa complexité (biologique, linguistique, sociale, culturelle et émotionnelle).

Après les précédentes éditions bi-anuelles du workshop WACA, organisées successivement à Grenoble (2005), à Toulouse (2006), à Paris (2008), à Lille (2010) et WACAI à Grenoble (2012), cette nouvelle édition se déroulera à Rouen le 30 Juin et 1er Juillet 2014.

30 Juin et 1er Juillet 2014
Rouen

Gérard Bailly
Magalie Ochs
Alexandre Pauchet

Table of Contents

A user study on a new Super-Wizard of Oz platform explored in a long-distance survey context 1
Ritta Baddoura, Gentiane Venture and Guillaume Gibert

Virtual conversational agents and social robots: converging challenges ... 7
Gerard Bailly, Magalie Ochs, Alexandre Pauchet and Humbert Fiorino

Implantation d un ACA narrateur (Démonstration) 16
William Boisseleau, Ovidiu Serban and Alexandre Pauchet

Alignement par Production d'Hétéro-Répétitions chez un ACA 18
Sabrina Campano, Nadine Glas, Caroline Langlet, Chloé Clavel and Catherine Pelachaud

MOCA-RT : Une Plateforme Distribuée et Ouverte pour les Compagnons Arti ciels 24
Matthieu Courgeon, Caroline Faur, Wafa Johal, Céline Jost, Florian Pecune, Sylvie Pesty and Dominique Duhaut

Laughing Body 26
Yu Ding, Thierry Artière Artieres and Catherine Pelachaud

The characterization of emotional body expression in di erent movement tasks 28
Nesrine Fourati and Catherine Pelachaud

Approche basée sur les traces modélisées pour agents socio-émotionnels dans les jeux vidéo 30
Joseph Garnier, Jean-Charles Marty, Karim Sehaba and Elise Lavoue

Engagement-based Politeness Indications for Virtual Agents 36
Nadine Glas, Ken Prepin and Catherine Pelachaud

E et Proteus et amorçage : Quand l apparence des personnages virtuels in uence les comportements et les attitudes des utilisateurs 41
Jérôme Guegan and Stephanie Buisine

Procedural Animation Pipeline For Virtual Agent System 49
Jing Huang and Catherine Pelachaud

Les Styles pour la Plasticité des Robots Compagnons 51
Wafa Johal, Sylvie Pesty and Gaelle Calvary

Intentions stratégiques basées sur un modèle a ectif et une théorie de l esprit 59
Hazael Jones and Nicolas Sabouret

Expressing social attitudes in virtual agents for social coaching 65
Hazaël Jones, Mathieu Chollet, Magalie Ochs, Nicolas Sabouret and Catherine Pelachaud

Etablissement de relations entre émotions, couleurs subjectives et couleurs objectives à partir d'annotations spontanées 71
Marie-Jeanne Lesot and Marcin Detyniecki

Modeling sensory-motor behaviors for social robots 77
Alaeddine Mihoub, Gerard Bailly and Christian Wolf

Robots sociaux : design et recherche aux frontières de l'expérimentation .. 83
Ioana Ocnarescu and Frédérique Pain

Modélisation de l'influence de la personnalité d'un compagnon artificiel sur ses attitudes sociales 85
Florian Pecune, Caroline Faur, Magalie Ochs, Céline Clavel, Catherine Pelachaud and Jean-Claude Martin

Virtual Interactive Behavior : une architecture modulaire pour ACA 91
André-Marie Pez, Pierre Philippe, Brice Donval and Catherine Pelachaud

Machine Learning for Interactive Systems : Challenges and Future Trends 93
Olivier Pietquin and Manuel Lopes

Nonverbal behavior of a virtual agent expressing attitudes in a group 101
Brian Ravenet, Angelo Cafaro, Magalie Ochs and Catherine Pelachaud

Caractérisation d'unités gestuelles en vue d'une interaction humain-avatar . 107
Ilaria Renna, Sebastien Delacroix, Fanny Catteau, Coralie Vincent and Dominique Boutet

Un modèle logique de théorie de l'esprit pour un agent virtuel dans le contexte de simulation d'entretien d'embauche 114
Nicolas Sabouret and Marwen Belkaid

Évaluation subjective d'un modèle BDI de Théorie de l'esprit 120
Nicolas Sabouret, Atef Ben Youssef and Sylvain Caillou

Perception, langage, curiosité : éléments clés pour la conceptualisation de connaissances en robotique interactive 126
Christophe Sabourin and Kurosh Madani

Adaptation in an Interactive Model designed for Human Conversation and Music Improvisation: a preparatory outline 132
Kevin Sanlaville, Frédéric Bevilacqua, Catherine Pélachaud and Gérard Assayag

MyBlock/AgentSlang : une plateforme pour le déploiement d'ACA 138
Ovidiu Serban and Alexandre Pauchet

A user study on a new Super-Wizard of Oz platform explored in a long-distance survey context

Ritta Baddoura[1, 2] Gentiane Venture[3] Guillaume Gibert[1, 2]

[1]INSERM U846, Stem-Cell and Brain Research Institute, Bron, France
[2]Université de Lyon, Université Lyon 1, 69003 Lyon, France
[3]Tokyo University of Agriculture and Technology, Tokyo, Japan

rittabaddoura@yahoo.fr

Abstract

SWoOZ is a new super Wizard of Oz (WoZ) research platform developed to study human-robot interaction (HRI) and mediated human-human interaction. A humanoid robot is used as a proxy between two humans. An experimenter is bound with this proxy and fully controls its head motion with his own movements (live and free of attached sensors). Manipulations can be applied to any motion leaving the rest of the dynamics untouched. This paper presents preliminary results of a user study aiming at evaluating the platform's usability, efficiency and likability. The experimental scenario consists of a realistic long-distance survey conducted by a researcher who interviews Japanese participants on cultural topics (non-deceptive WoZ). The study addresses the possible effects of the remote user's previous experience with robotics (naive vs. non-naive) on the participants' evaluation of the platform.

Keywords

Wizard of Oz; Telerobotics; Social Robotics; Naive vs. non-naive user; Head motion.

Résumé

SWoOZ est une nouvelle plateforme de Super magicien d'Oz (WoZ) visant à l'étude des interactions homme-robot (HRI) et des interactions humain-humain médiatisées. Un robot humanoïde est utilisé comme intermédiaire entre deux humains. L'expérimentateur est lié au robot dont il contrôle entièrement les mouvements de la tête à partir de ses propres mouvements (en direct et sans l'usage de capteurs attachés). Des manipulations peuvent être appliquées à n'importe quel mouvement, laissant intact le reste de la dynamique. Cet article présente les résultats préliminaires d'une étude visant à évaluer la facilité d'utilisation de la plateforme, ainsi que son efficacité et son appréciation du point de vue des utilisateurs. Le scénario expérimental consiste en une enquête menée à distance par un chercheur qui interroge, par le biais de la plateforme, des participants japonais sur des sujets culturels (les participants savent que le robot est téléopéré par un humain). L'étude envisage les effets possibles de l'expérience préalable avec les robots de l'enquêteur (naïf vs. non-naïf) sur les participants.

Mots-Clés

Magicien d'Oz ; Télé-robotique ; Robotique sociale ; Utilisateur naïf vs. non-naïf ; Mouvement de la tête.

1 Introduction & Motivation

During the last few years, there has been a growing attention on exploring teleoperation and telepresence as well as the effects of culture in the social robotics and in the human-robot interaction (HRI) fields. In today's global village where distances are a major component of many personal and professional daily realities, many recent studies focus on showing the interest and the added value of using teleoperated [1, 2] and telepresence robots in many various fields such as remote education [3], health care environments, independent living for the elderly, offices [4], and industrial or military operations that are lead in uncertain and unknown environments [5]. As for the effects of culture, studies address topics such as verbal and non-verbal communication styles [6, 7], user preferences and attitude [8], user beliefs [9] and perception of the robot particularly of its social presence [10, 11], user attribution of personality traits to it [12], and interpretation of facial expressions [13], head motion [14, 15], gaze and gestures [7] and body posture [16] expressed by the robot.

Currently, different kinds of robots with different appearances, capacities and autonomy-levels, are being developed and studied in order to achieve various tasks in different environments. Thus, the importance of building socially-competent robots and mastering the key components of a satisfying and successful interaction with humans has grown wider. The Wizard of Oz (WoZ) technique has been frequently used in this perspective by researchers in the fields of HRI. More precisely, as underlined by [1], WoZ is usually employed to compensate for the robot's insufficient social and/or technical abilities, hence allowing for a smoother interaction and an enhanced vision of future design improvements.

To study human-robot interaction and human-human mediated interaction, and rather than proposing a set of predefined behaviors to be selected by the wizard as in the classical WoZ [1], we developed an enhanced WoZ called SWoOZ (which stands for Super WoZ) setup that has the capacity of mirroring face, eye and head motion on a robot and consequently allowing the generation of a spontaneous movements in order to support a more genuine and realistic interaction [17, 18]. In this

platform, a humanoid robot is used as a proxy between two humans involved in dyadic interactions. One human, called here the remote user [1] (or the interviewer in relation to the interview task performed in our experiment) is bound with the humanoid robot as he controls in real-time and free of attached sensors, by simply performing his own movements, the eye, and face and head motion of this robot. The remote user perceives the scene almost as if he was present, instead of the robot, in the same room as his human interlocutor called here the local user (or the interviewee). The humanoid head motion for instance, as the human interaction partner sees it, is the direct translation of the wizard's motion which is accurately tracked and replicated by the robot with less than 200 ms delay.

SWoOZ can be used to manipulate specific movements without modifying the rest of the dynamics, thus giving an insight on the acceptable limits for the human partner for various parametric manipulations and interactions. Following the study proposed by [19] in a human-human interaction mediated by an avatar, we have first investigated the role of damping head movements during a human-human interaction mediated by a humanoid robot and found as expected that damping head movements affects the interaction [18]. Indeed, naive subjects interacting with a robot controlled in real-time by a confederate's head motion, increased their head movements when the robot's head motion was attenuated.

As reported in [20], many interactions take place simultaneously when humans are communicating through a telerobotic system as the one deployed by SWoOZ. These interactions include HRI between the human users and the remotely controlled communication humanoid proxy, human-human interaction, and human-computer interaction between the remote user and the local user's image on the screen, as it is the case in our setup. In order to further explore these different levels of interaction, as well as to evaluate the SWoOZ platform usability and efficiency when operated by different confederates in the frame of realistic interactive scenarios, we started a study whose preliminary results are presented here.

2 Methods

2.1 Experimental Setup & Equipment

The SWoOZ platform consists of: a) a system able to estimate the remote user's head pose (orientation and location) and rigid/non-rigid motion: The Random Forests Head Tracking system [21] is used in this experiment together with a consumer depth camera (ASUS/Xtion sensor) ; b) a software program to apply online manipulation to specific parameters; c) a humanoid robot: SWoOZ is compatible with the robot NAO (Aldebaran) and iCub (http://www.icub.org/), given that NAO is used in the current study. Once the data are estimated, they are sent to the robot that mimics the estimated remote user's head motion. Further information about the SWoOZ platform can be found in [17, 18] as well as on the SWoOZ Github page: https://github.com/GuillaumeGibert/swooz.

The remote user's voice captured by a microphone is transmitted to the local user interacting with the teleoperated robot through a small speaker positioned behind it. The head movements of both the remote and the local users are recorded synchronously with an IMU (Inertial Measurement Unit) system. These recorded data will be analyzed later. The IMU sensor is attached around the remote user's and the local user's heads for specific motion analysis such as intensity, jerkiness, velocity and frequency of the human users' movements (IMU sensors are different from the ASUS/Xtion sensor which is only used for the tracking and transmission of the remote user's head motion). To bind the remote user to the robot and enable him as much as possible to sense the scene as if he was seated in its place, auditory and visual feedbacks are transmitted to him using a High Definition (HD) webcam (Creative Live Cam Socialize HD) positioned behind the robot and binaural microphones (MS-TFB-2, The Sound Professionals, Inc.) discreetly placed on the robot's body.

2.2 Participants

Two remote users (one naive and another non-naive) and 20 naive participants (previous exposure to robots was controlled prior to the experiment) volunteered to take part in the study. The 22 candidates are Japanese students from Tokyo University for Agriculture and Technology (TUAT). All of them range in age from 19 to 25 years old. The naive interviewer (never used or interacted with robots or with WoZ setups) interviewed 14 participants (9 males; 5 females); this group will be referred as Participants X in the rest of the text. The non-naive interviewer (previously exposed to manipulating robots and to HRI, he has used NAO previously for research purposes) interviewed 6 participants (5 males; 1 female); this group will be referred as Participants Y in the rest of the text. The remote users/interviewers (both males) have both verbal and non-verbal (head motion) control on the robot, the latter having no autonomy at all. Both were trained to perform the interview (e.g. learning the questions, keeping their behavior consistent with the one a researcher would have, monitoring the participants' answers' duration) and were similarly instructed regarding the technical requirements necessary for the proper functioning of the SWoOZ setup (remain in the field of vision of the depth camera, sit straight etc.).

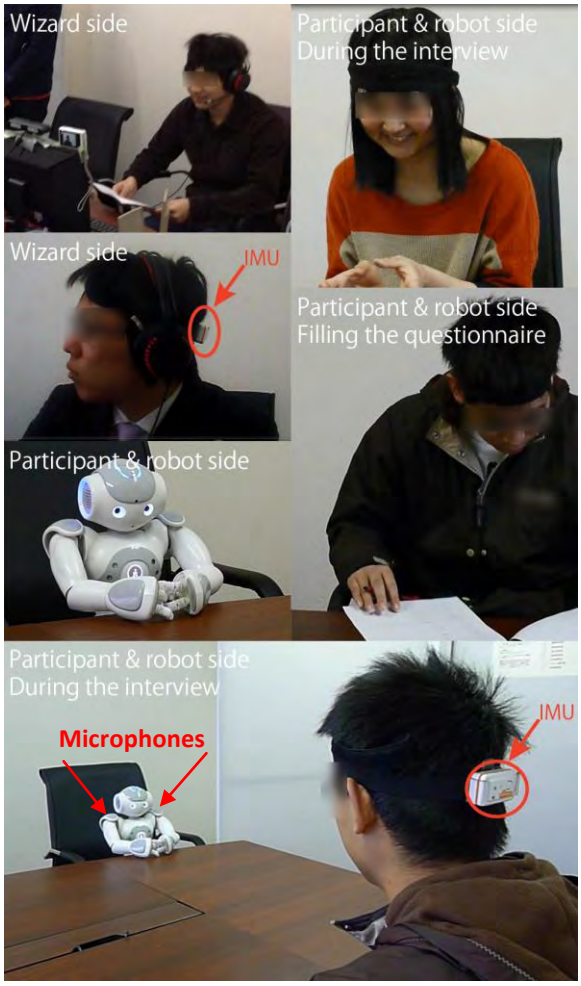
2.3 Materials, Procedure & Data collection

The scenario design aims at providing a realistic context to the experiment and consists of a cultural user study taking place on a Japanese university campus. The participants volunteered to participate in an anonymous survey lead by a Japanese researcher working in France. The survey investigates, through the participants' answers to an interview and a questionnaire, how the Japanese youth perceives the French and the Japanese cultures. The scenario mainly targets a fluid interaction between the remote user and the local user via the proxy, knowing that the contents of the interviewees'

[1] They are not mainly referred to as « wizards » to insist on the fact that the setup is used in a non-deceptive way.

answers (such as e.g. answer's duration, personal opinions or the amount of exact information on the French culture) are not important for the study.
The participants are informed about the following: a) The researcher is unable to be physically present and the interview will be live mediated by a humanoid robot – therefore the scenario involves no deception; b) The interview room is filmed using two cameras, the interviewer's and the interviewee's voices are recorded and IMU sensors are used for head motion capture; c) The same survey will be lead in France for cross-cultural comparison. The remote user is in room A while the robot NAO and the local user are facing each other, seated on either side of a table in a real University meeting room (room B) (see Figure 1).

Figure 1. The remote user (wizard) is in room A while the local user (interviewee) and the robot are in room B.



The interview scenario and the related questionnaire were carefully designed for this study, in accordance with Riek's reporting guidelines for WoZ studies in HRI [1], regarding various issues including social deception, rigorous and repeatable design, wizard training, constrained wizard recognition and production abilities, wizard errors, specified user instruction and behavior hypotheses. A pilot study was done on 4 participants in France and 3 participants in Japan to test and improve the questionnaire and the interview questions (including the syntax that had to be in accordance with the researcher's status and role, as well as with the Japanese cultural specificities).
When the participant is seated, the researcher/remote user presents himself and provides a recapitulation of the survey. The participant/local user is reminded that there are no false and right answers: only personal opinions are expected. Then the interview starts.
This oral part of the survey consists of 15 questions revolving around the specificities of the Japanese and French cultures, on their common points and their differences with some focus on communication and interaction questions. Both interviewers/remote users have been trained during three days to master the interview in regards to its contents, general duration, and to the style of questioning, but also in regards to the position in front of the depth camera sensor. More generally, the experimental design aims at defining a precise and repeatable conversational context in which head motion is spontaneously produced by both users; a context that is the same for the 20 interviewees and that validates the comparison between each experimental session.
The interview's overall duration is 10 min while the questionnaire needs 5 to 10 min to be filled (depending on the participant). The whole experiment lasts around 15 to 20 min. When the interview is completed, the remote user asks the participant to fill the questionnaire placed on the table. The questionnaire consists of 40 items divided into 5 sets. 4 sets use a 5-point Likert scale (where 0 = not at all and 4 = to a very high degree). The sets addressed in this paper assess the participant's evaluation of the robot as a proxy, more particularly the participants' evaluation of the robot's efficiency, likability and credibility. An open-ended question ends the questionnaire to allow the local user to express more freely and personally his/her feedback. The Cronbach's alpha of the participants' questionnaire is 0.91 which is above the generally acceptable level 0.7 [22] and shows a very good internal reliability. Additionally, each wizard was asked at the end of the whole round of interviews to fill a questionnaire in 9 items divided into 3 sets, in order to get his feedback on the experiment and on the SWoOZ platform.

3 Hypotheses & Preliminary Results

First of all, we are interested in getting a feedback on the SWoOZ platform efficiency from local users who are not familiar with it and who have no prior experience with robots or with WoZ setups. As for the remote users, we are interested in observing the possible effects of their previous experience with robots and HRI on the local users' feedback. Thus, we explore this experimental scenario with two samples: one is interviewed by a naive remote user and the other by a non-naive remote user. Therefore, we first hypothesize that the remote user's previous exposure to HRI will impact the interviewees' experience of the interaction as well as their ratings (H1) and that X evaluations (interviewed by the naive interviewer) of the proxy will

be significantly different from Y's (interviewed by the non-naive interviewer).

Regarding the local users' evaluation of the humanoid proxy, we focus in this paper on their ratings of its efficiency, usefulness, likability, engagingness, human-likeness and on their satisfaction with it. We are also interested in their evaluation of the humanoid robot's credibility as a proxy/mediator between them and the remote user. Based on the prior study performed with SWoOz [18]; as well as on its ability to enable the proxy to mirror the remote user's motion, we expect the participants' ratings of efficiency, usefulness and satisfaction to be above the average score (H2) (2 being the average on this 5-point Likert scale).

Also, given the fact that the robot has a close to natural head motion as it mirrors the remote user's spontaneous motion, and given the fact that the voice the interviewees hear is the remote user's human voice, we expect the local users to find the proxy engaging and likable thus rating it above the average score (H3). From another perspective, we make the assumption that NAO's credibility as a proxy representing a researcher (in regards to NAO's appearance and given role), as well as its human-likeness (NAO only moves its head but its eyes are rigid, and it has no mouth) to be poorly rated (H4). For instance, [23] showed that a robot's appearance affects its likability and that participants expect the robot appearance to match its task during an interview context.

We calculated the descriptive statistics (95% CI) based on the interviewees' ratings of the proxy's performance. Participants X and participants Y gave generally medium-to-low ratings (see Figure 2). Y (interviewed by the non-naive remote user) seem to have given more generous scores than X.

X found the human-like aspect of the robot to be very poor (X: M= 1.00, SD= 1.13) and considered that the robot failed in being credible (X: M= 1.28, SD= 1.03). Nevertheless, they found the proxy rather satisfying (X: M= 2.14, SD= 1.30), likable (X: M= 2.07, SD= 1.10) and useful (X: M= 2.5, SD= 1.11). Y gave also low scores to the robot's human-likeness (Y: M= 1.5, SD= 1.11) and efficiency (Y: M= 1.66, SD= 0.94) and found it unsatisfactory (Y: M= 2.00, SD= 0.81). Despite these ratings, they considered it to be likable (Y: M= 2.83, SD= 0.68), engaging (Y: M= 2.5, SD= 0.95) and useful (Y: M= 2.5, SD= 1.11).

We did a Mann-Whitney test to ascertain if the differences between X and Y scores are statistically significant, thus implying an effect of the remote user's previous experience with robots and HRI. The observed U-values failed to be significant at $p \leq 0.05$ (see Table 1), thus invalidating a possible effect of the remote user's previous experience on the participants' ratings.

As some recent studies have underlined the human partner gender effects in HRI [10, 24, 25], we also did a Mann-Whitney test to rule out possible gender effects on the participants' ratings. Results showed the gender factor to be statistically not significant, which can be possibly attributed to the low representation of women in both samples as well as to the small size of the samples.

Figure 2. X's and Y's evaluation of the SWoOZ proxy on a 5-point Likert scale (0 = not at all; 4 = to a very high degree)

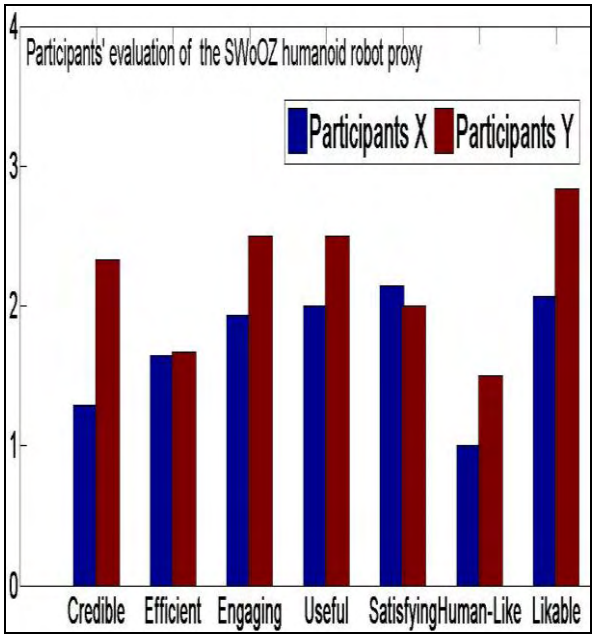


Table 1. Mann-Whitney test results based on X's and Y's evaluations of the proxy (U critical = 17)

The Proxy is / **U observed**	Credible	Efficient	Engaging
	22	41.5	32
Useful	Human-Like	Likable	Satisfying
32	32.5	31	38

4 Discussion and Conclusion

Generally, the participants showed a moderate to low appraisal of the proxy's performance. H3 was validated as the participants found the proxy's rather likable and engaging (among the highest scores for X and Y). H2 was partly validated since the participants were moderately satisfied with the proxy (the highest score for X). They found it rather useful as well. But despite that, they judged it as poorly efficient. H4 was partly validated: as expected, the proxy's human-likeness was poorly rated, but X and Y did not seem to have similar feedback on the robot's credibility. X average rating of this dimension was very low, whereas Y gave moderate ratings to it.

Nevertheless, the Mann-Whitney test results failed to validate the remote user's previous experience effect on the participants, thus infirming H1 and showing that the observed differences between X and Y ratings are

statistically insignificant. This might be probably due to the small size of the samples, especially of the one interviewed by the non-naive remote user. Therefore, it would be of real interest to proceed to further experiment with more participants, at least with the non-naive remote user, in order to reexamine the previous experience effect. Another explanation would be that the preparation/training phase of both remote users' has reduced the gap between them, but this does not seem to be a sufficient reason. Also, the mediation/teleoperation characteristics of the SWoOZ platform might mitigate the remote user's effect. Indeed, the proxy's mediation interferes in the human dyad and is in favor of rendering a rather homogeneous/similar behavior of the robot, especially that the remote user's head motion is only mirrored here. Thus, using a humanoid robot with richer face expressions, such as the iCub for example (it is able to move its eyes and mouth) might better render some differences between the remote users and enable us to more precisely assess the impact of their previous experience with robots on the participants.

These results are only preliminary and have to be completed with further analysis of the remaining experimental data. However, some important aspects are underlined such as the failure to prove significant effect of the remote user's previous experience with robots. Nevertheless, this result is interesting in regards to the platform's usability as it suggests that any naive person could, with some preparation, successfully use the SWoOZ platform and be as equivalently effective as a more experienced person. The question of the proxy's choice, in relation to its appearance and to its assigned task needs to be given more attention for credibility reasons. Last but not least, the participants' satisfaction and their moderate appreciation of the proxy's (and therefore of the SWoOZ device) usefulness and of its engaging and likable behavior, are encouraging feedbacks to work towards improving the features of the SWoOZ platform for more efficiency and smoother ability to mediate human-human interaction.

5 Perspectives and Future works

The analysis of the remaining parts of the questionnaire (that were not addressed here) and the processing of the IMU data to obtain various head motion characteristics, will be addressed in future papers to shed more light on this study's results. Furthermore, running other experiments using human-human non-mediated interaction, or videoconference mediated interaction, or another robot, would open up to constructive comparisons with our collected data. Investigating technical aspects such as the localization of the camera in the setup: behind the robot vs. on the head of the robot, could be also of interest to understand the impact of having egocentric vs. non-egocentric visual input on the remote user's sense of "immersion".

Acknowledgments

This work was supported by the ANR SWoOZ project (11PDOC01901) and the Tokyo University for Agriculture and Technology Techno-Innovation Park, Japan.

We wish to mostly thank Ryo Matsukata, Hiroshi Takaiwa, Manfred Corbeau, Takamune Izui from TUAT and Florian Lance from INSERM U846 for their valuable help in setting and running the experiment.

References

[1] L. D. Riek, "Wizard of Oz Studies in HRI: A systematic Review and New Reporting Guidelines," *Journal of Human-Robot Interaction* vol. 1, pp. 119-136, 2012.

[2] S. Nishio, H. Ishiguro, M. Anderson, and N. Hagita, "Representing Personal Presence with a Teleoperated Android: A Case Study with Family," in *AAAI Spring Symposium: Emotion, Personality, and Social Behavior*, 2008, pp. 96-103.

[3] F. Tanaka, T. Takahashi, S. Matsuzoe, N. Tazawa, and M. Morita, "Telepresence robot helps children in communicating with teachers who speak a different language," presented at the Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction, Bielefeld, Germany, 2014.

[4] A. Kristoffersson, S. Coradeschi, and A. Loutfi, "A review of mobile robotic telepresence," *Advances in Human-Computer Interaction,* vol. 2013, p. 3, 2013.

[5] V. Harutyunyan, V. Manohar, I. Gezehei, and J. W. Crandall, "Cognitive Telepresence in Human-Robot Interactions," *Journal of Human-Robot Interaction,* vol. 1, pp. 158-182, 2012.

[6] P. Rau, Y. Li, and D. Li, "Effects of communication style and culture on ability to accept recommendations from robots," *Computers in Human Behavior,* vol. 25, pp. 587-595, 2009.

[7] M. Fukushima, R. Fujita, M. Kurihara, T. Suzuki, K. Yamazaki, A. Yamazaki, K. Ikeda, Y. Kuno, Y. Kobayashi, and T. Ohyama, "Question strategy and interculturality in human-robot interaction," in *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, 2013, pp. 125-126.

[8] C. Bartneck, T. Nomura, T. Kanda, T. Suzuki, and K. Kennsuke, "A cross-cultural study on attitudes towards robots," in *HCI International*, 2005.

[9] F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "Interacting with a human or a humanoid robot?," in *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, 2007, pp. 2685-2691.

[10] P. Schermerhorn, M. Scheutz, and C. R. Crowell, "Robot social presence and gender: Do females view robots differently than males?," in *Proceedings of the 3rd ACM/IEEE*

international conference on Human robot interaction, 2008, pp. 263-270.

[11] A. Kristoffersson, K. S. Eklundh, and A. Loutfi, "Towards measurement of interaction quality in social robotic telepresence," in *Proceedings of the Ro-Man Workshop on Social Robotic Telepresence*, 2012, pp. 24-31.

[12] A. Weiss, B. van Dijk, and V. Evers, "Knowing me knowing you: Exploring effects of culture and context on perception of robot personality," in *Proceedings of the 4th international conference on Intercultural Collaboration*, 2012, pp. 133-136.

[13] C. Becker-Asano and H. Ishiguro, "Intercultural differences in decoding facial expressions of the android robot Geminoid F," *Journal of Artificial Intelligence and Soft Computing Research*, p. 215, 2011.

[14] G. Trovato, T. Kishi, N. Endo, M. Zecca, K. Hashimoto, and A. Takanishi, "Cross-Cultural Perspectives on Emotion Expressive Humanoid Robotic Head: Recognition of Facial Expressions and Symbols," *International Journal of Social Robotics*, vol. 5, pp. 515-527, 2013 2013.

[15] C. L. Sidner, C. Lee, L.-P. Morency, and C. Forlines, "The effect of head-nod recognition in human-robot conversation," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 2006, pp. 290-296.

[16] A. Kleinsmith, P. R. De Silva, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture," *Interacting with Computers*, vol. 18, pp. 1371-1389, 2006.

[17] G. Gibert, M. Petit, F. Lance, G. Pointeau, and P. F. Dominey, "What makes human so different? Analysis of human-humanoid robot interaction with a super Wizard of Oz platform," in *International Conference on Intelligent Robots and Systems*, Tokyo, Japan, 2013.

[18] G. Gibert, F. Lance, M. Petit, G. Pointeau, and P. F. Dominey, "Damping robot's head movements affects human-robot interaction," presented at the Human-Robot Interaction (HRI), Bielefeld, Germany, 2014.

[19] S. M. Boker, J. F. Cohn, B. J. Theobald, I. Matthews, T. R. Brick, and J. R. Spies, "Effects of damping head movement and facial expression in dyadic conversation using real-time facial expression tracking and synthesized avatars," *Philosophical Transactions of the Royal Society B-Biological Sciences*, vol. 364, pp. 3485-3495, Dec 12 2009.

[20] A. Kiselev and A. Loutfi, "Using a mental workload index as a measure of usability of a user interface for social robotic telepresence," in *Workshop in Social Robotics Telepresence*, 2012.

[21] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 617-624.

[22] J. C. Nunnally, *Psychometric theory*: McGraw-Hill, 1978.

[23] D. Li, P. P. Rau, and Y. Li, "A cross-cultural study: effect of robot appearance and task," *International Journal of Social Robotics*, vol. 2, pp. 175-186, 2010.

[24] F. Eyssel, D. Kuchenbrandt, S. Bobinger, L. d. Ruiter, and F. Hegel, "'If you sound like me, you must be more human': on the interplay of robot and user features on human-robot acceptance and anthropomorphism," presented at the Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, Boston, Massachusetts, USA, 2012.

[25] C. R. Crowell, M. Scheutz, P. Schermerhorn, and M. Villano, "Gendered voice and robot entities: perceptions and reactions of male and female subjects," presented at the Proceedings of the 2009 IEEE/RSJ international conference on Intelligent robots and systems, St. Louis, MO, USA, 2009.

Virtual conversational agents and social robots: converging challenges

Gérard Bailly[1], Magalie Ochs[2], Alexandre Pauchet[3], Humbert Fiorino[4]
[1]GIPSA-Lab, Grenoble
[3]LTCI, Paris
[4]LITIS, Rouen
[5]LIG, Grenoble

gerard.bailly@gipsa-lab.grenoble-inp.fr, magalie.ochs@telecom-paristech.fr, pauchet@insa-rouen.fr, humbert.fiorino@imag.fr

Résumé
Ce papier identifie les thématiques de recherches à la croisée des défis des robots sociaux et des agents conversationnels animés, engagés dans des interactions situées avec des agents humains ou artificiels.

Mots Clefs
Agents conversationnels animés ; robotique sociale ; interaction homme-machine ; interaction située.

Abstract
This paper deals with converging challenges faced by research communities that have largely evolved in parallel, namely the domain of virtual conversational agents and social robotics.

Keywords
Embodied conversational agents; social robots; human-computer interaction; situated interaction.

Introduction

An embodied agent is an artificial agent that interacts with the physical environment through a physical body within that environment. By extension, virtual metaphors of the physical world have enriched the concept of embodiment with graphical representations of real or virtual human agents and environments which users can perceive and with which they can interact.

One of the main challenges of social robots and virtual agents is to exhibit "intelligent" behaviors, notably the ability to interact in a comprehensive way with humans, alter egos and the environment as perceived by human participants or external observers. Interaction with human beings is particularly challenging because "natural" interactional patterns – including conversational skills – are not only complex, adaptive and highly context-sensitive but also quite lawful, shaped by biological, linguistic, social, emotional and cultural backgrounds.

Empowering robotic or virtual agents with the ability to interact autonomously with human beings and the environment is a challenge that builds on common ground and shares common research challenges (empowering agents with cognitive skills such as perception, motor control and planning, memory, evaluation of linguistic, paralinguistic or non linguistic signals, etc.), techniques (such as signal processing or machine learning, etc.) and technological "bolts" (development and evaluation of actuators and sensors, etc.).

We here attempt to sketch an overview of the some of the common problems and challenges that both communities face. We do not cover all topics: learning & development is treated in depth in Oudeyer's talk [1] and by Pietquin et al [2] , language understanding would require a lengthy section, etc. The objective of this paper is to trigger discussions between two communities. Extensive reviews already published are referenced in each section when necessary. The first section is dedicated to the human model: how biological, linguistic, social, emotional and cultural skills that humans so naturally exhibit can be used and scaled to build convincing artificial agents and what lessons can be drawn from recent experimental paradigms such as the beaming of robotic or virtual avatars by human pilots. The second section deals with artificial cognition and situated interaction: how far have we gone in building artificial theory-of-mind models and how agents can handle and learn from situated interaction? The third section is dedicated to the ultimate dimensions of human interaction: our ability of adapting our behavior according to our conversational partners, their emotional states and the social context of the interaction. The fourth section is dedicated to the current technical challenges and emerging technologies, notably dealing with the modeling of multimodal behaviors and their on-line exploitation. Finally, we will focus on users' studies dealing acceptance of robots and virtual agents as assistants.

1 The human model

Endowing virtual or robotic avatars with socio-communicative skills and multimodal behavior is a crucial issue for agents engaged into face-to-face conversations or, more generally joint activities with human partners. Capturing, understanding and modeling human verbal, co-verbal and non verbal behaviors are thus important challenges for the development of social agents.

1.1 Characterizing multimodal behavior

The first step towards the modeling of perception-action loops is to collect sensorimotor scores that characterize both the time-varying audiovisual scene that the target human agent explores and the actions he/she performs to get/push information into it. Note that the perceptual and motor parts of the score are mutually dependent since scene perception depends intrinsically on action (e.g. visual perception is paced by endogenous gaze shifts) and action is motivated by perception (e.g. exogenous gaze shifts partially depends on audiovisual saliency).

The sensorimotor scores can be supplied by three types of data: intrusive or non intrusive motion capture,

manual annotations and behavioral rules. Intrusive motion capture devices include tracking of facial or body degrees of freedom via passive and active markers attached to body joints as well as head-mounted eye trackers. Non intrusive motion capture devices concern recordings of physiological signals (incl. speech), video-based motion analysis such as model-based facial/body analysis with (e.g. Kinect) or without depth information [3] or remote eye trackers.

Such raw information is often supplemented by manual annotation. This additional indexing is necessary for semi-automatically trimming multimodal scores with intermediate discrete representations (e.g. regions of interest, gestures, etc), providing functional labels [4], or evaluating system performance. Several annotation tools have been proposed: Praat [5] and Transcriber [6] are popular tools for speech annotation. Multimodal tools include Anvil [7], Elan [8], MMAX, Dialogue Tool, ILSP, NITE Workbench, DAT, etc. [see a comparative evaluation in 9]

More explicit knowledge can be found in the psychophysical/psychological literature. Work conducted by Kendon et al [10, 11] on gestures, Kita et al [12] on pointing, etc. are illustrative examples of socially conditioned behavioral rules often implemented in current interactive systems.

These rich sources of information about human behaviors may be exploited for endowing autonomous agents with context-sensitive social behaviors. This exploitation is however not straightforward for at least two reasons: 1) human behaviors should be down-scaled to the perceptuo-motor affordances of the target agent. Agility, processing abilities as well as cognitive resources of the agent have no possible comparison with those of the human teacher; 2) the observed behaviors of human partners in face of a human agent vs. an artificial agent are also very different. It is thus difficult to rely on observed human behavior during human training in terms of both content and form.

Recent works have explored the possibility to teleoperate agents by human pilots to artificially endow them with social skills. This could be achieved both with virtual clones [13] and robots [14-16]. Note however that more immersive teleoperation – known as beaming – can be achieved with robotic incarnations.

1.2 Modeling human dialog

The simplest approach is the finite-state approach [for instance see 17] that represents the structure of the dialogue as a finite-state automaton where each utterance leads to a new state. This approach describes the structure of the dialogue but do not explain it. In practice, this approach is limited to system-directed dialogues.

The frame-based approach represents the dialogue as a process of filling in a frame (also called form) which contains a series of slots [18]. Slots usually correspond to information that the system needs to acquire from the user. It is less rigid than the finite-state approach. Indeed, the dialogue manager includes a control algorithm which determines the response of the system. For instance, the user can fill several slots in one utterance unlike the finite-state approach.

The plan-based approach [19] comes from classic AI. It combines planning techniques such as plan recognition with ideas from the speech act theory [20]. An example of implementation is TRAINS [21]. This approach is rather complex from a computational perspective, and requires advanced NLU components in order to infer the speaker's intentions.

The Information State Update (ISU) framework [22] proposed by the TRINDI project, implements different kinds of dialogue management models. The central component of this approach is called the Information State (IS). It is a formal representation of the common ground between the dialogue participants as well as a structure to deal with agent reasoning. Dialogue act triggers update the IS. GoDIS is an example of system based on this approach [23].

The logic-based approach represents the dialogue and its context in some logical formalism and takes advantage of mechanisms such as inference [see 24, 25]. Most of the logic based approach works are only on a theoretical level.

Dialogue management remains a major deadlock in ECAs [26]. Most of the existing ECAs only integrates basic dialogue management processes, such as a keyword spotter within a finite-state approach or a frame- based approach (for instance, see the SEMAINE project [27]. It is mainly due to the complexity of all the components that compose a dialogue system, the addition of fuzziness along the processing flow and the multidimentionality and multimodality of dialogues.

2 Artificial cognition & situated interaction

Endowing agents with cognitive models is of major importance to reason on the verbal and non-verbal behavior of oneself and others [28]. Several *theory of mind* (ToM) models [29] are been proposed and embodied in various agents, robots [30] as well as virtual characters [31]. ToM is the ability of agents to attribute mental states (beliefs, intents, desires, etc.) to oneself and others and to understand that others have ToM that are different from one's own. We estimate the other's mental states via their verbal and non-verbal behavior: most ToM models [29, 32] proposed that ToM builds on basic processing modules such as face detection, eye direction detection (EDD) or general imitation/simulation capabilities, notably of hand gestures [33]. Evolutionary psychology [34] in fact claims that ToM development is *innately constrained and programmed* in the same way as the development of language ability or face recognition. More recently, embodied cognition [35, 36] tends to put forward the sensori-motor expertise as the basis of the understanding of other minds. Whatever the theoretical claims, ToM is basically recruited to decode agent's intentions: Gallagher et al [37] showed that viewers of cartoons actually monitor intentions of virtual agents.

We have more and more evidence that higher mental processes are grounded in early experience of the physical world via active perception [38]. Robots are surely better equipped to probe the environment and ground their internal representations of agents and

objects via sensorimotor experience. Recent technological advances have been performed in machine learning so that to enable incremental learning via intelligent selection of sensorimotor experience [39].

3 Emotional and social skills

Social robots and virtual agents can be used for similar application (coaching, tutor, etc). They can elicit similar social and emotional states to human partners, e.g. empathy with Kismet [40] and FearNot! [41]. They both require emotional and social skills. Verbal and non-verbal behaviors displayed by virtual agent or social robot facilitate communication (understanding, believability, etc.). Understanding and displaying emotions or some internal psychological states (such as thinking, doubting or being surprised) are important skills during an human-machine interaction. The ECA community has particularly worked on the expression of multimodal behavior: expression of psychological states, coverbal behaviors and its synchronization with speech [42] as well as the display of emotions and links with personality in intelligent environments [43, 44].

3.1 Recognition and display of socio-emotional attitudes

To endow virtual agents or robots with an illusion of life, one of the key elements is their capacity to express socio-emotional attitudes, including emotions but also interpersonal attitudes such as dominance or friendliness.

In the domain of virtual agents, several models have been developed to give the capacity to agents to display emotions. Most of them proposes a repertoire of facial expressions designed based on empirical and theoretical studies in psychology, that have highlighted the morphological and dynamic characteristics of human's emotional facial expressions. In particular, the models are mainly based on the categorical approach proposed by Ekman and Friesen [45]. This approach is based on the hypothesis that humans categorize facial expressions of emotions into a number of categories similar across cultures: happy, fear, anger, surprise, disgust, and sadness (also known as the "big six" basic emotions). The Moving Pictures Experts Group MPEG-4 standards support facial animation by providing Facial Animation Parameters (FAPs) as well as a description of the expression of these six basic emotions [46]. Note however that other taxonomies have been proposed that promote a larger set of socio-emotional attitudes (Baron-Cohen et al [47] cluster more than 400 facial expression in 23 groups) which are connected to the evaluation and display of mental states (see section 2).

Whereas the muscles of the face of a virtual agent can be easily manipulated, these computational models, based on human findings, cannot systematically be applied to robot. Some robots, as for instance Kismet [40], have been constructed specifically to enable the expression of facial expression with a particular focus on the elements of the face conveying emotions (eyebrows, mouse, etc.). However, on some robot, as for instance Nao, the face cannot be controlled. In this case, other elements away from human models can be explored to convey emotions, by using for instance the color of the eyes or the skin.

To gather more subtle, multimodal and natural expressions, some computational models are based on the analysis of annotated corpus to identify the characteristics of the expressions of emotions [48, 49] but also of social attitudes [50, 51]. The corpus can correspond to acted or natural situations in which humans expressed some socio-emotional states, ideally with a motion capture system to automatically compute the socio-emotional characteristics of the facial and corporal movements. Another method consists in collecting a corpus of virtual agents expressions directly created by users [52]. Such an approach intrinsically takes into account the expressive capabilities of the virtual agent and collects a large amount of data based on the users' perception of the virtual agent. However, this method is not well adapted to robots that cannot be easily manipulated.

To summarize, while the computational model of socio-emotional expressions of virtual agents can be largely inspired from the human findings, the specific expressive capabilities of robots entails to explore novel methods to identify how robots may convey socio-emotional attitudes. The beaming, as described in the previous section (§ 1.1) could be this novel method.

3.2 Alignment & social engagement

It is well-known that people in interaction mutually adapt their behaviors. This accommodation occurs via multiple sensory-motor loops operating at various levels of the interaction and this closed-loop process in turn induces modifications in all levels of representation, from social and psychological evaluation to low-level gestural behaviors such as gaze, respiratory patterns, or speech.

The monitoring of space and distance [53] is also very important for the regulation of face-to-face social interaction. Several authors have attempted to model the personal space of virtual agents [54] and robots [55]. Brainbridge et al [56] conducted an original work comparing virtual vs. physical presence: subjects collaborated on simple book-moving tasks with a humanoid robot that was either physically present or displayed via a live video feed. This combination of interactive behavior, and post-interaction, self-reported perception, indicates that participants afford greater trust to the physically present than to the video-displayed robot, making participants more willing to follow through with an unusual request from the robot.

Mutual adaptation implies also temporal coordination [see 57 for a review] of individual's behaviors during social interactions. Benus [58] has notably shown that people sync at turn-taking in collaborative dialogues.

Several research works have shown that people are sensitive to accommodation patterns in HCI [59-62]. Van Vugt et al [44] have notably shown that users prefer to interact with ECA that have facial features similar to theirs. Moreover, adaptive behavior of ECA increases familiarity [63]. Researchers also explore architectures endowing robots with multi-level context- and user-sensitive adaptive behaviors [64, 65].

The detection, generation and monitoring of engagement is a key issue in both robotics/ECA.

4 Emerging technologies

The conception of interactive systems able to deal with a large multimodal observation and state spaces both for analysis and generation has triggered the development of several key technologies.

4.1 Platforms for Interactive Systems

An Interactive System contains multiple components. The components processing the user's inputs can be formalized as Knowledge Extractors, the Dialogue Manager as an Interaction Manager and the output components are Behavior Generators associated to Players. The component that interprets the behavior, the Player, can either be a simple Speech Interface, an Embodied Conversational Agent (ECA) or a robot. We restrict here the presentation of projects that propose a system with a player that is general enough, i.e. not strongly linked with the interpreter.

The first example is the MULTIPLATFORM project [66] which served as a component integration platform for two well known projects: Verbmobil [67] and Smartkom [68].

Another generation of systems is based on the Psyclone middleware platform, which implements a classic blackboard communication protocol. Two platforms use the Psyclone protocol: Mirage [69] and GECA [70]. The system proposed by Cavazza et al [71] is not only an ECA but also a companion, engaged into a long term interaction process to forge an empathic relation with its user. The system uses several proprietary platforms, designed by industrial partners: a middleware platform, Inamode, developed by Telefonica I+D; an Automatic Speech Recognition (ASR) and a Text To Speech (TTS) engine, developed by Loquendo; and a Virtual Character, developed by As An Angel.

Semaine [27] is a Sensitive Artificial Listener (SAL), built around the idea of emotional interaction. The project focuses on a Virtual Character that perceives human emotions through a multi-modal set-up and answers accordingly. Several virtual characters with different personalities are proposed, each having a different reactive model to the perceived emotion. The affect detection part is a fusion of low level speech features extracted using OpenSMILE [72] and face gestures classified using iBug [73]. The behavior of the agent is managed by two components: a Text-to-Speech synthesizer: MaryTTS [74] and a gesture synthesis component, which converts the data into Greta BML code [75].

Virtual Human Toolkit (VHToolkit) [76] is a generic platform designed to support ECA systems, developed around a component-based design methodology. It has been used successfully in many applications varying from e-learning to military training. It provides a collection of components for all the major tasks of an interactive system: speech recognition, text-to-speech, dialogue management [using the NPCEditor component proposed by 77] non-verbal body movement generator [78] and an uniform perception layer, formalized as PML [79]. The project uses the SmartBody Embodiment (Shapiro, 2011) as a visual BML interpreter for verbal and non verbal behavior.

Finally, AgentSlang [80] consists in a series of original components integrated with several existing algorithms, to provide a development environment for interactive systems. The platform is efficient and provides action execution feedback and data type consistency check.

4.2 Learning behavioral models

Recent approaches aim to learn dialogue policies with machine learning techniques such as reinforcement learning [81]. In this approach, the dialogue management is seen as a decision problem and the dialogue system is modeled as a Markov Decision Process (MDP). Young et al [82] have notably proposed to augment a (Partially Observable MDP) POMDP-based spoken dialogue system with perception and action cues that are directly observed in the speech signals. Speech recognition and synthesis as well as dialog management are all embedded into a large statistical framework. Expectation-Maximization (EM) training and statistical inference are thus used to adjust the model parameters and monitor spoken interaction.

Several recent works have proposed to map perceptual cues to multimodal action given the on-line decoding of underlying joint sensori-motor states. As an example, Otsuka et al [83-85] proposed to use Dynamic Bayesian Networks (DBN) to estimate head and gaze directions and underlying interaction *regimes* given speech activities in multiparty conversations. Zhang et al [86] used a two-layered Hidden Markov Model (HMM) to model individual and group actions in meetings. Machine-learning techniques and statistical models are now mature to address both estimation and inference in high-dimensional observation and state spaces. Perception-action mappings that are functionally aware of the underlying socio-communicative states generally perform better than pure signal-based classifiers such as SVM or decision trees [87].

4.3 Incremental technologies

When used by interactive systems, models that interpret and generate verbal and non-verbal behaviors should be endowed with incremental processing capabilities [see 88 for a review]: human interlocutors in dialogue typically gesture and produce speech in a piecemeal fashion and on-line as the dialogue progresses. When starting their dialog turns, participants typically do not have a complete plan of how to say something or even what to say. They manage to rapidly integrate information from different multimodal sources in parallel and simultaneously plan and realize new behavioral contributions. Most dialog systems analyze and process interaction by speech acts or talk-spurts and have difficulty in processing other's behaviors and generate appropriate feedbacks on the fly. This often results in large response delays [89] that impair interaction and conversation.

Incremental technologies often confine the processing time span to a limited horizon of few frames ahead [90] or use predictive coding or parsing to provide classical technologies with context.

5 Coaching by virtual and robotic agents versus human intervention

Several research works have shown the benefits of interactions with physically present robot compared to virtual agent [91] or robot represented on a screen [92]: physical and social presence enables more natural interaction [see detailed state of art in 91].

Coaching by virtual and robotic agents is equally or even more effective notably with autistic subjects [93, 94]. Robots can be used to understand pathologies, for instance, learning social signatures during robot–children imitation task [95]. Coaching by virtual and robotic agents are often equally effective and sometimes compete with human coaching, notably with autistic subjects [93, 94].

More generally, broadening access to healthcare and improving prevention and patient outcomes are the main societal drivers for healthcare robotics. As the world's population is growing older, new challenges are arising. An increasing number of people needs healthcare while the number of people providing that care (doctors, nurses, physical therapists) is dropping. The value of healthcare robotics in increasing life-long independence becomes a key issue: care-taking for elderly people [96], promoting ageing in place [97], motivating cognitive and physical exercise [98] delaying the onset of dementia [99], and to mitigate isolation and depression. Socially Assistive Robotics (SAR) assists users through social rather than physical interactions. The robot's physical embodiment is at the heart of SAR's assistive effectiveness, as it leverages the inherently human tendency to engage with lifelike social behavior. An effective socially assistive robot must understand and interact with its environment, exhibit social behavior, focus its attention and communication on the user, sustain engagement with the user, as well as achieve specific assistive goals. Socially assistive robots are promising as diagnosis and therapeutic tool for children (autism, eye-tracking for ASD diagnosis) [100], the elderly [98, 99, 101], stroke patient [102, 103] and other special-needs populations requiring personalized care.

Some research results show that people are more likely both to fulfill an unusual request and to afford greater personal space to a robot when it was physically present, than when it was shown on live video [92]. Moreover, a robot to conduct interviews [104] or give instructions [105] can be as good as a human. Such results raise new research challenges for SAR: How do patients' non-verbal and verbal behaviors as well as their opinions differ in their interactions with a robot vs. a clinician? Can robots be as efficient as clinicians in the long term?

6 Migration

Story characters that migrate between virtual and real worlds have elicited interest in literature and cinema, among other forms of art: Alice going Through the Looking-Glass, Neo diving in and out of The Matrix, etc. are such examples. The concept of cross-embodiment has been named *agent migration* by Gomes et al [106] as: the "Process by which an agent moves between embodiments, being active in only one at a time". They conducted a user study with 51 elementary school children that interacted with an artificial pet that can migrate from two different embodiments, namely a robot pet and a virtual pet on a mobile phone. 43.3% of the children understand that both embodiments were actually the same entity.

Koay et al [107] investigated the use of three different visual cues (moving bars, moving face and flashing lights) to support the user's belief that they are still interacting with the same agent migrating between different robotic embodiments. The 21 primary school male participants support moving face as a strong cue for migration and that they had some expectations concerning the duration of the migration process.

Interlocutors of artificial agents have a mental model for the process of migrating personalities between different physical embodiments. The general challenge of agent migration is thus to preserve the agent personality across migration [108] despite the change of sensory-motor capabilities.

Conclusions

If virtual characters and humanoid robots exhibit large differences in the perceptual and motor cues they can be used to decode and encode socio-communicative intentions, they all face the challenge of monitoring short and long-term interactions. The internet of things will certainly be architected with intelligent agents. No doubt that human beings will need to interact with a limited number of agents that will be endowed with social skills. Depending on availability of resources, interaction needs and purposes, these intelligent agents will have various embodiments, from a plain humanoid robot to a disembodied voice. One challenge of the conception of intelligent agents is to mediate dialog whatever the available resources while adapting to the user, the situation and these available resources. Maintaining a form of personality is surely the perquisite of long-term acquaintance and mutual trust between humans and agents.

References

[1] Oudeyer, P.Y. *Curiosity-driven learning and development: How robots can help us understand humans*. in *Workshop Affect, Compagnon Artificiel, Interaction (WACAI)*. 2014. Rouen - France.

[2] Pietquin, O. and M. Lopes. *Machine learning for interactive systems: challenges and future trends*. in *Workshop Affect, Compagnon Artificiel, Interaction (WACAI)*. 2014. Rouen - France.

[3] Boker, S.M. and J.F. Cohn, *Real-time dissociation of facial appearance and dynamics during natural conversation*, in *Dynamic faces: Insights from experiments and computation*, C. Curio, H.H.B. lthoff, and M.A. Giese, Editors. 2011: Cambridge, MA. p. 239-254.

[4] Heylen, D., et al. *The next step towards a functional markup language*. in *Intelligent Virtual Agents (IVA)*. 2008. Tokyo. p. 37-44.

[5] Boersma, P. and D. Weenink, *Praat, a System for doing Phonetics by Computer, version 3.4*, in *Institute of Phonetic Sciences of the University of Amsterdam, Report 132. 182 pages*. 1996.

[6] Barras, C., et al., *Transcriber: development and use of a tool for assisting speech corpora production.* Speech Communication - special issue on Speech Annotation and Corpus Tools, 2001. **33**(1-2): p. 5–22.

[7] Kipp, M. *Anvil - a generic annotation tool for multimodal dialogue*. in *European Conference on Speech Communication and Technology (Eurospeech)*. 2001. Aalborg. p. 1367-1370.

[8] Hellwig, B. and D. Uytvanck, *EUDICO Linguistic Annotator (ELAN) Version 2.0.2 manual*. 2004, Max Planck Institute for Psycholinguistics: Nijmegen - NL.

[9] Garg, S., et al. *Evaluation of transcription and annotation tools for a multi-modal, multi-party dialogue corpus*. in *International Conference on Language Ressources and Evaluation (LREC)*. 2004. Lisbon.

[10] Kendon, A., *Gesture: Visible action as utterance*. 2004, Cambridge: Cambridge University Presspages.

[11] Kendon, A., *Does gesture communicate? A Review.* Research on Language and Social Interaction, 1994. **2**(3): p. 175-200.

[12] Kita, S., *Pointing: Where Language, Culture, and Cognition Meet*. 2003, Mahwah, NJ: Lawrence Erlbaum Associates. 339 pages.

[13] Boker, S.M., et al., *Something in the way we move: Motion, not perceived sex, influences nods in conversation.* Journal of Experimental Psychology: Human Perception and Performance, 2011. **37**(3): p. 874-891.

[14] Normand, J.-M., et al., *Beaming into the rat world: enabling real-time interaction between rat and human each at their own scale.* PLoS ONE, 2012. **7**(10): p. e48331.

[15] Steed, A., et al., *Beaming: an asymmetric telepresence system.* IEEE Computer Graphics and Applications, 2012. **32**(6): p. 10-17.

[16] Nishio, S., et al., *Body ownership transfer to teleoperated android robot*, in *Social Robotics*, S. Ge, et al., Editors. 2012, Springer Berlin Heidelberg. p. 398-407.

[17] McTear, M., *Spoken dialogue technology: toward the conversational user interface*. 2004, New York: Springer-Verlag 374 pages.

[18] Aust, H., et al., *The Philips automatic train timetable information system.* Speech Communication - special issue on Silent Speech Interfaces, 1995. **17**(3-4): p. 249-262.

[19] Allen, J. and C. Perrault, *Analyzing intention in utterances.* Artificial Intelligence, 1980. **15**(3): p. 143-178.

[20] Searle, J.R., *Speech Acts: An Essay in the Philosophy of Language*. 1969, Cambridge, UK: Cambridge University. 203 pages.

[21] Allen, J., et al., *Dialogue systems: From theory to practice in TRAINS-96.* Handbook of Natural Language Processing, 2000: p. 347-376.

[22] Larsson, S. and D.R. Traum, *Information state and dialogue management in the TRINDI dialogue move engine toolkit.* Natural language engineering, 2000. **6**(3&4): p. 323-340.

[23] Larsson, S., *Issue-based dialogue management*, PhD Thesis in *Department of Linguistics*. 2002, Göteborg University. 312 pages.

[24] Maudet, N., *Modéliser les conventions des interactions langagières: la contribution des jeux de dialogue*, in *IRIT*. 2001, Université Paul Sabatier: Toulouse, France. 195 pages.

[25] Hulstijn, J. and N. Maudet, *Uptake and joint action.* Journal of Cognitive Systems Research: Special issue on Cognition and Collective Intentionality, 2006. **7**(2&3): p. 175-191.

[26] Swartout, W.R., et al., *Toward virtual humans.* AI Magazine, 2006. **27**(2): p. 96-108.

[27] Schröder, M. (2010) *The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems*. Advances in Human-Machine Interaction **2010**, pages DOI: 10.1155/2010/319406.

[28] Vinayagamoorthy, V., A. Steed, and M. Slater. *Building characters: lessons drawn from virtual environments. toward social mechanisms of android science*. in *CogSci Workshop*. 2005. Stresa, Italy. p. 119-126.

[29] Baron-Cohen, S., A. Leslie, and U. Frith, *Does the autistic child have a "theory of mind"?* Cognition, 1985. **21**: p. 37-46.

[30] Scassellati, B., *Foundations for a theory of mind for a humanoid robot*, in *Department of Computer Science and Electrical Engineering*. 2001, MIT: Boston - MA. 174 pages.

[31] Peters, C., *A perceptually-based theory of mind model for agent interaction initiation.* International Journal of Humanoid Robotics, 2006. **3**(3): p. 321 - 340.

[32] Leslie, A.M., *ToMM, ToBY, and Agency: Core architecture and domain specificity*, in *Mapping the Mind: Domain specificity in cognition and culture*, L.A. Hirschfeld and S.A. Gelman, Editors. 1994, Cambridge University Press: Cambridge. p. 119–148.

[33] Rizzolatti, G., L. Fogassi, and V. Gallese, *Mirror neurons: Intentionality detectors?* International Journal of Psychology, 2000. **35**: p. 205-205.

[34] Gerrans, P., *The theory of mind module in evolutionary psychology.* Biology and Philosophy, 2002. **17**: p. 305-321.

[35] Adams, F., *Embodied cognition.* Phenomenology and the Cognitive Sciences, 2010. **9**(4): p. 619-628.

[36] Varela, F.J., E. Rosch, and E. Thompson, *The Embodied Mind: Cognitive Science and Human Experience*. 1992, Boston, MA: MIT Press. 299 pages.

[37] Gallagher, H.L., et al., *Reading the mind in cartoons and stories: An fMRI study of 'theory of mind' in verbal and nonverbal tasks.* Neuropsychologia, 2000. **38**: p. 11-21.

[38] Williams, L.E., J.Y. Huang, and J.A. Bargh, *The scaffolded mind: Higher mental processes are*

grounded in early experience of the physical world. European Journal of Social Psychology, 2009. **39**: p. 1257-1267.

[39] Baranes, A. and P.-Y. Oudeyer, *Active learning of inverse models with intrinsically motivated goal exploration in robots*. Robotics and Autonomous Systems, 2013. **61**(1): p. 49-73.

[40] Breazeal, C., *Towards sociable robots*, in *Robotics and Autonomous Systems*. 2003. p. 167-175.

[41] Aylett, R., et al., *Unscripted narrative for affectively driven characters*. IEEE Computer Graphics and Applications, 2006. **26**(3): p. 42-52.

[42] Kopp, S., *Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors*. Speech Communication, 2010. **52**(6): p. 587-597.

[43] Pesty, S. and D. Duhaut. *Acceptability in interaction - from robots to embodied conversational agents*. in *Computer graphics theory and applications*. 2011. Algarve : Portugal.

[44] Van Vugt, H.C., et al., *Effects of facial similarity on user responses to embodied agents*. ACM Transaction on Human-Computer Interaction, 2010. **17**(2): p. 1-27.

[45] Ekman, P. and W.V. Friesen, *Unmasking the Face*. 1975, Palo Alto, California.: Consulting Psychologists Presspages.

[46] Ostermann, J., *Face animation in MPEG-4*, in *MPEG-4 Facial Animation - The Standard Implementation and Applications*, I.S. Pandzic and R. Forchheimer, Editors. 2002, Wiley: Oxford, UK. p. 17-55.

[47] Baron-Cohen, S., et al., *Mind Reading: The Interactive Guide to Emotions* 2004, University of Cambridge: UK.

[48] Niewiadomski, R. and C. Pelachaud. *Towards multimodal expression of laughter*. in *International Conference on Intelligent Virtual Agents (IVA)*. 2012. Santa Cruz, CA. p. 231-244.

[49] Bailly, G., et al. *Degrees of freedom of facial movements in face-to-face conversational speech*. in *International Workshop on Multimodal Corpora*. 2006. Genoa - Italy. p. 33-36.

[50] Chollet, M., M. Ochs, and C. Pelachaud. *Mining a multimodal corpus for non-verbal signals sequences conveying attitudes*. in *Language Resources and Evaluation Conference (LREC)*. 2014. Reykjavik, Iceland. p. paper 235.

[51] Ravenet, B., M. Ochs, and C. Pelachaud, *From a user-created corpus of virtual agent's non-verbal behaviour to a computational model of interpersonal attitudes*, in *International Conference on Intelligent Virtual Agent (IVA)*. 2013: Edinburgh.

[52] Ochs, M., B. Ravenet, and C. Pelachaud. *A crowdsourcing toolbox for a user-perception based design of social virtual actors*. in *Intelligent Virtual Agent Conference (IVA)*. 2013. Edinburgh.

[53] Hall, E.T., *A system for the notation of proxemic behaviour.* American Anthropologist, 1963. **85**: p. 1003–1026.

[54] Amaoka, T., H. Laga, and M. Nakajima. *Modeling the personal space of virtual agents for behavior simulation* in *International Conference on CyberWorlds*. 2009. Bradford, UK. p. 364-370

[55] Mumm, J. and B. Mutlu. *Human-robot proxemics: physical and psychological distancing in human-robot interaction*. in *Human-Robot Interaction (HRI)*. 2011. Lausanne, Switzerland.

[56] Bainbridge, W.A., et al. *The effect of presence on human-robot interaction*. in *IEEE International Symposium on Robot and Human Interactive Communication (RoMAN)*. 2008. Munich. p. 701-706.

[57] Delaherche, E., et al., *Interpersonal synchrony: a survey of evaluation methods across disciplines.* IEEE Trans. on Affective Computing, 2012. **3**(3): p. 349-365.

[58] Benus, S. *Are we 'in sync': Turn-taking in collaborative dialogues*. in *Interspeech*. 2009. Brighton, UK. p. 2167-2170.

[59] Bell, L., J. Gustafson, and M. Heldner. *Prosodic adaptation in human-computer interaction*. in *International Congress of Phonetic Sciences*. 2003. Barcelona. p. 2453-2456.

[60] Suzuki, N. and Y. Katagiri, *Prosodic alignment in human-computer interaction.* Connection Science, 2007. **19**(2): p. 131-141.

[61] Suzuki, N. and Y. Katagiri. *Prosodic synchrony for error management in human-computer interaction*. in *ISCA Workshop on Error Handling in Spoken Dialogue Systems*. 2003. p. 107-111.

[62] Oviatt, S., C. Darves, and R. Coulston, *Toward adaptive conversational interfaces: modeling speech convergence with animated personas.* ACM Transactions on Computer-Human Interaction, 2004. **11**: p. 300-328.

[63] Yaghoubzadeh, R. and S. Kopp. *Creating familiarity through adaptive behavior generation in human/agent interaction*. in *International Conference on Intelligent Virtual Agents (IVA)*. 2011. Reykjavík, Iceland. p. 195-201.

[64] Baxter, P.E., J. de Greeff, and T. Belpaeme. *Cognitive architecture for human–robot interaction: Towards behavioural alignment*. in *International Conference on Biologically Inspired Cognitive Architectures (BICA)*. 2013. Kiev, Ukraine. p. 30–39.

[65] Clair, A.S. and M.J. Matarić. *Studying coordinating behavior in human-robot task collaborations using the PR2*. in *PR2 workshop at Intelligent Robots and Systems (IROS)*. 2011. san Franscisco, CA.

[66] Herzog, G., et al., *Large-scale software integration for spoken language and multimodal dialog systems.* Natural Language Engineering, 2004. **10**(3-4): p. 283-305.

[67] Wahlster, W., *Verbmobil: Foundations of Speech-to-Speech Translation*. 2000 Berlin: Springer. 679 pages.

[68] Wahlster, W., *SmartKom: Foundations of Multimodal Dialogue Systems* 2006, New York: Springer Verlag. 643 pages.

[69] Thörisson, K.R., et al., *Constructionist design methodology for interactive intelligences.* AI Magazine, 2004. **25**(4): p. 77.

[70] Huang, H.-H., et al., *Integrating embodied conversational agent components with a generic framework.* Multiagent and Grid Systems, 2008. **4**(4): p. 371-386.

[71] Cavazza, M., R.S. de la Camara, and M. Turunen. *How was your day? a companion ECA*. in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 2010. Toronto. p. 1629-1630.

[72] Eyben, F., et al., *Opensmile: the munich versatile and fast open-source audio feature extractor*, in *Proceedings of the international conference on Multimedia*. 2010, ACM: Firenze, Italy. p. 1459-1462.

[73] Soleymani, M., M. Pantic, and T. Pun, *Multimodal emotion recognition in response to videos.* IEEE Transactions on Affective Computing, 2012. **3**(2): p. 211-223.

[74] Pammi, S.C., M. Charfuelan, and M. Schröder. *Multilingual voice creation toolkit for the MARY TTS Platform*. in *LREC*. 2010. Valleta, Malta.

[75] Poggi, I., et al., *GRETA. A believable embodied conversational agent*, in *Multimodal intelligent information presentation*, O. Stock and M. Zancarano, Editors. 2005, Kluwer: Dordrecht. p. 3-26.

[76] Chan, A.D.C., et al., *Myo-electric signals to augment speech recognition.* Medical & Biological Engineering & Computing, 2001. **39**: p. 500-504.

[77] Leuski, A. and D. Traum. *NPCEditor: a tool for building question-answering characters*. in *International Conference on Language Resources and Evaluation (LREC)*. 2011. Valletta, Malta.

[78] Lee, J. and S.C. Marsella. *Nonverbal behavior generator for embodied conversational agents*. in *International Conference on Intelligent Virtual Agents (IVA)*. 2006. Marina del Rey, CA.

[79] Scherer, S., et al. *Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors*. in *International Conference on Intelligent Virtual Agents (IVA)*. 2012. Santa Cruz, CA.

[80] Serban, O. and A. Pauchet, *AgentSlang: A fast and reliable platform for distributed interactive systems*, in *international conference on Intelligent Computer Communication and Processing (ICCP)*. 2013: Cluj-Napoca, Roumania.

[81] Frampton, M. and O. Lemon, *Recent research advances in reinforcement learning in spoken dialogue systems.* Knowledge Engineering Review, 2009. **24**(4): p. 375-408.

[82] Young, S., et al., *POMDP-based statistical spoken dialogue systems: a review.* Proc IEEE, 2013. **101**(5): p. 1160-1179.

[83] Otsuka, K., Y. Takemae, and J. Yamato. *A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances*. in *International Conference on Multimodal Interfaces (ICMI)*. 2005. Torento, Italy.

[84] Otsuka, K., J. Yamato, and H. Murase. *Conversation scene analysis with dynamic Bayesian network based on visual head tracking*. in *ICMI*. 2006. p. 949-952.

[85] Otsuka, K. *Multimodal Conversation Scene Analysis for Understanding People's Communicative Behaviors in Face-to-Face Meetings*. in *International Conference on Human-Computer Interaction (HCI)*. 2011. Orlando, FL. p. 171-179.

[86] Zhang, D., et al., *Modeling individual and group actions in meetings with layered HMMs.* IEEE Transactions onMultimedia, 2006. **8**(3): p. 509-520.

[87] Mihoub, A., G. Bailly, and C. Wolf. *Social behavior modeling based on Incremental Discrete Hidden Markov Models*. in *Human Behavior Understanding*. 2013. Barcelona, Spain. p. 172-183.

[88] Schlangen, D., et al. *Middleware for incremental processing in conversational agents*. in *Annual Meeting of the Special Interest Group in Discourse and Dialogue (SIGDIAL)*. 2010. Tokyo, Japan. p. 51-54.

[89] Fraser, N.M. and G.N. Gilbert, *Simulating speech systems.* Computer Speech and Language, 1991. **5**(1): p. 81-99.

[90] Bloit, J. and X. Rodet. *Short-time viterbi for online HMM decoding: Evaluation on a real-time phone recognition task*. in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2008. Las Vegas, NE. p. 2121-2124.

[91] Segura, E.M., et al. *How do you like me in this: user embodiment preferences for companion agents*. in *Intelligent Virtual Agents (Lecture Notes in Computer Science n°7502)*. 2012. SantaCruz, CA. p. 112-125

[92] Bainbridge, W., et al., *The benefits of interactions with physically present robots over video-displayed agents.* International Journal of Social Robotics, 2011. **3**(1): p. 41-52.

[93] Huskens, B., et al., *Promoting question-asking in school-aged children with autism spectrum disorders: effectiveness of a robot intervention compared to a human-trainer intervention.* Developmental Neurorehabilitation, 2013. **16**(5): p. 345-356.

[94] Duquette, A., F. Michaud, and H. Mercier, *Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism.* Autonomous Robots, 2008. **24**(2): p. 147-157.

[95] Boucenna, S., et al., *Learning of social signatures through imitation game between a robot and a human partner.* IEEE Transactions on Autonomous and Mental Development 2014.

[96] Broadbent, E., R. Stafford, and B. MacDonald, *Acceptance of healthcare robots for the older population: review and future directions.* International Journal of Social Robotics, 2009. **1**(4): p. 319-330.

[97] Johnson, D.O., et al., *Socially assistive robots: a comprehensive approach to extending independent living.* International Journal of Social Robotics, 2013: p. 195-211.

[98] Fasola, J. and M. Mataric, *A socially assistive robot exercise coach for the elderly.* Journal of Human-Robot Interaction, 2013. **2**(2): p. 3-32.

[99] Tapus, A., C. Tapus, and M.J. Mataric. *The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia*. in *Rehabilitation Robotics (ICORR)*. 2009. Kyoto.

[100] Cabibihan, J.-J., et al., *Why Robots? A survey on the roles and benefits of social robots in the therapy of children with autism.* International Journal of Social Robotics, 2013. **5**(4): p. 593-618.

[101] Heerink, M., et al., *Assessing acceptance of assistive social agent technology by older adults: The Almere Model.* International Journal of Social Robotics, 2010. **2**(4): p. 361-375.

[102] Abdullah, H., et al. (2011) *Results of clinicians using a therapeutic robotic system in an inpatient stroke rehabilitation unit*. Journal of Neuroengineering and Rehabilitation **8**, pages DOI: 10.1186/1743-0003-8-50.

[103] Mazzoleni, S., et al., *Acceptability of robotic technology in neuro-rehabilitation: Preliminary results on chronic stroke patients.* Computer Methods and Programs in Biomedicine, 2014: p. 1-7.

[104] Wood, L.J., et al. (2013) *Robot-mediated interviews - how effective Is a humanoid robot as a tool for interviewing young children?* PLoS ONE, 13 pages DOI: 10.1371/journal.pone.0059448.

[105] Giuliani, M. and A. Knoll, *Using embodied multimodal fusion to perform supportive and instructive robot roles in human-robot interaction.* International Journal of Social Robotics, 2013. **5**: p. 345–356.

[106] Gomes, P.F., et al., *Migration between two embodiments of an artificial pet.* International Journal of Humanoid Robotics, 2014: p. accepted.

[107] Koay, K.L., et al. *A user study on visualization of agent migration between two companion robots*. in *International Conference on Human-Computer Interaction (HCI)*. 2009. San Diego, CA.

[108] Kaplan, F. *Artificial attachment: Will a robot ever pass ainsworth s strange situation test?* in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*. 2001. Tokyo. p. 125-132.

Implantation d'un ACA narrateur (Démonstration)

William Boisseleau[1] Ovidiu Şerban[2] Alexandre Pauchet[1]

[1] LITIS, INSA de Rouen, France : alexandre.pauchet@insa-rouen.fr
[2] ISR Laboratory, University of Reading, United Kingdom : o.serban@reading.ac.uk

Résumé

Cet article décrit l'implantation d'un ACA narrateur sur AgentSlang, une plate-forme orientée composants dédiée à la conception de systèmes interactifs. Cet ACA est conçu pour interagir de manière aussi naturelle que possible avec un enfant, dans un contexte de narration collaborative. La version courante de cet ACA fait suite à une version validée par Magicien D'Oz. Le prototype proposé ici en démonstration intègre une série de composants et logiciels existants ou spécialement conçus pour la tâche de narration.

1 Introduction

La conception de systèmes permettant des interactions homme-machine naturelles est une tâche particulièrement compliquée. Dans ce cadre, le développement d'Agents Conversationnels Animés (ACA, ou ECA - Embodied Conversational Agents - en anglais [8]), imitant les humains, est un axe de recherche prometteur. Les ACA, en tant que personnages animés, tendent à reproduire le comportement humain à différents niveaux de communication : langue naturelle, expressions faciales, gestes et postures, regards, etc. [1]. Cependant, du fait de leur apparence hyper-réaliste, les ACA peuvent induire des attentes particulières, souvent déçues, quant à leurs capacités dialogiques [7]. Les projets de recherche les plus récents ont démontré qu'il était possible d'améliorer la perception qu'ont les utilisateurs des ACA simplement en augmentant leurs capacités d'interaction et leur expressivité [1, 9]. Le projet SEMAINE [10] ou VHTookit [6] donne un aperçu des performances actuelles des ACA.

Une des difficultés dans le développement d'ACA vient de la nécessité d'avoir un environnement de développement dans lequel les interactions sont les plus naturelles possibles, afin de pouvoir tester les modèles et d'instaurer une conception itérative. Une première expérimentation a été réalisée par Magicien d'Oz (MOz), intégrant un modèle de dialogue par automates à états finis. Les différents états étaient alors déclenchés manuellement par une psychologue, en fonction des retours audio et visuel de l'utilisateur. Le présent article présente une première version d'un ACA narrateur intégrant le modèle interactif testé durant cette expérimentation. Cet ACA est déployé dans AgentSlang [5], une plate-forme dédiée à la conception de systèmes interactifs intégrant un système de reconnaissance d'expressions rationnelles supportant la synonymie.

L'objectif de ce projet est de créer un environnement de narration dans lequel un enfant peut interagir de manière naturelle avec un ACA. L'idée principale du prototype décrit ici est de déclencher les actions qui étaient activées manuellement durant le MOz en utilisant des composants AgentSlang [5]. La même histoire est utilisée comme support ("Le ballon perché"). Notre plate-forme doit donc intégrer narration, courtes phrases interactives (réponses aux questions non prévues dans le scénario) et de synchroniser ces différents éléments avec un ensemble d'illustrations. M.A.R.C. a été intégré en tant qu'ACA narrateur [4].

Un ensemble d'expressions rationnelles ont été identifiées durant le MOz afin d'encoder les interventions dialogiques possibles des enfants. Elles permettent soit d'activer les transitions et de passer ainsi d'une étape de la narration à une autre (contiguë ou non), soit d'activer une réponse orale ou multimodale sans changer d'étape. Une absence de réponse peut également être considérée comme une expression valable. De plus, même si l'enfant exécute une action non prévue par le scénario (hors contexte), le système réagit en mentionnant la source d'incompréhension. Le système supporte également des ensembles d'expressions d'entrée liés à des ensembles de réponses, afin d'éviter les effets de répétition. Les motifs d'entrée sont définis à l'aide du langage Syn !bad [5], et fonctionnent uniquement sur du texte saisi ou retranscrit.

2 Composants AgentSlang

L'architecture proposée consiste en un ensemble de composants AgentSlang [5], liés entre eux par canaux de communication. La figure 1 décrit le modèle, pour lequel chaque composant possède un identifiant unique, équivalent à un port TCP dans cette configuration. Bien qu'AgentSlang soit une plate-forme générique pour la conception de systèmes interactifs, dont les ACA, un certain nombre de composants ont dus être construits spécifiquement pour ce projet :

1. L'enfant effectue des actions dialogiques sous une forme textuelle. Le texte est récupéré par le système soit par saisie (composant *Text*), soit par retranscription (composant *SpeechToText*).
2. Le composant *Text* envoie une chaîne de caractères au composant *Senna* [3], qui annote le texte.
3. Le texte annoté est alors traité par le composant *Pattern-*

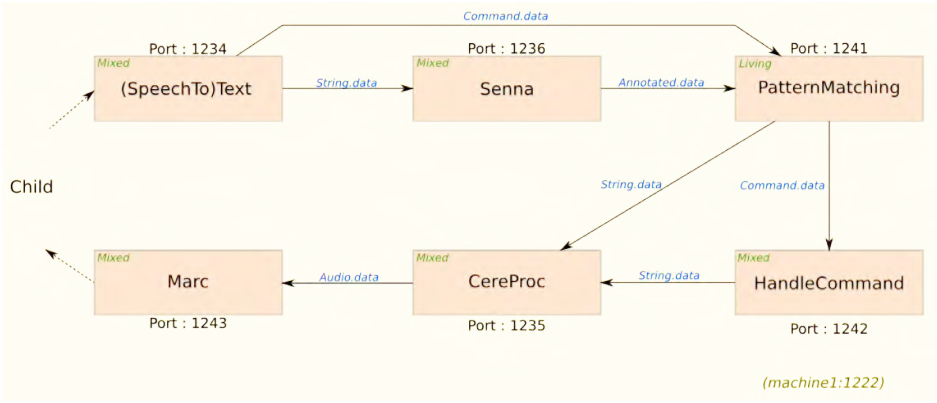


FIGURE 1 – L'architecture d'ACA narrateur proposée

Matching qui vérifie dans un premier temps si il correspond à un motif d'entrée sous la forme d'un automate Syn !Bad [5]. Deux listes de motifs Syn !Bad sont vérifiées successivement : la liste des réponses attendues et celle des réponses hors contexte, activées à n'importe quel moment. Chaque motif est couplé à un ensemble de réponses possibles ainsi qu'une étape à jouer (qui peut être l'étape actuelle). En fonction du motif détecté, une commande est alors envoyée aux composants suivants qui accomplissent les actions graphiques ou interactives sélectionnées. L'envoi se fait sous la forme d'une chaîne de caractères.

4. Le composant *CereProc* [2] transforme un texte produit en fichier audio, qui est finalement lu par le composant *MARC* [4]. Marc est l'ACA utilisé pour jouer des expressions faciales et prononcer des phrases.

Les données décrivant les différentes éléments d'entrée et sortie (étapes du scénario, commandes et éléments hors contexte) sont stockées dans des fichiers XML facilement compréhensibles, éditables et maintenables. Un outil de génération de script a également été développé pour générer rapidement ces fichiers types.

3 Conclusion et perspectives

Dans cet article de démonstration, nous avons présenté rapidement une implémentation d'un ACA narrateur à l'aide d'AgentSlang, une plate-forme de déploiement de systèmes interactifs. Le résultat principal est un premier prototype d'ACA narrateur dont le modèle de dialogue est un simple automate à états fini.

Il existe à l'heure actuelle plusieurs limitations à ce prototype. Tout d'abord, Syn !bad et la plate-forme AgentSlang sont maintenant multilingues, mais ne supportaient que l'anglais au moment du développement de cette démonstration. Deuxièmement, les éléments d'interaction sont à l'heure actuelle encore limités, mais devraient augmenter progressivement du fait de notre conception itérative. Enfin, aucune mémoire n'est pour l'instant intégrée aux modèles, ni en ce qui concerne les éléments hors contexte qui auraient déjà été activés, ni en ce qui concerne les informations éventuellement collectées sur l'enfant (nom, âge, etc.). Ces trois points sont actuellement en développement.

Remerciements

Cette démonstration fait partie du projet NARECA, financé par l'ANR (ANR-13-CORD-0015).

Références

[1] CASSELL, J., BICKMORE, T., CAMPBELL, L., VILHJÁLMSSON, H., AND YAN, H. Embodied conversational agents. MIT Press, 2000, ch. Human conversation as a system framework : designing embodied conversational agents, pp. 29–63.

[2] CEREPROC. Cerevoice sdk. `http://www.cereproc.com/`.

[3] COLLOBERT, R., WESTON, J., BOTTOU, L., KARLEN, M., KAVUKCUOGLU, K., AND KUKSA, P. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research 12* (2011), 2493–2537.

[4] COURGEON, M., MARTIN, J., AND JACQUEMIN, C. Marc : a multimodal affective and reactive character. In *Proceeding of Workshop on Affective Interaction on Natural Environment* (2008).

[5] ŞERBAN, O., AND PAUCHET, A. Agentslang : a new distributed interactive system. current approaches and performance. In *International Conference on Agents and Artificial Intelligence (ICAART)* (2014), p. 8.

[6] HARTHOLT, A., TRAUM, D., MARSELLA, S. C., SHAPIRO, A., STRATOU, G., LEUSKI, A., MORENCY, L.-P., AND GRATCH, J. All together now : Introducing the virtual human toolkit. In *International Conference on Intelligent Virtual Humans* (Edinburgh, UK, Aug. 2013).

[7] MORI, M. The uncanny valley. *Energy 7*, 4 (1970), 33–35.

[8] OGAN, A., FINKELSTEIN, S., MAYFIELD, E., D'ADAMO, C., MATSUDA, N., AND CASSELL, J. Oh dear stacy ! : social interaction, elaboration, and learning with teachable agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 39–48.

[9] PELACHAUD, C. Modelling multimodal expression of emotion in a virtual agent. *Philosophical Trans. of the Royal Society B : Biological Sciences 364*, 1535 (2009).

[10] SCHRÖDER, M. The SEMAINE API : towards a standards-based framework for building emotion-oriented systems. *Advances in HCI 2010* (2010), 2–2.

Alignement par Production d'Hétéro-Répétitions chez un ACA

Sabrina Campano Nadine Glas Caroline Langlet Catherine Pelachaud Chloé Clavel

Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI

prenom.nom@telecom-paristech.fr

Résumé

L'hétéro-répétition (HR) est la répétition intentionnelle par un locuteur des mots de son interlocuteur. Elle a plusieurs fonctions, comme transmettre qu'un message a été reçu, ou exprimer une attitude émotionnelle. C'est un phénomène d'alignement conversationnel fréquent, témoignant de l'implication du locuteur dans l'interaction. Dans cet article, nous proposons un modèle permettant à un Agent Conversationnel Animé de réaliser une HR ayant pour fonction d'exprimer une attitude affective, comme la surprise ou l'appréciation. Notre but est de faire paraître l'ACA impliqué dans la conversation, en faisant l'hypothèse que cela favorise l'engagement de l'utilisateur. La sélection d'une HR est établie grâce à un arbre de décision construit d'après des travaux existants en Analyse Conversationnelle, et la réalisation d'une HR est déterminée par son type et des variables contextuelles.

Mots Clef

Hétéro-répétition, engagement, alignement, Agent Conversationnel Animé.

Abstract

Other-Repetition (OR) is a current speaker's conscious reproduction of something a former speaker just said before. It has several functions, such as conveying that the message has been taken into account or expressing an emotional stance. It is an alignment process that occurs frequently in conversation, and is an indication that the speaker is involved in the interaction. In this paper we propose a model that provides a Conversational Embodied Agent with the capability of producing ORs that express an emotional stance, such as surprise or appreciation. Our goal is to model a conversational agent that seems involved in the conversation, hypothesizing that this will enhance user engagement. OR selection is determined by means of a decision tree which is defined according to existing literature on Conversational Analysis. The realization of an OR is guided by the OR type and contextual variables.

Keywords

Other-repetition, engagement, alignment, Embodied Conversational Agent.

1 Introduction

Favoriser l'*engagement* de l'utilisateur est un élément important dans une interaction avec un ACA. L'engagement d'un locuteur est défini comme la valeur qu'il attribue au fait rester avec les/l' autre(s) participant(s) et de poursuivre de la conversation [18]. Pour engager l'utilisateur, un ACA peut montrer qu'il est impliqué dans la conversation. L'implication est en effet une condition de l'engagement [17], et peut être montré par des phénomènes d'*alignement*. Au sens général, l'alignement est l'approbation par deux participants d'une activité en cours [9]. Les phénomènes d'alignement incluent les répétitions lexicales [26] [1], le style linguistique [14], l'activité vocale [4], les concepts [2], et l'inter-compréhension [21].

Dans cet article, nous nous intéressons à un phénomène d'alignement multi-modal qui est l'*hétéro-répétition* (HR). L'hétéro-répétition (HR) est une pratique fréquente dans le dialogue, qui consiste pour un individu à répéter ce que l'interlocuteur vient de dire [16]. L'HR se distingue de la simple répétition lexicale : c'est une répétition *volontaire*, ayant une *fonction* particulière. Plusieurs travaux en Analyse Conversationnelle sur l'HR décrivent ces fonctions en s'appuyant sur des corpus de langage parlé. L'HR permet par exemple de signifier qu'un message a été pris en compte, de demander une confirmation du message précédent, de montrer son accord ou son désaccord [16]. Elle facilite la compréhension, montre que l'on contribue à un sujet de discussion, et de façon plus générale, indique que l'on est impliqué dans la conversation [24]. Elle permet aussi d'exprimer une *attitude affective*, ou *évaluation*, en réaction à ce que l'interlocuteur vient de dire [23].

Il ne semble pas exister à ce jour de modèle informatique permettant à un ACA de réaliser des HRs. En revanche, il existe des modèles centrés sur d'autres phénomènes d'alignement. Par exemple, Rich et. al (2010) [19] ont proposé un système de détection d'engagement de l'utilisateur en étudiant l'alignement des regards, et Ochs et al. (2013) [15] ont travaillé sur l'alignement des sourires entre un ACA et un utilisateur. Au niveau verbal, Jong et al. [5] ont proposé un modèle d'alignement sur la politesse d'une phrase pour un ACA. Il prend en compte la structure de la phrase, le niveau de formalité du vocabulaire, et la distinction entre

le *vous* de vouvoiement et le *tu*. Kopp et al. (2008) [10] ont défini un modèle permettant à un ACA d'émettre des *feedbacks* (ex : « mhm », « oui »), qui sont de courts signaux multi-modaux émis par l'interlocuteur. Ils peuvent signaler une incompréhension, ou inciter le locuteur à continuer son récit.

Dans une étude que nous avons réalisée précédemment sur le corpus d'interaction humain-agent SEMAINE [3], nous avons identifié plusieurs types d'HRs. Nous avons étudié un sous-ensemble du corpus, où un opérateur humain joue le rôle d'un agent. Nous avons utilisé des méthodes de détection automatique sur les fichiers de transcription, et leur analyse nous a permis de tirer deux conclusions importantes : les HRs apparaissent fréquemment dans ce contexte d'interaction humain-agent, et les HRs permettant d'indiquer qu'un message a été reçu et d'exprimer une appréciation ou un affect étaient particulièrement représentées. Dans une interaction, un ACA (non joué par un être humain) devrait donc être capable de produire ce type d'HRs. Notre objectif est de proposer un modèle permettant à un ACA de réaliser des HRs ayant une fonction AFFective (HR-Aff), dans le but de favoriser l'engagement de l'utilisateur. Dans cet article, nous nous concentrons sur les aspects verbaux de l'HR-Aff. Définir un modèle de production d'HR-Aff demande de répondre à plusieurs questions de recherche : (i) Quels mots peut répéter l'ACA pour produire une HR-Aff ? (ii) Quelle fonction communicative peut-il exprimer à travers elle ? (iii) Comment peut-il exprimer cette fonction ? Nous répondons à ces questions en définissant un modèle formel permettant à un ACA de réaliser des HR-AFF. Nous donnons d'abord la définition des concepts utilisés et le principe général du modèle (Section 2), puis nous détaillons sa formalisation (Section 3). Enfin, des perspectives pour l'évaluation de ce modèle en cours d'implémentation sont proposées (Section 4).

2 Modèle d'Hétéro-Répétition

2.1 Contexte

Notre travail se déroule au sein du projet A1 :1, dans le cadre duquel un ACA à taille humaine va être placé au musée de la Vendée afin de discuter avec les visiteurs dans une interaction en face-à-face. Ce musée expose notamment une statue du sculpteur Jacques Bousseau, et un hélicoptère. L'ACA, nommé Léonard, possède une connaissance étendue des œuvres du musée (type d'œuvre, nom, artiste). Il doit donner des informations sur ces œuvres et leurs artistes, tout en favorisant l'engagement de l'utilisateur. L'environnement dans lequel se trouve l'ACA est équipé d'un micro et d'une caméra. L'interaction commence lorsqu'un visiteur s'approche de l'ACA, et que sa présence est détectée. La sélection des comportements verbaux et non verbaux de l'ACA est simulée sur la plate-forme GRETA [6], un agent conversationnel animé en 3D fonctionnant en temps-réel. Les intentions communicatives de l'agent et ses comportements sont décrits au format standardisé FML-BML [12, 25].

2.2 Concepts

Le modèle que nous proposons est fondé sur le travail de Svennevig [23], portant sur les *hétéro-répétitions* exprimant une *attitude affective* (que nous abrégeons HR-Aff). Une HR-Aff est une réaction à ce que l'interlocuteur vient de dire, qui exprime une appréciation positive / négative pour un sujet de conversation, de l'intérêt, ou de la surprise. Par exemple, dans l'extrait suivant, elle peut à la fois exprimer de la surprise ou une appréciation positive : A : « Oui, j'ai trois enfants » S : « **Trois enfants** ? *(sourires)* ». Selon Svennevig (2004), ce type d'HR peut avoir deux types d'effets : (i) inciter l'utilisateur à en dire plus, (ii) clore un sujet de conversation. Si le sujet de conversation a déjà été bien développé, une appréciation peut servir à le clore.

La définition de l'appréciation que nous adoptons est celle qui est présentée dans la théorie de Martin et White [13] sur l'*évaluation verbale*. Dans cette théorie, l'*appréciation* est une évaluation positive ou négative d'un objet (appelé *cible*), portant notamment sur des critères esthétiques. Par exemple « J'adore Jacques Bousseau » est une appréciation positive de la cible *Jacques Bousseau* [1]. Dans le cadre du projet A1 :1, la liste de *cibles* C contient des types d'œuvres (ex : sculpture), des noms d'œuvres (ex : Assiette Tripode), et des noms d'artistes (ex : Jacques Bousseau). Notre modèle utilise également une représentation des *préférences* de l'ACA sur l'ensemble des cibles de C, permettant de savoir si l'ACA apprécie ou non une cible $c \in C$.

2.3 Principe Général

Le module d'hétéro-répétition (HR) est intégré dans une architecture générale (montrée sur la figure 1) permettant de gérer une interaction. Il prend en entrée les sorties de différents modules :

- **DEV** : à partir d'informations verbales, détecte et renvoie en sortie les évaluations exprimées par l'utilisateur sur des cibles, dont les *appréciations*.
- **PSD** : détermine les *préférences* de l'ACA, planifie les sujets de dialogue abordés par l'ACA (un sujet représente une cible). Renvoie en sortie les *préférences* de l'ACA, et la décision de *continuer* ou arrêter un sujet de conversation.
- **Transcription de la Parole** : détecte et renvoie les mots prononcés par l'utilisateur, dont des noms de *cibles* qui peuvent être utilisés par le module HR.

Après chaque tour de parole de l'utilisateur, les différents modules envoient leurs sorties au module HR. La sélection et la construction d'une HR-Aff dépendent de ces entrées. Dans notre modèle, un ACA peut produire :

1. une HR-Aff de type surprise/intérêt (**HR-Aff-SI**) : l'agent exprime sa surprise / son intérêt par rapport une appréciation de l'utilisateur. Par exemple : « Ah bon vous n'aimez pas cette statue ? ».

1. Pour plus de détails sur ces notions dans un contexte d'interaction humain-agent, le lecteur peut se référer au travail de Langlet et Clavel (2014) [11]

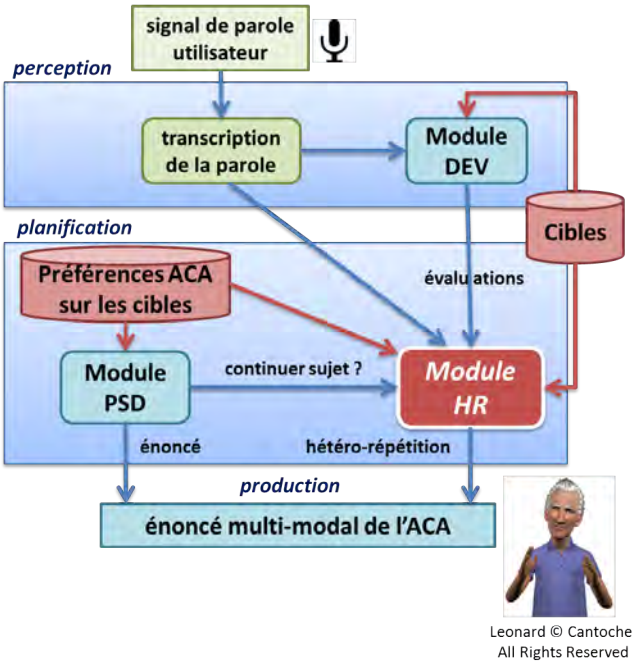


FIGURE 1 – Architecture générale du modèle d'interaction.

2. une HR-Aff de type appréciation (**HR-Aff-Appr**) : l'agent exprime une appréciation qui lui est propre. Par exemple : « Moi non plus je n'aime pas trop cette statue. ».

Un ACA choisit une HR-Aff-SI lorsque ses préférences divergent avec une appréciation formulée par un utilisateur, et que le sujet de conversation doit être poursuivi. Il choisit une HR-Aff-Appr dans les autres cas. Ce processus de sélection, guidé par un arbre de décision, est détaillé Section 3.2. Si une HR-Aff est sélectionnée par le module HR, elle est construite selon des spécifications verbales et non verbales dépendant de son type, que nous expliquons Section 3.3. Elle est ensuite jouée par l'ACA. Au même tour, PSD sélectionne également un énoncé à jouer par l'ACA lié à un sujet de conversation (ex : « Parlons de Jacques Bousseau »). Certaines contraintes assurent la compatibilité des sorties des modules HR et PSD (non détaillées ici par manque de place).

3 Formalisation

3.1 Entrées

Entrées du Module HR Fournies par le Module PSD. HR prend en entrée des informations fixes envoyées en début d'interaction, et des informations dynamiques envoyées au cours de l'interaction. Les informations fixes sont les préférences de l'agent a sur l'ensemble des cibles C, fournies par PSD. Une fonction de PSD associe à chaque cible $c \in C$ une valeur positive ou négative $v_a(c) \in \mathbb{R}$, représentant la valeur de c du point de vue de a. Une valeur négative ou positive signifie respectivement que a n'apprécie pas ou apprécie c. Si $v_a(c) = 0$, c est indifférent pour a.

Pendant l'interaction, PSD envoie à HR sa décision de poursuivre ou d'arrêter le sujet de conversation courant, représentée par une variable booléenne $continuer \in \{Vrai, Faux\}$.

Entrées du Module HR Fournies par le Module DEV. Soit I l'intervalle de temps correspondant au dernier tour de parole de l'utilisateur dans une interaction en cours. Le module DEV fournit au module HR la liste $A_u(I)$ des appréciations détectées chez l'utilisateur u sur I. Pour chaque appréciation $app \in A_u(I)$, DEV fournit une fiche sémantique de app qui est la suivante :

- $Pol_{app} = \{positive, negative\}$: *polarité* de l'appréciation.
- Src_{app} : *source* de l'appréciation, qui est la personne à l'origine de l'appréciation. Par exemple, si un utilisateur dit « Ma femme adore Klimt », alors la source est la femme de l'utilisateur. Dans notre modèle d'interaction, nous traitons les évaluations seulement lorsque la source est l'utilisateur.
- $Cible_{app} \in C$: *cible* de l'appréciation.
- $TermeLem_{app} \in T$: *terme d'appréciation* sous forme lemmatisée, parmi une liste T de termes d'appréciation.
- $TermeCat_{app} = \{ADJ, VERB\}$: *catégorie grammaticale* du terme d'appréciation, pouvant être un adjectif (ADJ) ou un verbe ($VERB$).

Nous détaillerons un exemple d'instanciation de cette fiche dans la section suivante.

Entrées du Module HR Fournies par l'outil de Transcription de la Parole. Soit I l'intervalle de temps correspondant au dernier tour de parole de l'utilisateur dans une interaction en cours. L'outil fournit la liste $W_u(I)$ des mots prononcés par l'utilisateur u sur I. Nous définissons $C_u(I)$ comme la liste des cibles détectées chez u sur l'intervalle I, définie par $W_u(I) \cap C$.

3.2 Sélection d'une HR-Aff

Notre système de sélection est un arbre de décision binaire inspiré de la théorie de Svennevig. Nous utilisons les attributs suivants :

1. $appreciation \in \{Vrai, Faux\}$: cette variable vaut $Vrai$ si le module DEV a renvoyé au moins une appréciation dans le dernier tour de parole de l'utilisateur ($\exists app \in A_u(I)$), et $Faux$ sinon. Si $appreciation = Vrai$, alors si l'agent formule une appréciation, elle sera *alignée* ou *désalignée* avec celle de l'utilisateur. Si $appreciation = faux$ alors l'agent formule une évaluation *simple*. Le fait qu'une HR-Aff-Appr soit alignée, désalignée ou simple a un impact sur sa forme verbale, comme expliqué Section 3.3.
2. $divergence \in \{Vrai, Faux\}$: indique si l'utilisateur a exprimé une appréciation sur une cible c qui diverge avec la valeur de c du point de vue de l'ACA. Soit app une appréciation formulée par un

utilisateur u sur une cible c, $divergence = Vrai$ si $(Pol_{app} = positive \wedge v_a(c) < 0) \vee (Pol_{app} = negative \wedge v_a(c) > 0)$, et $divergence = Faux$ sinon. Si $divergence = Vrai$ ou $divergence = Faux$ et que l'agent formule une HR-Aff-Appr, elle sera respectivement désalignée ou alignée avec l'appréciation de l'utilisateur.

3. $continuer \in \{Vrai, Faux\}$: la valeur de cette variable est fournie en entrée par le module PSD, qui décide d'interrompre ou de poursuivre un sujet de conversation. Si $continuer = Faux$, alors le module HR ne peut pas utiliser une HR-Aff-SI (qui inciterait l'utilisateur à poursuivre sur le sujet).

L'arbre de décision correspondant aux règles expliquées ci-dessus est montré sur la figure 2. Il est évalué seulement si l'utilisateur a prononcé le nom d'une cible $c \in C$. Dans le cas contraire, cela signifie que l'utilisateur n'a ni fait d'appréciation sur une cible, ni prononcé le nom d'une cible. Il est donc impossible pour l'agent de réaliser une HR.

Dans le cas où le module PSD renvoie plusieurs appréciations pour le dernier tour de parole de l'utilisateur ($card(A_u(I)) > 1$), ou lorsque l'utilisateur a prononcé le nom de plusieurs cibles ($card(C_u(I)) > 1$), c'est l'appréciation ou la cible prononcée le plus récemment qui est prise en compte dans l'arbre, et qui est utilisée dans la construction de l'HR-Aff (expliquée Section 3.3).

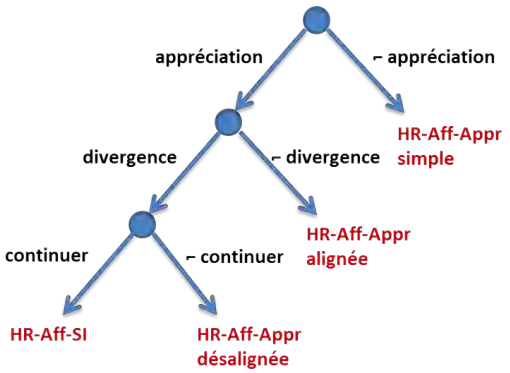


FIGURE 2 – Arbre de décision pour la sélection du type d'HR-Aff.

Exemple : soit la phrase « C'est une statue spectaculaire ! » représentant le dernier tour de parole de l'utilisateur. Le module DEV renvoie au module HR la liste $A_u(I) = \{app\}$, et la fiche sémantique de app est :

$Src_{app} = utilisateur$ $Pol_{app} = positive$
$TermeLem_{app} = spectaculaire$ $Cible_{app} = statue$
$TermeCat_{app} = ADJ$

$A_u(I)$ n'étant pas vide, l'attribut $appreciation$ vaut $Vrai$. La polarité de l'appréciation sur la cible $statue$ est positive ($Pol_{app} = positive$), et nous savons par le module PSD que la valeur de $statue$ pour l'agent a est positive ($v_a(statue) > 0$). L'attribut $divergence$ vaut donc $Faux$. D'après l'arbre de décision, c'est une HR-Aff-Appr alignée qui est sélectionnée.

3.3 Construction d'une HR-Aff

Spécifications Verbales Un ACA peut produire une HR-Aff en répétant le nom d'une cible et / ou en répétant un terme d'appréciation. Pour construire un schéma verbal autour du terme répété, nous nous inspirons des schémas d'appréciation qui sont utilisés par le module DEV à des fins de détection. Nous adoptons ici une version simplifiée de ces schémas.

Toute HR-Aff produite par l'ACA contient une appréciation app^{HR} qui représente : (i) dans une HR-Aff-Appr, l'appréciation de l'ACA sur une cible c en fonction de la valeur de c du point de vue de l'ACA ($v_a(c)$) (ii) dans une HR-Aff-SI, l'appréciation de l'utilisateur, qui est répétée par l'ACA en exprimant de la surprise. La fiche sémantique d'une app^{HR} comporte les informations décrites Section 3.1. De plus, une variable $Neg_{app^{HR}} \in \{Vrai, Faux\}$, indique si une forme négative doit être utilisée [2].

Nous définissons un patron pour chaque type d'HR présenté sur l'arbre de décision. Chaque patron ([]) est composé d'une forme fixe ayant des variantes (symbole |), et d'une appréciation app^{HR}. Ces patrons sont les suivants :

- **HR-Aff-SI** : [« Ah bon | Ah oui | Vraiment |∅ », app^{HR}, « ? »]. Dans ce cas, app^{HR} est identique à l'appréciation app formulée par l'utilisateur.
- **HR-Aff-Appr alignée** :
  - si $Neg_{app^{HR}} = Faux$: [« Moi aussi », app^{HR}, « . »] ;
  - si $Neg_{app^{HR}} = Vrai$: [« Moi non plus », app^{HR}, « . »].

  Dans ce cas, $Src_{app^{HR}} = agent$, et les autres attributs de app^{HR} sont égaux aux attributs de app.
- **HR-Aff-Appr désalignée** : [« Moi », app^{HR}, « . »]. Dans ce cas, $Src_{app^{HR}} = agent$, $Pol_{app^{HR}}$ est l'inverse de Pol_{app}, et les autres attributs sont identiques à app.
- **HR-Aff-Appr simple** : [app^{HR}, « . »]. Dans ce cas, $Src_{app^{HR}} = agent$, $Cible_{app^{HR}} = c$, $Pol_{app^{HR}}$ correspond à la valeur de c pour l'ACA ($v_A(c)$). Par défaut, $TermeLem_{app^{HR}} = aimer$ et $TermeCat_{app^{HR}} = VERB$, c'est à dire que l'ACA formule une appréciation de type « j'aime c » ou « je n'aime pas c ».

Lorsqu'il y a un choix entre plusieurs variantes pour un même patron, une variante est sélectionnée aléatoirement. Une fois que les caractéristiques de la variable app^{HR} sont déterminées, elle est réalisée sous sa forme verbale. Lorsque $TermeCat_{app^{HR}} = ADJ$, app^{HR} est réalisée dans une structure attributive de la forme « je trouve que $cible$ est ADJ » si $Src_{app^{HR}} = agent$, ou « vous trouvez que $cible$ est ADJ » si $Src_{app^{HR}} = utilisateur$. Lorsque $TermeCat_{app^{HR}} = VERB$, l'appréciation est réalisée sous la forme « je VB $cible$ » si $Src_{app^{HR}} = agent$, ou

2. Une forme négative est utilisée lorsqu'un terme d'évaluation a une polarité opposée à la polarité de l'appréciation. Par exemple le verbe aimer a une polarité positive. Dans ce cas, si l'appréciation est de polarité négative, alors la forme négative « ne...pas » doit être utilisée.

« vous VB *cible* » si $Src_{app^{HR}} = utilisateur$.
Une synthèse de ces schémas est montrée sur la figure 3.
Exemple : soit « Eh bien... je n'aime pas trop les statues. » représentant le dernier tour de parole de l'utilisateur. Le module DEV renvoie au module HR la liste $A_u(I) = \{app\}$, et la fiche sémantique de app est :
$Pol_{app} = negative$ $Src_{app} = utilisateur$
$Cible_{app} = statue$ $TermeLem_{app} = aimer$
$TermeCat_{app} = VERB$
A l'issue du processus de sélection, le module HR a décidé de produire une HR-Aff-SI. Dans ce cas, app^{HR} est identique à app. Comme $TermeCat_{app^{HR}} = VERB$ et $Src_{app^{HR}} = utilisateur$ l'appréciation est réalisée sous la forme « vous n'aimez pas les statues ». Le patron correspondant à une HR-Aff-SI est « Ah bon | Ah oui | Vraiment », app^{HR}, « ? ». La variante « Ah oui » est sélectionnée aléatoirement. La forme réalisée de app^{HR} est intégrée au patron avec la forme fixe, ce qui produit l'HR-Aff-SI « Ah oui vous n'aimez pas les statues ? ».

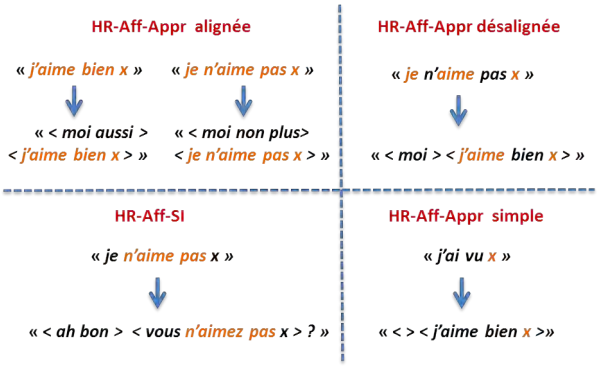


FIGURE 3 – Patrons de construction d'une HR-Aff : exemples types.

Spécifications Non Verbales Les spécifications non verbales utilisées sont des intentions communicatives issues de la plate-forme de l'agent GRETA [6], décrites au format standardisé FML-BML [12, 25]. Ces intentions et leur réalisation sont enregistrées dans un lexique (*Lexicon*), déterminé d'après des travaux en psychologie [20]. Par exemple, les intentions communicatives exprimant les émotions ont été établies d'après les travaux d'Ekman [8], qui a proposé un système de codage basé sur les muscles faciaux.
Pour notre modèle, nous utilisons 3 intentions communicatives correspondant à des expressions d'émotion :

- la surprise : cette intention est utilisée dans l'expression d'une HR-Aff-SI, dont la fonction est d'exprimer la surprise / l'intérêt [23].
- la joie : utilisée pour une HR-Aff-Appr de polarité positive. Dans notre travail, l'expression de joie correspond à l'expression d'un plaisir sensoriel esthétique, tel que défini par Eiseinbeger et al. (2010) [7].
- le dégoût : utilisé pour une HR-Aff-Appr de polarité négative.

4 Conclusion et Perspectives

Dans cet article, nous avons présenté un modèle informatique permettant à un ACA de réaliser des hétéro-répétitions exprimant une attitude affective. L'hétéro-répétition est un phénomène d'alignement conversationnel fréquent, témoignant de l'engagement des locuteurs dans l'interaction. Ce modèle est actuellement en cours d'implémentation dans la plate-forme GRETA, et nous prévoyons de réaliser une évaluation subjective à court terme. Pour cela, le modèle sera testé dans un scénario correspondant au cadre du projet A1 :1, dans lequel un utilisateur sera amené à discuter d'œuvres de musée avec l'agent. Afin de tester l'impact des hétéro-répétitions de l'ACA sur l'engagement de l'utilisateur, nous utiliserons un questionnaire d'engagement, comme celui précédemment décrit dans le travail de Sidner et al. [22].

Remerciements

Ce travail a été réalisé au sein du projet A1 :1 ainsi que dans le cadre du Labex SMART (ANR-11-LABX-65), et a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'Avenir portant la référence ANR-11-IDEX-0004-02.

Références

[1] Roxane Bertrand, Gaëlle Ferré, Mathilde Guardiola, et al. French face-to-face interaction : repetition as a multimodal resource. *Coverbal Synchrony in Human-Machine Interaction*, page 141, 2013.

[2] Susan E Brennan and Herbert H Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 22(6) :1482, 1996.

[3] Sabrina Campano, Jessica Durand, and Chloé Clavel. Comparative analysis of verbal alignment in human-human and human-agent interactions. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, 2014.

[4] Nick Campbell and Stefan Scherer. Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity. In *INTERSPEECH*, pages 2546–2549, 2010.

[5] Markus De Jong, Mariët Theune, and Dennis Hofs. Politeness and alignment in dialogues with a virtual guide. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pages 207–214. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[6] Etienne de Sevin, Radoslaw Niewiadomski, Elisabetta Bevacqua, André-Marie Pez, Maurizio Mancini, and Catherine Pelachaud. Greta, une plateforme d'agent conversationnel expressif et interactif. *Technique et science informatiques*, 29(7) :751, 2010.

[7] Robert Eisenberger, Ivan L Sucharski, Steven Yalowitz, Robert J Kent, Ross J Loomis, Jason R Jones, Sarah Paylor, Justin Aselage, Meta Steiger Mueller, and John P McLaughlin. The motive for sensory pleasure : Enjoyment of nature and its representation in painting, music, and literature. *Journal of personality*, 78(2) :599–638, 2010.

[8] Paul Ekman and Erika L Rosenberg. *What the face reveals : Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[9] Jan Gorisch, Bill Wells, and Guy J Brown. Pitch contour matching and interactional alignment across turns : An acoustic investigation. *Language and Speech*, 55(1) :57–76, 2012.

[10] Stefan Kopp, Jens Allwood, Karl Grammer, Elisabeth Ahlsen, and Thorsten Stocksmeier. Modeling embodied feedback with virtual humans. In *Modeling communication with robots and virtual humans*, pages 18–37. Springer, 2008.

[11] Caroline Langlet and Chloé Clavel. Modelling user's attitudinal reactions to the agent utterances : focus on the verbal content. In *5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data held at the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, 2014.

[12] Maurizio Mancini and Catherine Pelachaud. The fml-apml language. In *Proc. of the Workshop on FML at AAMAS*, volume 8, 2008.

[13] James R Martin and Peter RR White. *The language of evaluation*. Palgrave Macmillan Basingstoke and New York, 2005.

[14] Kate G Niederhoffer and James W Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4) :337–360, 2002.

[15] Magalie Ochs, Ken Prepin, and Catherine Pelachaud. From emotions to interpersonal stances : Multi-level analysis of smiling virtual characters. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 258–263. IEEE, 2013.

[16] Laurent Perrin, Denise Deshaies, and Claude Paradis. Pragmatic functions of local diaphonic repetitions in conversation. *Journal of Pragmatics*, 35(12) :1843 – 1860, 2003.

[17] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, Isabella Poggi, and Universita Roma Tre. Engagement capabilities for ecas. In *AAMAS'05 workshop Creating Bonds with ECAs*, 2005.

[18] Isabella Poggi. *Mind, hands, face and body : a goal and belief view of multimodal communication*. Weidler, 2007.

[19] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L Sidner. Recognizing engagement in human-robot interaction. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, pages 375–382. IEEE, 2010.

[20] Fiorella de Rosis, Catherine Pelachaud, Isabella Poggi, Valeria Carofiglio, and Berardina De Carolis. From greta's mind to her face : modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, 59(1) :81–118, 2003.

[21] Kevin Shockley, Daniel C Richardson, and Rick Dale. Conversation and coordinative structures. *Topics in Cognitive Science*, 1(2) :305–319, 2009.

[22] Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1) :140–164, 2005.

[23] Jan Svennevig. Other-repetition as display of hearing, understanding and emotional stance. *Discourse studies*, 6(4) :489–516, 2004.

[24] Deborah Tannen. *Talking voices : Repetition, dialogue, and imagery in conversational discourse*, volume 6. Cambridge University Press, 1992.

[25] Hannes Vilhjálmsson, Nathan Cantelmo, Justine Cassell, Nicolas E Chafai, Michael Kipp, Stefan Kopp, Maurizio Mancini, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, et al. The behavior markup language : Recent developments and challenges. In *Intelligent virtual agents*, pages 99–111. Springer, 2007.

[26] Arthur Ward and Diane Litman. Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education*, 2007.

The characterization of emotional body expression in different movement tasks

Nesrine Fourati Catherine Pelachaud

Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI

nesrine.fourati@telecom-paristech.fr
catherine.pelachaud@telecom-paristech.fr

Résumé

Dans cet article, nous étudions la caractérisation de l'expression émotionnelle dans différentes actions corporelles suivant un ensemble de paramètres. L'évaluation des paramètres a été réalisée par une étude perceptive.

Mots Clef

Mouvements corporels, Emotion, Caractèristiques corporelles

Abstract

In this paper, we investigate the characterization of emotional expression in different movement tasks using the same set of body cues. The evaluation of body cues was performed through a perceptual experiment.

Keywords

Body movement, Emotion, Body cues

1 Introduction

Since the last two decades, the study of emotional expression in movement received a lot of interest. Several studies have been conducted to associate distinct patterns of movement and postural behaviors with some emotions using Body Action/Posture Units and/or gestures (such as arms crossed in front of chest for Pride [1]). Bodily expression of emotions can also be signaled and described by the way a person is doing an action. Previous approaches mainly focused on one single movement task such as walking [2] or knocking at the door [3, 4]. The principal aim of our work is to build virtual characters able to express their emotional states through facial and body expressions while performing a large range of actions. To reach our aim we ought to characterize emotional body expression in various movement tasks based on body cues.

2 Emotional body behaviors collection

The collection of emotional behaviors is described in [5]. The actors were eleven (6 females and 5 males) graduate students. Although most of them have received theater courses since a long time, a professional acting director was hired to give them 7 training sessions regarding the use of body movements to express emotional states. We asked the actors to express 8 emotional states (Joy, Anger, Panic Fear, Anxiety, Sadness, Shame, Pride and Neutral) while performing different actions [5]; walking, sitting down, knocking at the door, lifting and throwing objects with one hand (a ball made of paper), and moving objects (books) on a table with two hands [5]. We recorded 3D motion capture data of the whole body as well as synchronized videos [5].

3 Perceptual experiment

A perceptual experiment was conducted to evaluate the emotion expressed by the actors and the characteristics of body posture and movement. First results of emotion perception task was reported in [5]. In this paper, we introduce only the results of body cues rating.

The popular crowd-sourcing website Amazon Mechanical Turk (AMT) was used to collect the results of body cues rating. 1008 participant took part in our study (56.01% of females and 43.98% of males). Since our database consists in a large set of emotional behaviors sequences (around 7000 sequences), we select a subset of motion sequences while considering one sample per actor, emotion and action (664 videos totally). Participants were asked to visualize 16 videos where the emotional body expression is reproduced through a computer avatar (the default 3D Studio MAX biped model of 3D Studio MAX software [5]). For each video, the participants were asked to rate the body expressive cues and perceived emotion using a 5-point scale (from 1 to 5). Each video was evaluated 24 times.

We defined a set of body cues that describe postural information (body shape), postural changes and the quality of movement dynamics based on the body movement coding

schema described in our previous work [6]. The proposed body cues are the straightness, the sagittal leaning and the openness of the whole body posture, the quantity and the regularity of arms movement and finally the speed, the fluidity and the power of body movement (See Figure 1). Low and high scores in the 5-point scale depict respectively low and high intensity of body cue rating (e.g. 1-5 refers to slow-fast in speed feature).

4 Results

Statistical results: Each body cue was subjected to one-way Anova for the set of the expressed emotions to evaluate its discriminative power. We found that each body cue is important for discriminating between emotions with a significant level ($p<0.001$). Besides, we conducted the Tukey test to investigate the difference of rating each movement feature from one emotion to another. We found that the rating of body cues is highly correlated with the expressed emotions. For instance, the rating of the power of body movement was significantly higher and different ($p<0.001$) for Anger expression (mean=3.78, see Fig. 1) than the other expressed emotions across all the actions. This result was also reported in previous studies [4].

Emotion expression characterization: The patterns of body cues used to characterize each emotion expression across all the actions are also congruent with previous works. For instance, as reported in previous studies [4, 1, 2], Sadness expression was characterized with light, smooth and slow movements, a small amount of arms movement, regular arm movements, a small body shape, a forward body leaning and a collapsed body posture. Anger expression was characterized with a different pattern of body cues. Figure 1 depicts the characterization of Anger and Sadness expressions across all the actions.

Emotions classification: We aim to study whether, for a given action, the expression of an emotion can be differentiated from the others through body cues ratings. For this purpose, we build, for each action 8 binary one-versus-all (OVA) SVM classifiers to classify an emotion against the others (8 emotions*7 actions=56 classifiers in total). The Gaussian Radial Basis Function was used to map the training data into the kernel space and the Sequential Minimal Optimization method was used to find the separating hyperplane. F-measure was calculated for each OVA classifier. The F-measure of each OVA classifier was above the chance level. The classification of Sadness against the other emotions across all the actions received the best scores (37% on average), followed by Neutral (30% on average) and Anger (28% on average). The F-measures of OVA classifiers related to most of the emotions received the best scores in walking action. However, the F-measure of Anger expressions classification against the other emotions was significantly better in Throwing action, while the F-measure of Panic Fear expressions classification against the others received the best scores in Moving books and Lifting action.

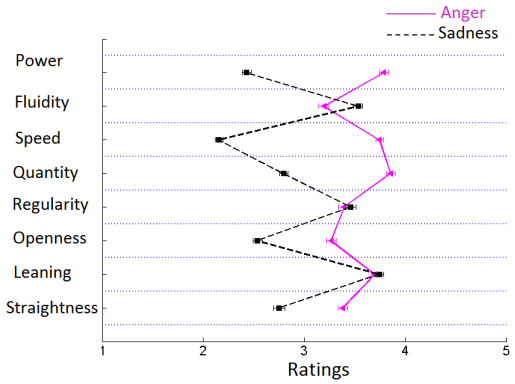


Figure 1: The characterization of Sadness (dashed line) and Anger (continuous line) expressions across all the actions

5 Conclusion and future work

This paper describes our attempt to characterize emotional body expression in different movement tasks (such as walking, sitting down). Emotional body behaviors were collected and the evaluation of body cues was conducted in a perceptual experiment. The first results that we obtained from the evaluation of body cues are congruent with previous studies and they are of a high interest since they introduce the characterization of emotional expression in different actions. The multiclass classification of all the emotions through body cues is part of our current goal. We aim also to investigate the classification between emotions using raters ground truth (emotions labeled as the most frequent emotion in emotion perception study) and the best body cues used to correctly classify each emotion. The detailed description of these results will be provided in our future work.

References

[1] Wallbott, H.G.: Bodily expression of emotion. J. of Social Psychology **28**(6) (1998) 879–896

[2] Montepare, J.M., Goldstein, S.B., Clausen, A.: The identification of emotions from gait information. J. of Nonverbal Behavior **11**(1) (1987) 33–42

[3] Pollick, F.E., Paterson, H.M., Bruderlin, A., Sanford, A.J.: Perceiving affect from arm movement. J. of Cognition **82**(2) (2001) B51–B61

[4] Gross, M.M., Crane, E.A., Fredrickson, B.L.: Methodology for Assessing Bodily Expression of Emotion. J. of Nonverbal Behavior **34**(4) (2010) 223–248

[5] Fourati, N., Pelachaud, C.: Emilya: Emotional body expression in daily actions database. LREC 2014 (2014)

[6] Fourati, N., Pelachaud, C.: Collection and characterization of emotional body behaviors. Int. Workshop on Movement and Computing (MOCO14) (2014)

Laughing Body

Y. Ding[1] T. Artières[2] C. Pelachaud[1,3]

[1] Institut Mines-Télécom, Télécom Paristech
[2] Pierre and Marie Curie University (Lip6)
[3] CNRS - LTCI UMR 5141

thierry.artieres@lip6.fr
{ding, catherine.pelachaud}@telecom-paristech.fr

Abstract

A statistical framework is proposed to control the laughing body of virtual character agent. The framework takes laughter sound intensity as input. Because of the periodicity of laughter body motion, body motion can be considered as the combination of three different frequency signals. During the training step, the framework captures the relation between intensity signal and output signals such as frequency signals. During the synthesis step, the trained framework takes as input intensity signal and works as generator of frequency signals. Generated body motion is used to animate a laughing virtual character agent.

Keywords

Laughter, Virtual Character Agent, Animation Synthesis, Data-driven.

1 Introduction

Laughter is an important social signal in human communication. It occurs frequently with positive emotions and even more to cheerful mood [1]; it is used as reaction to humorous stimuli or as marker of human pleasure when praised statements are received [2]; it is used to mask embarrassment [3] or to be cynical; it can play the role of social indicator of in-group belonging[2]; it can act as speech regulator during interlocutors [2] [4]; it is very contagious and can be used to elicit laughter in human conversation.
[5] categorizes laughter sound into 14 classes called pseudo-phoneme in reference to phoneme in speech. For simplicity, laughter pseudo-phoneme is called phoneme. [5] decomposes laughter sound stream as phonemes with their sequences of intensity and duration.
Darwin reported "During excessive laughter the whole body is often thrown backward and shakes, or is almost convulsed" [6]. Ruch and Ekman [7] described laughter movements as "rhythmic patterns", "rock violently sideways, or more often back and forth", "nervous tremor ... over the body", "twitch or tremble convulsively". Melo et al. [8] built a virtual character which "convulses the chest with each chuckle". It means that periodic motions of head and body are very prominent during laughing. The periodicity of body motion was used to distinguish between different videos of laughter in [9]. Ruch and Ekman [7] reported that rhythmical patterns during laughter usually were characterized by frequency around 5 Hz. Mancini et al. [9] observed 8 videos, which show actors laughing while watching funny images. Laughing actors produce rhythmic body movements with frequencies in the range of $[1.27Hz\ 3.66Hz]$.
Our aim is to build a body motion synthesis for a laughing virtual character agent. Body motion is inferred from phoneme intensity and duration sequences, which can be extracted from laughter sound stream by [5].

2 Methodology

We first gather a corpus of laughing movement. Participant watches funny movies. Their movements are recorded using motion capture. We recorded the data of 8 actors. We segmented the motion capture data to keep only the laughter episodes. We got 540 episodes. For each laughter episode, a phoneme sequence can be recognized by [5]. This sequence contain N phonemes. Phoneme intensity (I) and duration (D) sequences, $A = \{(I_n, D_n), n = 1...N\}$, are also calculated from recorded laughter sound signal by [5]. $\{D_n, n = 1...N\}$ in this sequence can be used to segment the whole body motion as concatenation of segmented motion M_n. So, a recorded sound and motion data can be considered as a sequence of phoneme segment, $AV = (I_n, D_n, M_n)$.
As described in Section 1, body motion is shaking periodically as rhythmic pattern. Therefore, body segmented motion, M_n, can be considered as the combination of several periodic signals, $\{PS\}$, which can be extracted by [10]. For simplicity, body motion is considered as the combination of only 3 periodic signals with the highest energy. These 3 periodic signals are noted by PS^H, PS^M and PS^L respectively for high, median and low energy. So, AV contains 5 elements, (I, D, PS^H, PS^M, PS^L).

Furthermore, all the phoneme segments, $\{AV\}$, in dataset are clustered by taking into account only intensity element, I. Clustering is conducted by Gaussian Mixture Model (GMM), Λ_Q, with Q states; I_n is taken as a state observation. Therefore, each state, q, is a set of AV. This set is noted by AV_q, where AV contains I with approximate values. For each AV_q, one can get a set of PS^H, a set of PS^M and a set of PS^L from AV_q. We note these 3 sets as $\Omega_q = \{PS^H, PS^M, PS^L\}$.
Then, a body generator, $\Phi = \{\Lambda_Q, \Omega_q, q = 1, 2, ..., Q\}$, is built. Λ_Q is a GMM. It is trained with observation with I. Ω_q is built for each GMM state and corresponds to 3 sets of periodic signals such as PS_q^H, PS_q^M and PS_q^L.
Finally, phoneme intensity and duration sequences, $A = \{(I_n, D_n), n = 1...N\}$, are taken as input to the trained GMM, Λ_Q, which determines state sequence and their durations, $S = \{(q_n, D_n), n = 1...N\}$. q_n in S is replaced by Ω_{q_n}. We obtain $S^\Omega = \{(\Omega_n, D_n), n = 1...N\}$; for each (Ω_n, D_n), one can extract randomly 3 periodical signals respectively from PS_n^H, PS_n^M and PS_n^L, which exist in Ω_n. D_n defines lasting duration of 3 extracted signals. Then, these extracted signals are combined into one signal, PS_n^{syn}. So, one can get a synthesized body motion by concatenating $\{PS_n^{syn}, n = 1...N\}$. In the connection between 2 PS_n^{syn}s, a low-pass filter is used to smooth motion stream. Therefore, a smoothing and shaking body motion can be produced from phoneme intensity and duration sequence.

References

[1] Ruch, W., Kohler, G., Van Thriel, C.: Assessing the 'humorous temperament': Construction of the facet and standard trait forms of the state-trait-cheerfulness- inventory - stci. Humor: International Journal of Humor Research **9** (1996) 303–339

[2] Provine, R.R.: Laughter: A scientific investigation. Penguin books edn. (2001)

[3] Huber, T., Ruch, W.: Laughter as a uniform category? A historic analysis of different types of laughter. In: Congress of the Swiss Society of Psychology. (2007)

[4] Provine, R.R.: Laughter Punctuates Speech: Linguistic, Social and Gender Contexts of Laughter. Ethology **95**(4) (1993) 291–298

[5] Urbain, J., Çakmak, H., Dutoit, T.: Automatic phonetic transcription of laughter and its application to laughter synthesis. In: Proceedings of ACII. (2013) 153–158

[6] Darwin, C.: The expression of the emotions in man and animals. London: John Murray (1872)

[7] Ruch, W., Ekman, P.: The Expressive Pattern of Laughter. Emotion qualia, and consciousness (2001) 426–443

[8] de Melo, C.M., Kenny, P.G., Gratch, J.: Real-time expression of affect through respiration. JVCA **21**(3-4) (2010) 225–234

[9] Mancini, M., Varni, G., Glowinski, D., Volpe, G.: Computing and evaluating the body laughter index. In: Proceedings of HBU. (2012) 90–98

[10] Sethares, W., Staley, T.: Periodicity transforms. Signal Processing, IEEE Transactions on **47**(11) (1999) 2953–2964

The characterization of emotional body expression in different movement tasks

Nesrine Fourati Catherine Pelachaud

Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI

nesrine.fourati@telecom-paristech.fr
catherine.pelachaud@telecom-paristech.fr

Résumé

Dans cet article, nous étudions la caractérisation de l'expression émotionnelle dans différentes actions corporelles suivant un ensemble de paramètres. L'évaluation des paramètres a été réalisée par une étude perceptive.

Mots Clef

Mouvements corporels, Emotion, Caractèristiques corporelles

Abstract

In this paper, we investigate the characterization of emotional expression in different movement tasks using the same set of body cues. The evaluation of body cues was performed through a perceptual experiment.

Keywords

Body movement, Emotion, Body cues

1 Introduction

Since the last two decades, the study of emotional expression in movement received a lot of interest. Several studies have been conducted to associate distinct patterns of movement and postural behaviors with some emotions using Body Action/Posture Units and/or gestures (such as arms crossed in front of chest for Pride [1]). Bodily expression of emotions can also be signaled and described by the way a person is doing an action. Previous approaches mainly focused on one single movement task such as walking [2] or knocking at the door [3, 4]. The principal aim of our work is to build virtual characters able to express their emotional states through facial and body expressions while performing a large range of actions. To reach our aim we ought to characterize emotional body expression in various movement tasks based on body cues.

2 Emotional body behaviors collection

The collection of emotional behaviors is described in [5]. The actors were eleven (6 females and 5 males) graduate students. Although most of them have received theater courses since a long time, a professional acting director was hired to give them 7 training sessions regarding the use of body movements to express emotional states. We asked the actors to express 8 emotional states (Joy, Anger, Panic Fear, Anxiety, Sadness, Shame, Pride and Neutral) while performing different actions [5]; walking, sitting down, knocking at the door, lifting and throwing objects with one hand (a ball made of paper), and moving objects (books) on a table with two hands [5]. We recorded 3D motion capture data of the whole body as well as synchronized videos [5].

3 Perceptual experiment

A perceptual experiment was conducted to evaluate the emotion expressed by the actors and the characteristics of body posture and movement. First results of emotion perception task was reported in [5]. In this paper, we introduce only the results of body cues rating.

The popular crowd-sourcing website Amazon Mechanical Turk (AMT) was used to collect the results of body cues rating. 1008 participant took part in our study (56.01% of females and 43.98% of males). Since our database consists in a large set of emotional behaviors sequences (around 7000 sequences), we select a subset of motion sequences while considering one sample per actor, emotion and action (664 videos totally). Participants were asked to visualize 16 videos where the emotional body expression is reproduced through a computer avatar (the default 3D Studio MAX biped model of 3D Studio MAX software [5]). For each video, the participants were asked to rate the body expressive cues and perceived emotion using a 5-point scale (from 1 to 5). Each video was evaluated 24 times.

We defined a set of body cues that describe postural information (body shape), postural changes and the quality of movement dynamics based on the body movement coding

schema described in our previous work [6]. The proposed body cues are the straightness, the sagittal leaning and the openness of the whole body posture, the quantity and the regularity of arms movement and finally the speed, the fluidity and the power of body movement (See Figure 1). Low and high scores in the 5-point scale depict respectively low and high intensity of body cue rating (e.g. 1-5 refers to slow-fast in speed feature).

4 Results

Statistical results: Each body cue was subjected to one-way Anova for the set of the expressed emotions to evaluate its discriminative power. We found that each body cue is important for discriminating between emotions with a significant level ($p<0.001$). Besides, we conducted the Tukey test to investigate the difference of rating each movement feature from one emotion to another. We found that the rating of body cues is highly correlated with the expressed emotions. For instance, the rating of the power of body movement was significantly higher and different ($p<0.001$) for Anger expression (mean=3.78, see Fig. 1) than the other expressed emotions across all the actions. This result was also reported in previous studies [4].

Emotion expression characterization: The patterns of body cues used to characterize each emotion expression across all the actions are also congruent with previous works. For instance, as reported in previous studies [4, 1, 2], Sadness expression was characterized with light, smooth and slow movements, a small amount of arms movement, regular arm movements, a small body shape, a forward body leaning and a collapsed body posture. Anger expression was characterized with a different pattern of body cues. Figure 1 depicts the characterization of Anger and Sadness expressions across all the actions.

Emotions classification: We aim to study whether, for a given action, the expression of an emotion can be differentiated from the others through body cues ratings. For this purpose, we build, for each action 8 binary one-versus-all (OVA) SVM classifiers to classify an emotion against the others (8 emotions*7 actions=56 classifiers in total). The Gaussian Radial Basis Function was used to map the training data into the kernel space and the Sequential Minimal Optimization method was used to find the separating hyperplane. F-measure was calculated for each OVA classifier. The F-measure of each OVA classifier was above the chance level. The classification of Sadness against the other emotions across all the actions received the best scores (37% on average), followed by Neutral (30% on average) and Anger (28% on average). The F-measures of OVA classifiers related to most of the emotions received the best scores in walking action. However, the F-measure of Anger expressions classification against the other emotions was significantly better in Throwing action, while the F-measure of Panic Fear expressions classification against the others received the best scores in Moving books and Lifting action.

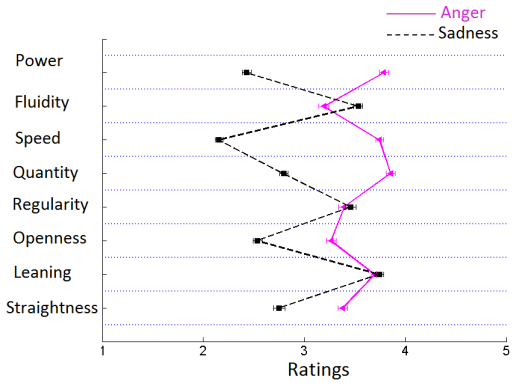


Figure 1: The characterization of Sadness (dashed line) and Anger (continuous line) expressions across all the actions

5 Conclusion and future work

This paper describes our attempt to characterize emotional body expression in different movement tasks (such as walking, sitting down). Emotional body behaviors were collected and the evaluation of body cues was conducted in a perceptual experiment. The first results that we obtained from the evaluation of body cues are congruent with previous studies and they are of a high interest since they introduce the characterization of emotional expression in different actions. The multiclass classification of all the emotions through body cues is part of our current goal. We aim also to investigate the classification between emotions using raters ground truth (emotions labeled as the most frequent emotion in emotion perception study) and the best body cues used to correctly classify each emotion. The detailed description of these results will be provided in our future work.

References

[1] Wallbott, H.G.: Bodily expression of emotion. J. of Social Psychology **28**(6) (1998) 879–896

[2] Montepare, J.M., Goldstein, S.B., Clausen, A.: The identification of emotions from gait information. J. of Nonverbal Behavior **11**(1) (1987) 33–42

[3] Pollick, F.E., Paterson, H.M., Bruderlin, A., Sanford, A.J.: Perceiving affect from arm movement. J. of Cognition **82**(2) (2001) B51–B61

[4] Gross, M.M., Crane, E.A., Fredrickson, B.L.: Methodology for Assessing Bodily Expression of Emotion. J. of Nonverbal Behavior **34**(4) (2010) 223–248

[5] Fourati, N., Pelachaud, C.: Emilya: Emotional body expression in daily actions database. LREC 2014 (2014)

[6] Fourati, N., Pelachaud, C.: Collection and characterization of emotional body behaviors. Int. Workshop on Movement and Computing (MOCO14) (2014)

Approche basée sur les traces d'interactions modélisées pour des agents socio-émotionnels dans les jeux vidéo

Joseph P. Garnier[1] Karim Sehaba[2] Elise Lavoue[3] Jean-Charles Marty[4]

Université de Lyon, CNRS
[1]Université Lyon 1, LIRIS, UMR5205, F-69622, France
[2]Université Lyon 2, LIRIS, UMR5205, F-69676, France
[3]Magellan, IAE Lyon, Université Jean Moulin Lyon 3, France
[4]Université de Savoie, LIRIS, UMR5205, F-69622, France

joseph.garnier@liris.cnrs.fr
Domaine principal de recherche : IA
Papier soumis dans le cadre de la journée commune : OUI ou NON

Résumé

Les concepteurs de jeux vidéo visent en permanence à créer un sentiment d'immersion aux joueurs afin de les plonger dans l'histoire qu'ils mettent en scène. Ce sentiment contribue grandement à la réussite d'un jeu vidéo. Pour ce faire, les concepteurs doivent créer un environnement, un monde, une histoire et des personnages non-joueurs (PNJ) crédibles. Chez les êtres vivants, en particulier les humains, les actions (et le comportement en général) sont guidées par les émotions et les relations sociales qu'ils entretiennent entre eux. Dans le but de rendre le comportement des PNJ plus crédible aux yeux des joueurs, nous proposons une approche basée sur les traces modélisées visant à doter des personnages virtuels d'émotions, en tenant compte de leur personnalité, et de relations sociales dynamiques. Dans un premier temps il s'agira de présenter un état de l'art sur les émotions et les relations sociales en sciences humaines et sociales, et en informatique. Dans un second temps, sur la base de ces recherches nous décrirons l'environnement, les acteurs et les interactions socio-émotionnelles au sein de notre jeu vidéo pour finalement exposer notre approche à base de traces modélisées. Enfin, nous discuterons des perspectives ouvertes par cette approche.

Mots Clef

émotions, relations sociales, informatique affective, agents, traces modélisées.

Abstract

The video game designers constantly aim to create a sense of immersion players to immerse in the story they depict. The feeling of immersion contributes greatly to the success of a video game. To do this, designers must create an environment, a world, a story and non-player characters (NPCs) credible. In living beings, particularly humans, actions (and behavior in general) are guided by emotions and social relationships they have with each other. In order to make the NPC's behavior more credible in the eyes of players, we propose a trace-based modeled approach to provide to virtual characters emotions, taking into account their personality, and dynamic social relations. Firstly, it will present a state of the art about emotions, and social relations in human and social sciences, and affective computing. Secondly, on the basis of this research we describe the environment, actors and socio-emotional interactions in our video game to finally present our trace-based modeled approach. Finally, we discuss the perspectives opened by this approach.

Keywords

emotions, social relations, affective computing, agents, trace-based modeled.

1 Introduction

On parle communément d'immersion comme d'une plongée dans l'eau, pour évoquer ainsi l'idée d'une expérience forte, absorbante, monopolisant toute l'attention de l'utilisateur ou du consommateur. L'immersion est au coeur de l'expérience vidéo ludique. Les jeux vidéo promettent en effet aux joueurs de vivre des situations "de l'intérieur". Une partie de l'immersion repose sur la crédibilité du comportement des personnages non-joueurs (PNJ). Hors, rares sont les jeux vidéo actuels où ceux-ci ont des réponses comportementales adaptées et convaincantes par rapport aux actions du joueur.

L'une des solutions envisagée est de doter les PNJ et les personnages joueurs (PJ) (personnages contrôlés par les joueurs) d'émotions et de relations sociales dynamiques afin de guider leurs comportements, comme c'est le cas

chez les êtres humains. Que ce soit en psychologie, en sciences humaines et sociales ou en physiologie, la littérature sur l'étude des émotions et des relations sociales est très riche, et peut alors servir de base à notre travail. Les PNJ et les PJ sont classiquement représentés par des agents. En informatique affective, la modélisation de processus émotionnels dans les systèmes computationnels est étudiée depuis les années 80, pour connaitre aujourd'hui une application de plus en plus importante en robotique notamment. Les recherches en sciences affectives de ces dernières années ont permis l'émergence de modèles intégrant la composante "sociale" aux processus émotionnels. Mais malgré les nombreux modèles socio-émotionnels existants, peu ont été conçus pour une application aux jeux vidéo. De plus, ces modèles se focalisent uniquement sur certains aspects de la problématique d'agents socio-émotionnels dans les jeux vidéo en ne tenant pas compte, par exemple, de la dynamique des relations sociales, ou encore de l'expérience acquise.

Ainsi, nous proposons d'intégrer totalement les interactions sociales dans la perception et l'expression émotionnelle des personnages joueurs et non-joueurs afin qu'ils aient un comportement plus cohérent face aux scénarios auxquels ils seront confrontés. Il s'agit, par exemple, d'intégrer des idéologies, des croyances ou des préjugés dans la perception et l'expression des émotions. Un personnage pourra réagir différemment envers d'autres personnages selon son expérience mais aussi selon l'appartenance à tel ou tel groupe social évoluant au gré de ses interactions. Contrairement aux approches existantes (par exemple [Ochs et al., 2009]) où l'émotion perçue par l'agent est déduite du potentiel émotionnel d'un évènement ou d'une action, nous proposons que les *agents perçoivent par inférence les émotions exprimées par leurs interlocuteurs à partir de leur expérience et de leur historique d'interactions*. Outre la perception d'émotions, notre approche *permettra aussi aux agents de choisir au mieux les comportements à adopter et les émotions à exprimer selon leurs buts, leurs croyance et leurs relations sociales*. Pour notre approche nous nous appuyons sur des traces modélisées d'interactions où à partir de règles de transformations, nous transformons des interactions de bas niveau, difficilement exploitables, en interactions de plus haut niveau. Ces interactions, riches en connaissances et pouvant être extraites, permettent à l'agent qui à partir d'une situation perçoit des émotions et adapte ses choix en vue d'atteindre ses buts.

Dans une première partie, un état de l'art sur les émotions et les relations sociales en psychologie sera présenté puis, dans une deuxième partie, nous exposerons notre approche pour finalement conclure par nos perspectives.

2 Etat de l'art

2.1 Emotions

La définition de l'émotion est souvent sujette à de nombreuses polémiques, à ceci près qu'il n'existe pas aujourd'hui de définitions consensuelles en psychologie pour expliquer la nature des émotions et pour définir leur représentation. Il n'est pas exagéré de considérer qu'il y a autant de définitions de l'émotion que de scientifiques travaillant sur le sujet. Dès 1981, plus de 140 définitions ont été relevées [Kleinginna and Kleinginna, 1981]. Les émotions étant principalement étudiées en psychologie, en neurosciences et sciences cognitives, la définition du mot "émotion" varie selon les disciplines et les époques. Le seul point sur lequel tous les chercheurs s'accordent, c'est que le concept est difficile à définir.

Au delà de son caractère personnel et individuel, l'émotion est ressentie par tous, humains et animaux [Scherer, 2001] [Darwin, 1872]. Cette permanence justifie son étude et son explication. Le champ de la psychologie propose un certain nombre de théories explicatives dont nous présentons ici les principales qui contribuent à la construction de notre approche.

Alors que les travaux de recherche sur la modélisation de l'émotion sont relativement récents et ont débutés réellement dans les années 50, depuis plus d'un siècle, quatre perspectives d'études ont été empruntées pour l'analyse du fonctionnement émotionnel :

- La perspective darwinienne où, les émotions sont *universelles* (on peut les trouver dans toutes les cultures et tous les pays), *adaptatives* (elles auraient favorisé la survie de l'espèce en permettant aux individus de répondre de façon appropriée aux exigences environnementales) et ont une fonction *communicative* (elles permettraient aux individus d'une même espèce d'être informés de ce que ressentiraient leurs congénères et des actions qu'ils seraient susceptibles d'entreprendre dans certaines situations). La fonction première des émotions est l'adaptation à l'environnement [Darwin, 1872].
- La perspective jamesienne où, *faire l'expérience d'une émotion c'est d'abord faire l'expérience des changements corporels ou physiologiques qui l'accompagne*. Sans la perception de ces changements, il est impossible de faire l'expérience des émotions. En effet, les changements périphériques suivent directement la perception du stimulus, et c'est la perception de ces changements qui constitue l'émotion [James, 1884].
- La perspective cognitiviste où, avant l'apparition d'une émotion, *le cerveau devait d'abord évaluer la situation* (théorie de l'appraisal) et décider si elle est potentiellement bénéfique ou néfaste pour l'organisme. Par la suite, le cerveau opterait pour une action conséquente avec son évaluation. C'est alors seulement que l'émotion émergerait, de cette prise de conscience de l'action d'approche ou de retrait [Arnold, 1960].
- La perspective socio-constructiviste où, *les émotions seraient des sortes de scripts applicables, régis par les normes socio-culturelles de références* et qui apparaîtraient de façon transitoire selon l'exigence des situations. Les réponses émotionnelles à la situation pourraient être automatiques du fait de l'intériorisation de ces scripts. C'est l'interprétation, dans la situation,

des liens qui unissent cette situation au système de valeurs et aux référents culturels qui feraient émerger l'émotion et les comportements subséquents. Ceci expliquerait notamment pourquoi les émotions diffèrent parfois d'une culture à l'autre [Averill, 1980].

Ces perspectives historiques ont permis aux théories contemporaines d'émerger et d'intégrer la cognition dans les débats. Ces théories sont issues des quatre perspectives historiques et peuvent être classées en trois catégories : les *approches catégorielles ou approches discrètes*, les *approches dimensionnelles* et les approches de l'*évaluation cognitive ou modèles componentiels*. La première approche s'est développée dès la fin des années 80 et est marquée par une inspiration néo-darwinienne. La seconde est apparue à la fin du 19^{e}siècle avec la théorie de Wundt. La troisième est apparue un peu plus tardivement dans le même temps que le développement des sciences cognitives en puisant dans les approches des perspectives jamesiennes, cognitives et socio-cognitives. Une des approches les plus actives et les plus influentes de la psychologie de l'émotion aujourd'hui est incarnée par les théories cognitives. Tout comme les théories des approches catégorielles, ces théories soulignent la fonction adaptative des émotions.

La plupart des théories des émotions utilisées actuellement comme base pour des modèles informatiques sont issues du courant cognitiviste, en particulier des théories de l'évaluation cognitive initiées par Arnold [Arnold, 1960]. Bien qu'aucun consensus n'existe sur le concept d'émotion, elle est généralement définie comme *un ensemble de variations épisodiques dans plusieurs composantes de l'organisme en réponse à des événements évalués comme importants pour l'organisme* [Scherer, 2001]. La définition de Scherer, en mettant l'accent sur la notion de changements synchronisés dans les différents sous-systèmes de l'organisme, permet de repenser la question de la séquence émotionnelle comme la question des inter-relations dynamiques entre les cinq composantes de l'émotion. Cette définition a également le mérite de ne pas se rapporter qu'à un aspect de l'émotion, ce qui est un problème récurrent dans les définitions qui ont été proposées pour définir l'émotion.

Nous nous appuyons sur la théorie de l'*évaluation cognitive* [Scherer, 2001] [Ortony et al., 1988]. Dans cette approche, l'émotion est vue comme un processus déclenché par une évaluation subjective d'un événement. Le *type* et l'*intensité* de cette émotion sont alors déterminés par la perception de cet événement ainsi que par l'évaluation d'un ensemble de variables appelées *variables d'évaluation* ou *SECs*. L'attention est particulièrement portée sur la détermination de ces variables dont les valeurs dépendent de l'état mental de l'individu (buts et croyances), de son profil psychologique (personnalité et préférences), et de facteurs situationnels et culturels [Lazarus, 1991][Scherer, 2001]. C'est ainsi qu'une même situation peut déclencher deux émotions différentes chez deux individus distincts. Il existe cinq composantes qui déterminent la succession de changements corporels et psychiques caractérisant une émotion [Scherer, 2001] :

- L'*évaluation cognitive* des stimuli ou des situations, qui permettent à l'individu d'évaluer à quel point un événement particulier, à l'origine du déclenchement de l'émotion, révèle une pertinence affective ;
- La *réaction du système nerveux périphérique*, qui prépare l'individu à une réaction urgente, par exemple l'augmentation de la fréquence cardiaque pour pouvoir courir et échapper à un agresseur, ou au contraire l'affronter ;
- La *tendance à l'action* ou *comportement d'adaptation* (également appelée *composante motivationnelle*), qui consiste à vouloir précipiter ou éviter un événement et qui permet aussi d'influer sur les interactions environnementales ;
- L'*expression motrice*, qui se caractérise par des modifications physiques adaptées, par exemple une expression à travers le visage, les paroles, la voix et les gestes ;
- Le *sentiment subjectif*, qui est la réflexion des changements se produisant dans toutes les composantes, permet la prise de conscience et la verbalisation du ressenti émotionnel.

2.2 Relations sociales

La situation de l'homme se représente à travers deux fonctionnements qui structurent sa vie et ses activités : l'*individuel* (de la perception de son environnement social au contrôle de l'action) et le *collectif* (processus de formation et de cohésion du groupe, établissement de liens affectifs entre membres d'un même groupe).

Le *fonctionnement collectif* concerne l'étude des petits groupes, l'analyse des interactions de toute nature entre l'individu et les groupes dont il fait partie, et la description de l'influence exercée par les groupes sociaux sur les fonctions psychologiques telles que la perception, la mémoire ou la motivation. La première théorie sur la *dynamique des groupes* est dû à Lewin [Lewin, 1959], à travers laquelle il met l'accent sur l'amélioration de l'efficacité individuelle et sociale par le groupe. Un groupe est *une association d'individus* entrant en interaction dans un contexte donné et poursuivant des buts communs. Les individus vont se doter de *rôles*, se *soumettre à des normes*, *partager des valeurs* et *réaliser des actions dans le cadre du groupe* auquel ils appartiennent. La force du groupe réside dans un système d'interdépendance. D'après Lewin, les forces au sein d'un groupe s'équilibrent naturellement et contribuent à sa dynamique. Le sentiment d'appartenance, la solidarité ou les échanges vont permettre d'orienter l'action du groupe dans deux directions : la pérennité de son existence et l'atteinte des objectifs fixés. L'appartenance à un groupe favorisera un processus d'apprentissage et l'adoption de certaines attitudes ou opinions chez ses membres. L'influence du groupe jouera sur les actions individuelles, chaque membre tenant compte de l'attitude des autres. En agissant sur un

élément particulier, par exemple en augmentant très fortement une force favorable au changement ou en diminuant le champ d'une force défavorable au sein du groupe, on peut modifier sa structure d'ensemble. La cohésion dans le groupe est maintenue par quelques facteurs : l'*affinité*, l'*attrait pour un objectif commun* et la *satisfaction de besoins personnels*. La cohésion peut se manifester par un ensemble de conduites collectives comme le conformisme, les conduites déviantes ou l'agressivité envers un autre groupe. Il existe en effet en permanence dans un groupe une certaine obligation de conformité. L'individu vit un conflit interne qui le partage entre ses propres convictions et les valeurs du groupe auquel il est supposé appartenir. Le fait de se conformer résulte d'une pression exercée par le groupe social. Ce sont les individus en mal d'estime de soi ou de confiance en soi qui sont les plus enclins à se conformer, simplement parce qu'ils recherchent la protection du groupe, ou veulent éviter d'en être exclus. Le degré de conformité d'un individu peut varier d'un groupe à l'autre ou d'une société à l'autre. L'attitude d'un *individu déviant* se caractérise par la non-conformité, il s'écarte délibérément des valeurs du groupe et privilégie ses propres valeurs ou celle d'un groupe de référence.

Le *fonctionnement individuel* concerne les mécanismes sous-tendant la génération des comportements individuels. Selon Moscovici [Moscovici, 1979] les *représentations sociales* sont essentielles à notre connaissance du sens commun et jouent un rôle tout aussi déterminant dans la vie mentale de l'individu que dans la vie des groupes. Elles sont une façon d'organiser notre connaissance de la réalité. Elles sont également un système de valeurs, de notions et de pratiques relatives à des objets ou à des aspects du milieu social, qui permet la stabilisation du cadre de vie des individus et des groupes, et qui constituent également un instrument d'orientation de la perception des situations et d'élaboration des réponses. Les représentations sont donc des régulateurs de la vie sociale. De plus, la *perception d'un individu* étant comprise comme un mécanisme régulateur de son action adaptative, elle prend nécessairement en compte tous les aspects du contexte social dans lequel l'action s'effectue. S'il veut pouvoir s'adapter, l'individu se doit donc de prendre en compte les comportements d'autrui afin d'ajuster les siens. Pour se faire, il devra anticiper ses actions en fonction du but à atteindre.

3 Proposition

Sur la base de notre état de l'art nous proposons une description générale de notre approche. A la différence de ce qu'il peut déjà exister en informatique affective (par exemple [Ochs et al., 2009]), cette approche propose que dans le but de ressentir et d'exprimer des émotions, les évènements et les actions ne soient pas porteurs de potentiel émotionnel. Au lieu de ça, c'est en inférant à partir de leur expérience et de leurs traces d'interactions (ou historique d'interactions), c'est-à-dire leurs connaissances, que les agents socio-émotionnels seront capable de ressentir les émotions exprimées par leurs interlocuteurs. Outre la perception, les traces d'interactions vont permettre aux agents d'exprimer des émotions adéquates et d'adapter leurs comportements en évoluant dans un jeu vidéo multijoueurs.

3.1 Présentation de l'environnement de jeu

Un jeu vidéo multijoueurs est composé de joueurs humains interagissant avec un système. Les joueurs dirigent, par exemple au moyen de leur clavier et de leur souris, leur personnage au sein du système, appelé *personnage joueur* (PJ). Les PJ évoluent dans un *environnement* composé d'entités et d'objets. Les *entités* sont les créatures peuplant le jeu ainsi que les *personnages non joueurs* (PNJ). Les PNJ sont contrôlés par une intelligence artificielle et réagissent aux événements provenant de l'environnement, c'est-à-dire des objets, des créatures, des PJ et des autres PNJ. Au cours de leurs interactions, les PJ et PNJ vont ressentir et exprimer des émotions qui vont faire évoluer leurs relations sociales. Quant aux relations sociales, elles vont moduler les émotions. La figure 1 présente l'environnement de jeu avec entre autres le graphe social entre PJ et PNJ.

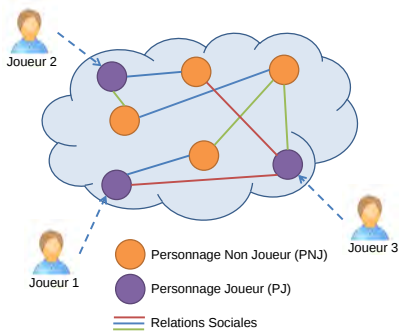


FIGURE 1 – Présentation de l'environnement de jeu. Les joueurs dirigent leur *personnage joueur* (PJ) au moyen du clavier de la souris. Les PJ interagissent avec les *personnages non joueurs* (PNJ), les autres PJ et le joueur. La figure ne représente pas les interactions mais les relations sociales (graphe social) entre les agents. La relation entre le joueur et le PJ représente le contrôle du joueur sur ce dernier.

3.2 Les interactions entre agents

Au sein du jeu, une interaction modélisée est orientée (elle va d'une source vers un destinataire) et on en distingue deux types : celles entre le joueur et son PJ et les autres, c'est-à-dire entre PJ-PJ, PNJ-PNJ, PJ-PNJ. Les interactions impliquant les créatures ainsi que les interactions entre agents et environnement ne seront pas abordées dans cet article.
Si l'on adapte les composants émotionnels décrits par Scherer [Scherer, 2001] à nos agents, nous avons : un *stimulus*, l'*évaluation cognitive*, l'*expression émotionnelle* et la *tendance à l'action*. Les interactions impliquant les PJ

et les PNJ peuvent être modélisées de la même manière. Au cours du processus émotionnel, il n'y a qu'une, deux ou trois interactions avec les autres agents. La première a lieu entre l'interlocuteur et l'agent lors de la réception du stimulus. Puis, dans la mesure où après l'évaluation cognitive le stimulus est significatif pour l'agent, une interaction a lieu entre l'agent et son interlocuteur pour lui signifier son expression émotionnelle. Une dernière interaction peut apparaitre si l'agent décide d'agir à l'encontre de son interlocuteur. Il est à noter que le composant *tendance à l'action* est particulier : après l'évaluation cognitive d'un PNJ le comportement est rationnel par rapport à son évaluation, alors que pour un PJ guidé par le joueur le comportement peut être irrationnel par rapport à l'évaluation et l'expression émotionnelle (par exemple le joueur peut décider d'attaquer malgré une expression de peur de la part du PJ). Avec un comportement irrationnel le joueur en subira les conséquences et comprendra qu'il devra prendre en compte lors de ses décisions, les réactions émotionnels de son PJ.

3.3 La composante sociale

Chaque PNJ a ses croyances, sa culture, ses aprioris et ses buts. Il en est de même pour les PJ mise à part que les buts sont déterminés et poursuivis par les joueurs à travers leur personnage. Comme les interactions ont une composante émotionnelle et sociale, les agents vont en tenir compte dans leurs buts et comportements pour ainsi faire évoluer leurs relations sociales qui sont de ce fait dynamiques. Lors de l'évaluation cognitive l'agent décide en fonction de ses buts, ses croyances, ses aprioris et sa culture d'exprimer (ou non) une émotion et réaliser une action qui, à eux deux, vont faire évoluer ses relations sociales. En voulant maitriser ses relations sociales, l'agent devra gérer l'intensité et le type de ses émotions de manière à ne pas aller à l'encontre de ses buts tout en veillant à rester cohérent afin de ne pas perturber l'immersion du joueur. Par exemple, si un agent A reçoit un coup d'épée de la part d'un agent "ami" B, l'agent A va exprimer une émotion de type et d'intensité différente du cas où l'agent l'ayant attaqué B, n'avait pas été pas son "ami".

3.4 L'expérience émotionnelle

De façon générale, au sein d'un système, les interactions ayant lieu entre les différents acteurs, que ce soit des humains ou des agents virtuels, sont capturées et conservées afin de constituer une trace d'interaction. Les traces d'interaction peuvent potentiellement contenir des connaissances pouvant être formalisées, partagées et réutilisées par le biais d'outils et de méthodes [Champin et al., 2013]. Dans notre approche, les agents vont avoir des interactions porteuses de "messages" émotionnels et sociales, exactement comme chez les humains. Chaque agent devra être capable d'enregistrer des traces d'interaction, puis d'en extraire, à partir de son expérience, de la connaissance telle que, l'émotion exprimée par ses interlocuteurs pour finalement exprimer lui même une émotion en fonction de ses buts et de ses relations sociales.

Afin de *percevoir les émotions* et de gérer au mieux leurs relations sociales, par l'*expression d'émotions adéquates* accompagnée de la *réalisation de bonnes actions*, les agents sont pourvus d'une expérience émotionnelle. L'*expérience émotionnelle* est représentée par les traces des interactions collectées au cours de la vie d'un agent. Ces traces vont contenir des informations sur les actions qu'il a subit, les actions qu'il aura effectuées, les émotions exprimées, les émotions ressenties, l'évolution de ses relations sociales, ... Formellement une *trace* est généralement définie comme un ensemble d'éléments observés temporellement situés. Les moyens courant permettant d'enregistrer et de conserver les traces sont les fichiers log, les flux RSS ou la mémoire, ... Associer un modèle aux traces, rend possible l'inférence.

Le *modèle de trace* est une description formelle de la structure et du contenu d'une trace. Le modèle fournit une information sur les propriétés de la trace à propos de ses éléments et des relations entre éléments.

Un élément observé est appelé un *obsel*, en référence à un élément de la trace. Un obsel est l'équivalent numérique d'un évènement produit dans le monde réel (par exemple un clic de souris). Le type d'un obsel est formellement décrit par le modèle de trace. Chaque type d'obsel est caractérisé par un nom, un "timestamp" et un ensemble de propriétés.

Un *M-Trace* (Modeled Trace) est une trace associée à un modèle. Si l'on veut faire l'analogie avec cette modélisation et la modélisation UML, le modèle de trace serait le diagramme de classe et le M-Trace serait le diagramme d'objet associé au diagramme de classe (une instanciation du diagramme de classe).

Les traces d'interaction capturées par le système sont appelées *Primary Traces* (ou traces primaires). Bien que riche en informations, cette trace est difficilement exploitable et nécessite des *transformations*. Les traces résultant d'une transformation sont appelées *Transformed Traces* (ou traces transformées), voir figure 2.

Les transformations permettent de créer une trace d'observés sur la base d'observés d'une ou de plusieurs traces selon des contraintes établies par avance. Une trace transformée représente un niveau de connaissance plus abstrait et plus interprétable qu'une trace non transformée ou de niveau de transformation inférieure. Une trace transformée peut subir d'autres opérations de transformation afin d'obtenir des traces de plus en plus haut niveau. Les opérations de transformation sont définies lors de la conception du modèle.

Dans notre exemple, nous considérons que l'agent a déjà de l'expérience et nous considérons uniquement les interactions socio-émotionnelles entre PNJ et PJ. Les interactions sont systématiquement composées d'une entité source et d'une observation : $\langle source, observation \rangle$. Celles-ci peuvent éventuellement être composées, en plus, d'un objet : $\langle source, objet, observation \rangle$. Lors de l'éva-

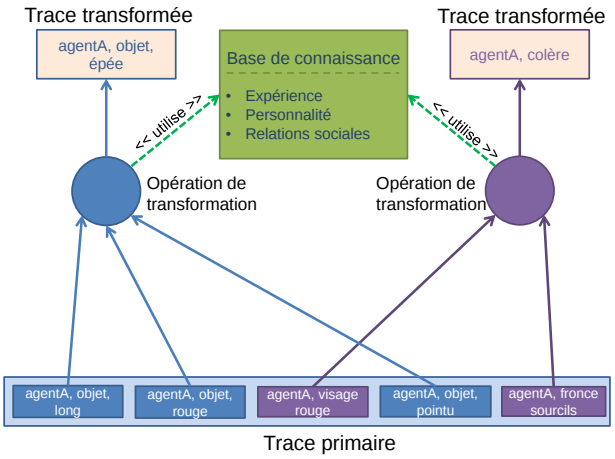


FIGURE 2 – Principe de transformation de traces primaire en trace transformée de plus haut niveau conceptuel à partir d'opérations de transformation dont les règles s'appuient sur la base de connaissance.

luation cognitive, l'agent va devoir traiter des *interactions primaires* difficilement interprétables de type visuel tel que $\langle AgentA, objet, long\rangle$, $\langle AgentA, objet, rouge\rangle$, $\langle AgentA, objet, pointu\rangle$, $\langle AgentA, objet, gris\rangle$, $\langle AgentA, visage\,rouge\rangle$, $\langle AgentA, fronce\,sourcils\rangle$ ou de type tactile tel que $\langle AgentA, tranche\rangle$. Le traitement va consister à conceptualiser les interactions à l'aide de règles de transformation, dans un contexte donné. Un *concept* est une interaction transformée de haut niveau (e.g $\langle AgentA, objet, épée\rangle$, $\langle AgentA, colère\rangle$, $\langle AgentA, ami\rangle$), créée à partir d'interactions de plus bas niveau (e.g $\langle AgentA, colère\rangle = \{\langle AgentA, visage\,rouge\rangle, \langle AgentA, fronce\,sourcils\rangle\}$. Comme le montre l'exemple, un concept peut tout aussi bien être un objet, une expression faciale ou une relation sociale. Quant au *contexte*, il est le reflet d'une situation vu au travers d'une séquence d'interactions primaires récemment reçues (e.g $\langle\langle AgentA, objet, pointue\rangle, \langle AgentA, objet, gris\rangle\rangle$). Les concepts, innés ou acquis, sont connectés les uns aux autres pour former l'*expérience*. L'expérience est donc un ensemble de concepts issus des traces d'interaction où chaque concept est relié directement ou non à un concept de type émotionnel. Au moment d'évaluer un stimulus, autrement dit une interaction, l'agent va s'appuyer sur le contexte afin d'utiliser son expérience en vue de percevoir au mieux l'émotion exprimée puis d'effectuer les meilleurs choix au niveau du comportement à adopter et de l'émotion à exprimer dans l'objectif d'atteindre ses buts (sociaux notamment).

4 Conclusion et perspectives

Dans l'objectif d'augmenter l'immersion du joueur en ayant un jeu cohérent, nous avons présenté un aperçu de l'état de l'art sur les émotions et les relations sociales suivi d'une proposition d'une approche basée sur les traces modélisées pour concevoir des agents socio-émotionnels. Bien que la littérature scientifique en psychologie, sciences humaines et sociales ainsi qu'en informatique affective sur les émotions et les relations sociales soit riche, peu de travaux en informatique proposent des modèles pour des agents mélangeant ces deux aspects. Dans cette première approche préliminaire, nous considérons que la connaissance est portée par les interactions et qu'il convient à l'agent de l'extraire afin d'adapter son comportement en fonction de ses buts et de l'exigence de l'environnement.

Remerciements

Ces travaux se déroulent dans le cadre du projet BeInG et d'une thèse CIFRE financée par l'entreprise Artefacts Studio et par l'ANRT.

Références

[Arnold, 1960] Arnold, M. B. (1960). *Emotion and personality*.

[Averill, 1980] Averill, J. R. (1980). A Constructivist View of Emotion. *Emotion : Theory, research and experience*, 1 :305–339.

[Champin et al., 2013] Champin, P.-A., Mille, A., and Prié, Y. (2013). Vers des traces numériques comme objets informatiques de premier niveau : une approche par les traces modélisées. *Intellectica*, (59) :171–204.

[Darwin, 1872] Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. Appleton.

[James, 1884] James, W. (1884). What is an emotion ? *Mind*, 9(34) :188–205.

[Kleinginna and Kleinginna, 1981] Kleinginna, P. R. and Kleinginna, A. M. (1981). A categorized list of motivation definitions, with a suggestion for a consensual definition. *Motivation and Emotion*, 5(3) :263–291.

[Lazarus, 1991] Lazarus, R. S. (1991). *Emotion and Adaptation*. Oxford University Press, New York.

[Lewin, 1959] Lewin, K. (1959). *Psychologie dynamique : les relations humains*. PUF, Paris, 3ème éditi edition.

[Moscovici, 1979] Moscovici, S. (1979). Social influence and social change. *European Journal of Social Psychology*, 9(4) :441–454.

[Ochs et al., 2009] Ochs, M., Sabouret, N., and Corruble, V. (2009). Simulation de la dynamique des émotions et des relations sociales de personnages virtuels. *Revue d'Intelligence Artificielle (RIA), Editions spéciale "Jeux vidéo"*, 23(2-3) :327–357.

[Ortony et al., 1988] Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press, first edit edition.

[Scherer, 2001] Scherer, K. R. (2001). Appraisal Considered as a Process of Multilevel Sequential Checking. In Oxford Press, U., editor, *Appraisal processes in emotion : Theory, methods, research*, pages 92–120.

Engagement-based Politeness Indications for Virtual Agents

Nadine Glas, Ken Prepin, Catherine Pelachaud
Institut Mines-Télécom
Télécom ParisTech
CNRS LTCI
{*firstname.lastname*}@telecom-paristech.fr

Abstract

We have looked at the impact of the perceived engagement level of an interaction participant on the perceived weight of Face Threatening Acts in human-human interaction. Our aim is to gather indications for modeling agents with human like behaviour regarding the question of whether or not agents that want to interact with a less engaged user should employ stronger politeness strategies than when they interact with more engaged users.

1 Introduction

For a range of applications we would like agents to engage its users. We consider engagement as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction" (Poggi, 2007 in: Peters et al., 2005). Numerous recent studies describe how an agent can influence user engagement by coordinating and synchronizing their behaviour with their user's. Such behaviour includes gaze (Sidner et al., 2005), gestures, postures, facial displays (Delaherche et al., 2012) and verbal behaviour (Campano et al., forthcoming). One of the verbal aspects that can be synchronized with the user's is the degree of expressed politeness, as described by De Jong et al. (2008). However, if the user has not expressed any politeness strategy in the current interaction so far, the agent would not know which one to use.

Therefore, we conduct a perceptive study to verify the existence of a link between the speaker's perceived engagement level of the hearer, and the speaker's politeness strategies in human-human interaction. This will give us an indication of whether or not a human-like agent who wants to continue the interaction with its user needs to speak with more caution to someone who is less engaged than to someone who is very engaged in the ongoing interaction.

Seen from another perspective, that of Brown & Levinson's politeness theory (1987), we hypothesize that the speaker's assessment of the hearer's level of engagement has an impact on the politeness level that the speaker employs in addressing the hearer.

In the following section we shall further specify this assumption while providing a short introduction to politeness theory. In section 3 we describe our evaluation method and in section 4 we present our results. In the final section 5 we conclude our findings.

2 Politeness Theory

Brown & Levinson's (1987) politeness theory is about saving the public self-image that every member wants to claim for himself, which is called this person's 'face'. They describe the concept of face as consisting of two components: a negative face which is the want of every 'competent adult member' that his actions be unimpeded by others; and a positive face which is the want of every member that his wants be desirable to at least some others.

According to Brown & Levinson (B&L) there are some acts that intrinsically threaten face. These acts are referred to as Face Threatening Acts (FTAs) and can be categorized into threats to the addressee's positive face (e.g. expressions of disapproval, criticism, disagreements) and threats to his negative face (e.g. orders, requests, suggestions, warnings). The speaker of an FTA can try to minimize the face threat of the FTA by employing a set of strategies. Brown & Levinson identified the following five strategies in increasing order of minimizing risk:

1. Without redressive action, baldly
2. Positive politeness

3. Negative politeness
4. Off record
5. Don't do the FTA

Roughly, the more dangerous the particular FTA x is, in the speaker's assessment, the more he will tend to choose the higher numbered strategy. Wx, the numerical value that measures the weightiness, i.e. danger, of the FTA x is calculated by:

$$W_x = D(S,H) + P(H,S) + R_x$$

where $D(S,H)$ is the value that measures the social distance between the speaker and the hearer, $P(H,S)$ is a measure of the power that the hearer has over the speaker, and R_x is a value that measures the degree to which the FTA x is rated an imposition in that culture.

The distance and power variables are intended as very general pan-cultural social dimensions. Brown & Levinson (1987) illustrate this by stating that $P(S,H)$ can be seen as great because H is for example a prince, and that $D(S,H)$ can be great because for example H speaks another dialect or lives in another valley.

However, the level of engagement can be seen as a measure for distance between participants in an interaction as well. Considering our definition of engagement (see above), a low level of engagement implies a temporally small value to *continue the interaction and be together with the other interaction participant(s)*, and vice versa. This distance may be comparable with Brown & Levinson's distance variable, only this time it has a more temporal and dynamic nature.

Having described the basics of politeness theory we can now reformulate our hypothesis into the following formula:

$$W_x = D(S,H) + P(H,S) + R_x - Eng(S(H))$$

where Eng(S(H)) is the speaker's perceived engagement level of the hearer.

3 Method

We are interested in the relation between the perceived level of engagement of the interaction partner and the perceived weight of an FTA. From Brown & Levinson's theory it is apparent that a straightforward way to infer the perceived threat of an FTA is by looking at the politeness strategy that is employed to formulate it. We thus need to create two conditions between which we can compare the employed politeness strategies; one interaction in which a participant is (highly) engaged and another in which he is less engaged.

We will model such conditions (interactions) for three different FTAs which we chose according to the context of this research: Building a virtual agent that represents a visitor in a French museum. The agent's goal is to engage human visitors in conversation about the museum and its objects. We therefore look at the FTAs: disagreement (in the preference for a painting), suggestion (to have a look at some other object) and request (for advice about what to see next).

For every FTA (and corresponding scenario) we design two interactions that represent the two conditions: one in which the perceived hearer's engagement is low and one in which it is higher. To ensure that the participant demonstrates the desired levels of engagement but all other variables of the interaction are kept as constant as possible, we prescribe the interactions. Human judgments regarding the appropriate politeness strategies will come from third party observers. For this study we will use solely written scenarios. In future studies we may add other modalities.

The design of our experiment can now be divided into two steps: 1) The formulation of a collection of dialogue strategies among which human judges can chose the most appropriate given one of both conditions; and 2) the design of the two different conditions (scenarios) in which the FTA will need to be judged. In section 3.2 we shall go into details about the latter issue. In section 3.1 we first explain our procedure for selecting and validating the politeness strategies.

3.1 Politeness strategies

According to Brown & Levinson's (1987) hierarchy of politeness strategies, and inspired by example sentences from De Jong et al. (2008), we constructed a maximum of French formulations for each FTA. The pronoun to address the hearer was kept constant to the less formal version *tu*. We did not design sentences with a mixture of strategies as this is a delicate matter that may even cause 'painful jerks' instead of an accumulating politeness (Brown & Levinson, 1987).

While theoretically we can rank our sentences according to their potential of minimizing the FTA's risk in the way B&L proposed, in practice

B&L's proposed hierarchy is not always entirely respected (e.g. De Jong et al., 2008). To deal with this issue we first validate our sentences as described below.

3.1.1 Strategy validation: Method

We developed a questionnaire to validate the perceived weights of the politeness strategies that will be used in our final experiment. This questionnaire consists of three parts, corresponding to the three different FTAs. Every part first introduces the context in which the sentences are supposed to be uttered: *Two young women in a museum meet thanks to a mutual friend*. This context corresponds to the scenarios presented in the final experiment. We also communicate the intention of the speaker (expressing his contradictory opinion, making a suggestion to have a look at a painting, or asking for advice about what to see next). In this way we avoid any speculation regarding the 'distance', 'power' and 'ranking' variables from B&L's formula. Below every context description we list all corresponding sentences, each followed by 6 questions regarding their plausibility and politeness level. We made sure that the order of the FTAs fluctuates between participants and that every participant is presented with a unique order of sentences. Also repetitions of sentence sequences and positions are avoided. The first question 1) asks for a simple 'yes' or 'no' to the question of whether or not the question seems plausible. The questions regarding politeness ask for rankings on a 7 point scale regarding respectively 2) the sentences' politeness level directly (De Jong et al., 2008); 3) the degree to which the speaker allows the hearer to make his own decision (Mayer et al., 2005, negative politeness); 4) the degree to which the speaker wants to work with and appreciates the hearer (Mayer et al., 2005, positive politeness); and 5) the degree to which the speaker spares the hearer's needs or face[1].

3.1.2 Strategy validation: Results

13 Native speakers of French have participated to this questionnaire. 8 of them are male, 5 female, and they range between the age of 23 and 40. ANOVAs for repeated measures conducted on the ranking of overall politeness (respectively question 2 and 5) show that for each FTA the sentences differ significantly from each other (disagreement: $F = 4.07$, $p < 0.01$ and $F = 5.84$, $p < 0.001$; suggestion: $F = 7.19$, $p < 0.001$ and $F = 8.01, p < 0.0001$; request: $F = 14.38, p < 0.0001$ and $F = 13.32, p < 0.0001$). Our results confirm other studies that B&L's ranking of politeness according to their strategies is not completely respected. Similar to the observations of De Jong et al. (2008) indirect strategies are rated much less polite than expected.

Out of all the sentences we make a selection for the final experiment. We first eliminate those sentences that are judged more than once as implausible in the sense of not correct or 'weird' French. For the remaining ones we look primarily to the mean score of their politeness level (question 2). We select one sentence for every observed level of politeness.

The described validation procedure cannot take into account B&L's heaviest risk minimizing strategy of not doing the FTA at all. However, considering the nature of this strategy as completely avoiding the FTA, we will indeed assume that there is no strategy heavier than this one and add it as such to the strategies that will be used in the final experiment.

3.2 Engagement conditions

We have the politeness strategies for which their use will be tested in both conditions. For every FTA we now need the two versions of dialogue that will provide a context for these strategies. The attributions of one of the dialogue participants, say Person A, will stay constant over both conditions, while the attributions of the other, say Person B, differ considerably in form in order to communicate a high or low level of engagement. As in the validation procedure we fix the power and distance variables by specifying that two young women meet in a museum through a mutual friend and start to talk. After about 15 turns, human judges are asked to recommend one of the sentences (each representing a politeness strategy) to Person A, under the instruction that this participant wants to place the FTA (communicate his disagreeing opinion, do a suggestion, or ask for advice) but also wants to absolutely continue the conversation with his interaction participant, Person B. The difference in which politeness strategy will be recommended then tells us whether the perceived weight of the FTA differs between both conditions.

For the dialogue transcriptions in which Person B is minimally engaged we keep her utterances as

[1] literally *mnager la susceptibilit*"

brief and uninterested as possible. Her engagement level is just high enough to participate in the interaction so far.

In the transcriptions where Person B is suppose to demonstrate a high level of engagement we use cues that have been linked to engagement in former studies and which are easily recognizable in written text: We longer Person's B reactions as to longer the interaction time (e.g. Bickmore et al., 2013), we add more backchannels (Gratch et al., 2006), add expressions of emotion (Peters et al., 2005.), and show interest in Person A (Peters et al., 2005)).

We verify whether or not we have successfully communicated the difference between the engagement levels of person B by asking the observers to judge B's attitude on several dimensions that are directly or indirectly linked to engagement. We ask 1) to rank on a scale from 1 to 7 the value that B attributes to the goal of being together with A; and 2) the value that B attributes to the goal of continuing the interaction (Peters et al., 2005.). These are the direct measurements for engagement. We also ask respectively at Person B's level of involvement, rapport and interest: 3) to what extend the interaction seems engaging for Person B (Lombard & Ditton, 1997 & 2000 in: Sidner et al., 2005); 4) if Person A and Person B want to become friends (Ringeval et al., 2013); and 5) if Person B seems interested in the interaction.

4 Results

89 people participated to our final experiment: 31 men and 58 female, ranging in age between 18 and 75. They are all native speakers of French. Every participant was exposed to one version (engaged or less engaged) of each scenario (FTA). The choice and order of the scenarios as well as the order of the suggested FTA sentences varied among the participants.

For every FTA, we performed t-tests on the questions that give us insight into the perceived engagement levels of participant B. Each one of them show us significant differences between the two conditions. This means that we have successfully created the engaged and less engaged conditions between which we compare the preferences for particular politeness strategies.

To test our final hypothesis, we performed a Man-Whitney U test (for ordinal data and no normal distribution assumed) on every FTA, regarding the sentences that were recommended to place (or even avoid) the FTA. We have found no significant difference between the two conditions.

5 Conclusion and discussion

In this study we have tested if the speaker's perceived engagement level of the hearer influences the speaker's perceived weight of his FTAs. We have done this by verifying if there is a difference between the weight of the politeness strategies that human would use in interaction with an engaged and less engaged person. The results presented in the former section have shown that we have found no significant evidence to support this hypothesis. This means that we have not proven that an agent that wants to continue the interaction with its user needs to speak politer to someone who is less engaged than to someone who is very engaged in the ongoing interaction. We can also reverse this statement by saying that we have found no proof that agents that speak with more engaged users can be less polite.

There are however some parameters to take into account when interpreting the results. First of all, the results may be distorted due to the fact that the FTAs can be interpreted as not really face threatening. While B&L categorize disagreements, suggestions and requests as FTAs, asking someone for his advice on what to see next can also be taken as showing interest in his values and knowledge, which is quite the opposite of a threat. Similarly, doing a suggestion can be interpreted as a remark that is placed purely in the interest of the hearer, which deletes its face threatening aspect as well.

It must further be noted that our experiment is based on written text and third party judges. In face-to-face interactions non verbal behaviour such as prosody, facial expressions and gestures play a big role. Non-verbal behaviour can influence the way in which verbal behaviour is interpreted and can reveal a range of information about the person's attitude and perceptions. Future research will have to show if multi-modal –and so perhaps more vivid– interactions lead to similar results.

We also leave for future work the analysis of the choice for one strategy over another within one condition. Perhaps we can make some interesting observations regarding for example the preference for either positive or negative strategies.

6 Acknowledgments

We would like to thank Brice Donval for his help with developing the websites that present the questionnaires; Nesrine Fourati for her 'matlab' help; and Sabrina Campano for translating our dialogues. We would also like to thank all the participants of our experiment. This research was funded by the French DGCIS project "Avatar 1:1".

References

Timothy W. Bickmore, Laura M. Vardoulakis Pfeifer and Daniel Schulman. 2013. *Tinker: a relational agent museum guide.* Autonomous agents and multi-agent systems.27(2):254–276. Springer

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage.* Cambridge University Press.

Sabrina Campano, Jessica Durand, Chloé Clavel. 2014. *Comparative analysis of verbal alignment in human-human and human-agent interactions* LREC Forthcoming.

Markus De Jong, Mariët Theune and Dennis Hofs. 2008. *Politeness and alignment in dialogues with a virtual guide.* Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1.,207–214.

Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux and David Cohen. 2012. *Interpersonal synchrony: A survey of evaluation methods across disciplines* Affective Computing, IEEE Transactions on,3(3):349–365. IEEE

Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, Rick J. Van der Werf, and Louis-Philippe Morency. 2006. *Virtual rapport.* Intelligent Virtual Agents,14–27. Springer

Matthew Lombard, and Theresa Ditton. 1997. *At the heart of it all: The concept of presence.* Journal of Computer-Mediated Communication. Wiley Online Library3(2)

Matthew Lombard, Theresa B. Ditton, Daliza Crane, Bill Davis, Gisela Gil-Egui, Karl Horvath and Jessica Rossman. 2000. *Measuring presence: A literature-based approach to the development of a standardized paper-and-pencil instrument.* Third international workshop on presence, Delft, The Netherlands. Wiley Online Library3(2)

Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini and Isabella Poggi. 2005. *Engagement capabilities for ecas.* AAMAS05 workshop Creating Bonds with ECAs.

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

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer and Denis Lalanne. 2013. *Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions.* Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on.,1–8 IEEE

Candace L. Sidner, Christopher Lee and Cory Kidd. 2005. *Engagement during dialogues with robots.* AAAI spring symposia.

Effet Proteus et amorçage : Quand l'apparence des personnages virtuels influence les comportements et les attitudes des utilisateurs

Jérôme Guegan[1] Stéphanie Buisine[1]

[1]Arts et Métiers ParisTech, LCPI
151, boulevard de l'Hôpital 75013 Paris, France

jeromeguegan@live.fr

Domaine principal de recherche: Psychologie
Papier soumis dans le cadre du WACAI

Résumé

À travers l'analyse des conceptions récentes et des théories classiques en psychologie, cette revue de question vise à fournir des éléments clés permettant de mieux comprendre l'influence que les personnages virtuels peuvent exercer sur les utilisateurs. Pour ce faire, cet article aborde l'étude des représentations digitales de soi (avatars) et des effets potentiels de l'apparence des agents conversationnels. Dans son ensemble, cette analyse permet d'éclairer les spécificités des interactions sociales en environnement virtuel.

Mots Clef

Effet Proteus, Amorçage, Avatars, Agents conversationnels Animés

Abstract

Through the analysis of recent concepts and classical theories in psychology, this review aims to provide key elements to better understand the influence that the virtual characters can have on users. To do this, this paper discusses the digital representations of self (avatars) and the effects of the appearance of conversational agents. Overall, this analysis provides insight into the characteristics of social interactions in virtual environments.

Keywords

Proteus effect, Priming, Avatars, Embodied conversational agents

1 Introduction

Le caractère malléable de l'auto-présentation dans les interactions en ligne est une composante essentielle de ce que signifie le fait d'avoir une identité virtuelle [38]. Les avatars (i.e., représentations digitales de soi) sont un exemple particulièrement représentatif de cette malléabilité du soi dans les environnements virtuels. Les possibilités offertes par la personnalisation de l'avatar permettent en effet de moduler les rôles sociaux, l'identité ou encore le genre des interlocuteurs [42]. Dans ce type d'interaction, une question centrale consiste à savoir si les avatars sont susceptibles de moduler les comportements des utilisateurs qui les incarnent.

Par ailleurs, l'analyse des interactions en environnement virtuel ne se limite pas à l'influence exercée par les avatars que l'utilisateur est susceptible d'incarner. Les sources d'influence sont en effet multiples puisqu'elles peuvent reposer sur l'apparence de l'avatar de l'utilisateur, de son (ses) interlocuteur(s), ou sur l'apparence des agents conversationnels susceptibles d'être intégrés à l'environnement. Dans cette perspective, il s'agit donc d'appréhender l'influence globale que l'apparence des personnages virtuels peut exercer sur les comportements et les attitudes des utilisateurs amenés à interagir en environnement virtuel.

Après avoir abordé les considérations théoriques classiques de l'anonymat et de ses effets sur la perception de soi, nous développerons des conceptions plus récentes appliquées aux environnements virtuels. Cette analyse permettra (1) de développer l'influence des représentations digitales de soi et (2) d'évoquer l'influence des autres représentations virtuelles (avatar d'autres utilisateurs, agents conversationnels animés) sur les comportements et les attitudes de l'individu. Par la mise en perspective de ces différentes sources d'influence, nous tenterons d'éclairer les spécificités des interactions sociales en environnement virtuel.

2 Bases théoriques

L'étude des effets que les représentations digitales de soi exercent sur les individus repose sur une tradition de recherche déjà ancienne. De nombreuses expériences classiques en psychologie sociale ont en effet abordé cette problématique, notamment via l'étude des effets des costumes. Sur un plan théorique, elles ont permis d'étayer et/ou de nuancer les propositions des théories de l'auto-perception et de la déindividuation.

2.1 Théorie de l'auto-perception

Selon la *théorie de l'auto-perception* [4], les individus observent leurs propres comportements pour inférer les attitudes et les dispositions personnelles qui peuvent en être à l'origine. Comme le souligne Bem, si les états internes de l'individu sont ambigus ou difficilement interprétables, ce dernier base son observation sur des indices externes, en adoptant la même position qu'un

observateur extérieur, « who must necessarily rely upon those same external cues to infer the individual's inner states » [4]. En accord avec ce postulat, l'expérience de Valins [39] a révélé que des participants amenés à croire que les battements de leur cœur augmentent à la vue de certaines photographies de femmes dénudées jugent les personnes apparaissant sur ces images comme significativement plus attractives. En accord avec la théorie de l'auto-perception, un observateur extérieur ayant accès aux mêmes informations aurait pu produire des inférences sensiblement similaires (à ce propos, voir aussi [32]). Frank et Gilovich [13] ont aussi montré que des participants portant des uniformes noirs présentent des comportements plus agressifs que des sujets vêtus de blanc – la couleur noire étant généralement associée à la mort et au mal [1] tandis que le blanc est davantage associé au bien, à l'entraide [23]. Ce phénomène a été observé en laboratoire, mais également suite à l'analyse des fautes et pénalités infligées à des équipes sportives selon les couleurs de leurs maillots : les joueurs de football américain et hockey sur glace se comportent de façon plus agressive lorsqu'ils portent des maillots noirs que lorsqu'ils portent des maillots blancs. Selon Frank et Gilovich [13], le processus mis en évidence est que l'observation de son apparence (i.e., je porte un costume noir) conduit les participants à faire des inférences implicites concernant leurs dispositions personnelles (i.e., je suis une personne agressive) qui aboutissent à des changements comportementaux effectifs (i.e., je choisis un type de jeu plus agressif).

L'influence des costumes sur le comportement a été étudiée de façon encore plus directe par Johnson et Downing [19]. Dans cette étude, des participantes vêtues d'une tenue rappelant le Ku Klux Klan (KKK) ou d'un costume d'infirmière devaient administrer des chocs électriques fictifs à un individu. Selon les conditions expérimentales, les participantes pouvaient donc revêtir un costume associé à des indices comportant une tonalité négative ou positive. Les résultats révèlent que les participantes vêtues en « KKK » administraient des chocs électriques de forte intensité tandis que les participantes vêtues en « infirmières » se montraient plus clémentes.

2.2 Théorie de la déindividuation

Initialement, la déindividuation correspond à un état caractérisé par une altération de la conscience de soi et de la capacité à raisonner sur ses actions d'une manière autocritique [11]. Plus précisément, les situations de foule ou d'anonymat en groupe provoqueraient une perte d'identité personnelle engendrant une baisse des inhibitions, du sentiment de responsabilité et d'exposition. Dans ce contexte, les individus peuvent donc réaliser des comportements qu'ils n'auraient pas mis en jeu s'ils avaient été seuls ou personnellement identifiables [11]. Zimbardo [44] a insufflé une tonalité plus négative à la notion de déindividuation en y intégrant un ensemble de facteurs qui minimisent le contrôle de soi lié à la honte, la culpabilité ou la peur et conduisent à des comportements qui transgressent les normes en vigueur : impulsion, irrationalité, comportement régressif, comportement antisocial, etc. Plusieurs expériences ont en effet montré que l'anonymat peut faciliter l'apparition des comportements négatifs ou contre-normatifs tels que l'administration de chocs électriques ou l'expression d'un discours obscène [7, 34, 44]. Dans la cadre d'une étude menée en milieu naturel le soir d'Halloween, Diener, Fraser, Beaman et Kelem [6] ont observé le comportement des enfants venus chercher des friandises. Lorsque le costume garantissait l'anonymat et que les enfants arrivaient en groupe, ils avaient significativement plus tendance à voler des pièces de monnaie et des bonbons supplémentaires (comportements transgressifs allant à l'encontre de la consigne donnée par l'adulte venu les accueillir) que lorsqu'ils étaient seuls ou personnellement identifiables.

Ceci étant, plusieurs travaux suggèrent que les conséquences de la déindividuation ne sont pas toujours négatives (pour une méta-analyse voir [29]). Par exemple, l'expérience de Johnson et Downing évoquée plus haut manipulait également la déindividuation (présence vs. absence de l'affichage du nom du participant). Les résultats révèlent que si la déindividuation augmente bel et bien l'intensité des chocs électriques en condition KKK, les « infirmières » se montrent quant à elles encore plus clémentes en condition de déindividuation que d'individualisation. Or, la théorie classique de la déindividuation ne peut rendre compte de cette accentuation du comportement pro-social des infirmières. Gergen, Gergen, et Barton [14] ont également montré que des participants déindividués (anonymes dans une chambre noire) ne se montraient pas plus agressifs mais avaient au contraire tendance à aborder des thèmes de conversation plus personnels et à se montrer davantage affectueux. Ces résultats sont donc également opposés à la conception initiale de la déindividuation portée par les propositions de Festinger et al. [11] et plus particulièrement de Zimbardo [44]. Autrement-dit, le fait que la déindividuation produise des changements comportementaux semble empiriquement bien établi [22, 31], mais ces changements ne sont pas nécessairement négatifs. Plusieurs auteurs ont donc réinterrogé les interprétations fatalistes et négatives de la théorie [19, 29]. Selon Gergen et al. [14], la déindividuation peut faire apparaitre des comportements pro-sociaux ou antisociaux, le facteur le plus déterminant étant en réalité la tonalité des différents indices identitaires présents dans la situation. Dans une acception assez proche, Diener [7] considère que la réactivité de l'individu déindividué aux stimuli immédiats n'est pas médiatisée par la conscience, mais se rapprocherait plus d'un schéma béhavioriste de type stimulus-réponse. Le corollaire de ce processus est que la déindividuation accentue la focalisation des individus sur les sollicitations immédiates de la situation. C'est donc la tonalité des indices contextuels qui oriente les effets du processus de déindividuation, et non l'anonymat en tant que facteur isolé. Ainsi, selon Spivey et Prentice-Dunn [36], la déindividuation est avant tout une condition neutre, mais qui maximise la sensibilité des individus aux influences de l'environnement. Il en résulte que l'impact des indices identitaires est

particulièrement fort lorsque les individus sont déindividués. En conséquence, l'anonymat peut maximiser l'effet des costumes portés puisque « la déindividuation augmente l'impact des indices identitaires sur l'auto-perception » [42]. Les costumes peuvent donc agir comme des points de repères identitaires qui constitueront une image de soi ponctuelle (pro-sociale ou antisociale), d'autant plus propice à moduler les comportements que les individus sont déindividués.

En résumé, l'auto-perception permet de créer les conditions de la mise en place d'un lien psychologique qui modulera les comportements ultérieurs d'un individu via les indices identitaires associés à son apparence. Parallèlement, il apparait que l'influence des indices identitaires est renforcée en situation de déindividuation, maximisant de fait leur impact sur l'auto-perception et donc sur la modulation des attitudes et des comportements. Dès lors, on distingue le pouvoir identitaire que les avatars peuvent exercer sur les utilisateurs qui les incarnent.

3 L'Effet Proteus : Effet de l'avatar

Selon Yee et Bailenson [40], les environnements virtuels qui rendent les interlocuteurs anonymes peuvent être vus comme des versions numériques de la chambre noire de l'expérience « *Deviance in the dark* » évoquée plus haut [14]. Il s'agit en effet d'espaces propices à la déindividuation en raison de l'anonymat et de l'isolement physique des différents utilisateurs. En outre, dans ces environnements, l'avatar n'est pas un simple costume mais une « représentation de soi pleine et entière » [40]. En d'autres termes, le costume est un indice identitaire parmi d'autres, mais l'avatar constitue le premier indice identitaire dans les environnements virtuels. De fait, les individus déindividués dans ces environnements devraient être particulièrement sensibles aux indices sociaux associés à la nouvelle identité qu'ils infèrent à partir de leur avatar. De la même façon que les individus en uniformes noirs se conforment à une identité plus agressive [13], les utilisateurs des environnements virtuels se conformeront aux attentes qui découlent du prototype identitaire auquel renvoie l'avatar. Sous l'action de sa propre image, de la représentation de soi virtuelle, l'individu va donc s'auto-influencer et rationnaliser ses comportements dans le sens de l'identité constituée par l'avatar. Cette modulation comportementale issue de l'apparence de l'avatar est appelée effet Proteus (du nom du Dieu de la mythologie grecque qui possédait la faculté de métamorphose). En accord avec la théorie de l'auto-perception, ce phénomène peut apparaitre même lorsque l'individu est seul. L'utilisateur peut en effet agir « à la troisième personne », comme le ferait un observateur extérieur. S'il est trivial de rappeler que l'utilisateur exerce un contrôle et une influence directe sur l'avatar qu'il incarne, ce dernier peut aussi infléchir les comportements et les attitudes de l'utilisateur. De fait, on peut considérer que « la relation utilisateur/avatar s'initie de façon circulaire et conduit à l'élaboration d'une identité spécifique » [16].

L'effet Proteus a été testé expérimentalement à travers différentes études. Dans une première étude de Yee et Bailenson [40] (Figure 1), les participants incarnaient un avatar (attractif vs. moyennement attractif vs. non attractif) et entraient en contact avec un interlocuteur de sexe opposé. Durant l'expérience, le participant ne voyait jamais la réelle apparence de l'interlocuteur et réciproquement (un rideau noir séparait la salle).

Figure 1. Exemple d'avatar utilisé par Yee & Bailenson [40]

Les résultats mettent en évidence que les participants de la condition attractive diminuent la distance interpersonnelle et révèlent significativement plus d'informations sur eux-mêmes que les participants incarnant un avatar non-attractif. Autrement-dit, le fait d'incarner un avatar d'apparence attractive conduit les participants à se montrer plus intimes dans leurs interactions sociales. Il est à noter que ce phénomène résulte de la simple exposition à un miroir virtuel permettant au participant d'observer son avatar durant environ une minute. Puisque les participants percevaient l'environnement en vue subjective (i.e., à travers les yeux de l'avatar), l'avatar n'était plus visible après cette phase d'exposition au miroir. Les auteurs considèrent donc que l'effet Proteus s'initie quasi-instantanément.

Dans une seconde étude, les auteurs ont manipulé la taille de l'avatar, un facteur identifié comme relevant davantage de l'estime de soi et de la compétence que de l'attractivité [43]. Les participants (hommes vs. femmes) incarnaient un avatar (grand vs. petit vs. de même taille que l'interlocuteur) et entraient en contact avec un interlocuteur du sexe opposé. L'interlocuteur était en réalité un expérimentateur qui se comportait de la même façon dans toutes les conditions expérimentales. L'étude reposait sur une tâche de négociation dans l'environnement virtuel. Selon cette tâche (i.e., « *Ultimatum Game* » [9]), deux interlocuteurs décident à tour de rôle de la façon de répartir une somme d'argent entre eux. L'un des interlocuteurs opère la répartition et l'autre accepte ou rejette cette décision. En cas d'acceptation l'argent est attribué à chacun selon les termes de la répartition, en cas de rejet personne n'obtient d'argent. Les résultats révèlent que les avatars de grande taille amènent les individus à se montrer plus confiants et déloyaux dans une négociation que les avatars de petite taille. Il apparait que les participants incarnant des avatars de grande taille avaient le plus tendance à proposer des répartitions inéquitables. De façon corollaire, les

participants incarnant des avatars de petite taille avaient significativement plus tendance à accepter une répartition inéquitable que les participants des conditions grande taille et taille moyenne.

Cette étude a ensuite été répliquée afin d'observer si la modulation comportementale initiée dans l'environnement virtuel pouvait perdurer lors d'un échange en présentiel [42]. Si la première partie de l'expérience était similaire, la seconde partie amenait les participants à interagir en face-à-face. Les deux interlocuteurs réalisaient donc une nouvelle fois la tâche de négociation. Les résultats reproduisent le phénomène mis en évidence par Yee et Bailenson [40]. Mais il apparait également que les participants qui incarnaient un avatar de grande taille dans l'environnement virtuel sont aussi ceux qui se montrent le plus déloyaux lors de la négociation en présentiel. Ainsi, les effets des indices identitaires sur l'auto-perception se maintiennent à l'issue de la situation de déindividuation en environnement virtuel. Il semble enfin important de souligner que les participants n'ont pas conscience de l'influence exercée par l'apparence de leur avatar et que l'effet Proteus s'initie de manière relativement implicite.

Cette première série d'expériences comporte toutefois un certain nombre de limites. Dans ces études, les participants intègrent en effet l'environnement virtuel via un dispositif d'immersion (visiocasque à détection de mouvement) apportant une perception en vue subjective. Bien que ces dispositifs d'immersion connaissent un essor considérable, ils restent éloignés de la réalité actuelle des échanges en ligne où les utilisateurs perçoivent le plus souvent leurs avatars à la troisième personne (vue de dos) et par le biais d'un écran d'ordinateur. Des études plus récentes ont permis d'observer l'influence des avatars, avec des résultats sensiblement similaires mais dans des situations plus usuelles [27]. En outre, Yee et al. [42] ont réalisé une recherche quasi-expérimentale étudiant une population de joueurs du jeu de rôle en ligne World of Warcraft (Figure 2). Les auteurs ont recensé plusieurs dizaines de milliers d'avatars au moyen d'un script automatique. L'attractivité des différents types d'avatars disponibles dans le jeu a d'abord été évaluée par une population de juges. La taille des avatars en fonction de leur race (humain, troll, gnome, etc.) a également été mesurée.

Les résultats révèlent que les avatars attractifs de grande taille constituent les avatars les plus puissants (niveaux d'expérience au sein du jeu les plus élevés). Les auteurs en concluent que le choix de l'avatar détermine la manière de jouer. Il est à noter que l'attractivité, la taille ou plus largement l'apparence de l'avatar n'impliquent aucun bénéfice fonctionnel pour le joueur. Par exemple, les avatars de grande taille ne sont pas plus puissants ou ne se déplacent pas plus rapidement que les avatars de petite taille.

Figure 2. Différents avatars de World of Warcraft (repris de [42])

Les études présentées ont permis d'apprécier l'influence que l'avatar peut exercer sur les comportements, mais les avatars peuvent également impacter les attitudes des utilisateurs. L'étude de Fox, Bailenson et Tricaze [12] fournit un bon exemple de ce phénomène. Cette expérience examine l'influence que le caractère plus ou moins sexualisé et suggestif d'un avatar féminin peut exercer sur les utilisatrices qui l'incarnent. En particulier, les auteurs ont manipulé deux facteurs, la tenue (sexualisée vs. non-sexualisée) et le visage de l'avatar (ressemblant à soi vs. ne ressemblant pas à soi). Durant l'expérience, chaque participante entrait en interaction avec un avatar masculin. A l'issue de l'interaction, les auteurs mesuraient les pensées liées au physique, c'est-à-dire les éléments d'*auto-objectification* impliquant de considérer les femmes comme des objets réduits uniquement à des attributs sexuels. Les résultats révèlent que les participantes incarnant un avatar sexualisé expriment davantage d'éléments d'auto-objectification que les participantes incarnant un avatar non-sexualisé. Les résultats de cette expérience soulignent l'influence que l'effet Proteus peut exercer au niveau attitudinal. Au regard des travaux antérieurs, il est également possible de supposer que l'image fortement sexualisée de la femme – notamment véhiculée via certaines héroïnes de jeu vidéo incarnées par les joueurs/joueuses – peut faciliter le développement d'attitudes négatives à l'égard des femmes, même au-delà du cadre des environnements virtuels.

L'étude de Peña et al. [27] fournit une autre illustration de l'influence des avatars sur les attitudes. Cette étude (inspirée de [19]) amenait les participants à incarner un avatar représentant un docteur, un membre du KKK, ou un corps transparent (i.e., avatar de la condition contrôle n'impliquant aucun stéréotype). Cette expérience n'utilisait pas la vue subjective, l'avatar étant perçu à la

troisième personne. Les participants pouvaient pivoter la caméra et utiliser un miroir pour obtenir une vue complète de l'avatar. Durant l'expérience, les participants se déplaçaient dans un musée virtuel et devaient inventer des histoires sur la base des planches exposées. Il apparait que les participants incarnant un avatar du KKK, dont les attributs stéréotypés sont clairement négatifs, imaginaient les histoires les plus négativement connotées (i.e., meurtre, vengeance, crime et mépris). Les participants de la condition KKK composaient également des histoires comportant moins d'éléments positifs que les participants de la condition docteurs et de la condition contrôle. Ainsi, selon Peña et al., en plus de maximiser l'apparition d'éléments négatifs, « le fait d'utiliser des avatars liés à des associations agressives inhibe davantage les pensées positives » [27].

Enfin, sur un plan plus général, il semble que l'effet Proteus puisse s'initier quel que soit le réalisme, la qualité graphique ou le degré de sophistication comportementale de l'avatar : les études citées plus haut ont été réalisées au sein de différents types d'environnements virtuels, utilisant différents moteurs graphiques et avec différents dispositifs d'immersion. Il ne semble donc pas nécessaire de disposer d'un haut niveau de réalisme pour engendrer les effets attitudinaux et comportementaux : ce sont les indices identitaires qui déterminent les modulations comportementales et attitudinales et non la qualité du rendu.

4 De l'avatar à l'agent conversationnel

Plusieurs processus liés à l'apparence de l'avatar sont susceptibles d'aboutir à des changements comportementaux mais ne relèvent pas de l'effet Proteus tel qu'il a été initialement formalisé. Il convient donc d'établir plusieurs distinctions conceptuelles. En outre, l'analyse de ces processus nous permet d'inférer un certain nombre d'effets potentiels des agents conversationnels.

Ainsi, l'objectif de cette partie est de développer ces processus sociocognitifs communs aux avatars (représentation de soi) et aux représentations d'autrui (agents conversationnels autonomes ou avatars d'autres utilisateurs). Cette analyse nous permettra ensuite de suggérer en perspectives des pistes de recherche dans le domaine des agents conversationnels.

4.1 Confirmation comportementale

Le premier processus, *la confirmation comportementale* concerne l'influence qu'une personne (i.e., observateur) peut avoir sur une autre (i.e., cible). Dans ce système, les comportements de la cible peuvent être modulés de manière à confirmer les attentes de l'observateur (voir notamment *effet pygmalion* et *prophétie auto-réalisatrice* [24, 30]). Par exemple, dans une étude de Snyder, Tanke et Berscheid [35], des étudiants de sexes opposés étaient placés dans une situation de communication téléphonique. Les résultats révèlent que les hommes qui entraient en contact avec une femme en considérant cette cible comme attirante amenaient leur interlocutrice à se comporter d'une façon sensiblement plus amicale et charmante. Dans un environnement virtuel, un observateur entrant en interaction avec une cible incarnant un avatar attractif pourrait, de la même façon, amener cette dernière à se comporter de manière plus sympathique. Il est à noter que dans l'optique de la confirmation comportementale, la source du changement de comportement dépend bien plus largement de l'observateur que de la cible elle-même. En effet, les comportements et les attentes de l'observateur initient les modulations au niveau du comportement de la cible.

A l'échelle de la société, on trouve un type de fonctionnement apparenté dans le phénomène de *menace du stéréotype* [37]. En effet, les croyances stéréotypées et assimilatrices associées à certains groupes sociaux peuvent générer des attentes exerçant une influence sur la performance des membres de ces groupes. Dans cette perspective, la menace du stéréotype pourrait s'initier sur le support de l'apparence des personnages virtuels. Plus largement, il a en effet été montré que les stéréotypes ethniques et les biais raciaux maintiennent leur influence délétère lors d'interactions en environnement virtuel [10].

Afin de dissocier effet Proteus et confirmation comportementale, plusieurs études ont contrôlé la perception que l'interlocuteur avait de l'avatar. Aussi, dans l'étude princeps sur l'attractivité de l'avatar [40], l'interlocuteur ne percevait pas les caractéristiques du visage de l'avatar, il percevait un visage humain non texturé en noir et blanc. Il en va de même concernant la taille de l'avatar, l'avatar du participant conservant la même taille aux yeux de l'interlocuteur quelle que soit la condition. L'effet Proteus constitue donc bel et bien un processus autonome. Cependant, il semble évident que ce processus peut être combiné à la confirmation comportementale pour maximiser encore davantage l'impact des représentations digitales de soi. Un plan complet tenant compte de l'influence de l'observateur semble donc essentiel pour appréhender le poids respectif de ces processus sur les comportements (e.g., avatar neutre pour le participant, mais attrayant pour l'observateur, etc.). Dans cette optique, il est aussi possible de placer l'effet Proteus et la confirmation comportementale en concurrence (e.g., avatar perçu comme attractif par l'individu mais non-attractif par son (ses) interlocuteur(s), et inversement). De nombreuses recherches seraient encore nécessaires pour caractériser ces effets respectifs sur le comportement de l'utilisateur.

4.2 Effet d'amorçage

Le deuxième processus concerne l'*effet d'amorçage* et *l'activation automatique des stéréotypes*. La théorie de l'auto-perception constitue le principal support de l'effet Proteus, mais l'influence des avatars pourrait aussi être expliquée par l'action implicite des stéréotypes sur les comportements. Selon Bargh, Chen, et Burrows [3], l'amorçage correspond « à l'activation de structures de connaissances, comme des concepts ou des stéréotypes, dans le contexte situationnel ». Or, les effets de

l'amorçage peuvent avoir une incidence sur les perceptions sociales [17], mais aussi sur les comportements des individus et leurs interactions avec autrui. En effet, puisque les cognitions sont organisées dans une structure de connaissances, l'activation d'un certain concept/stéréotype (e.g., KKK) peut activer les réseaux d'informations associés (e.g., racisme, agression, violence) influençant l'individu de façon implicite sur le plan comportemental. Ce phénomène aurait pour effet d'activer les réseaux proches du concept amorcé, et d'inhiber les réseaux lointains (e.g., la gentillesse dans le cas de l'amorçage du KKK). Par exemple, Bargh, Chen, et Burrows [3] ont démontré que la présentation de mots liés à la vieillesse (i.e., amorçage du stéréotype de la personne âgée) conduisait des participants à marcher significativement moins rapidement que des participants exposés à des mots neutres. Dans une autre étude, des participants exposés à l'amorce visant à activer le stéréotype du « professeur » obtenaient de meilleures performances dans une tâche de culture générale que les participants exposés à l'amorce « hooligan » [8].

Il en résulte que si l'effet Proteus relève d'un effet d'amorçage, il serait dû au fait d'être exposé à l'avatar et donc aux concepts que son apparence est susceptible d'activer. Selon Peña et al. [27], l'amorçage serait d'ailleurs le principal mécanisme qui sous-tend l'effet Proteus. Par exemple, le fait de voir un avatar attractif (son propre avatar) peut conduire les participants à se comporter de manière plus sympathique en raison d'une association stéréotypée entre les individus attractifs et les comportements amicaux [9]. L'effet d'amorçage peut donc également expliquer les effets des stéréotypes ethniques sur le comportement des individus [3], que ceux-ci incarnent ou observent le personnage portant ce stéréotype. De fait, il est important de noter que l'effet d'amorçage met l'avatar et l'agent conversationnel au même niveau en termes d'influence sur le comportement de l'utilisateur.

4.3 Auto-perception vs. amorçage

Les explications de l'influence des avatars en termes d'auto-perception ou d'amorçage font aujourd'hui débat dans la littérature. Dans des proportions équivalentes, ces explications alternatives sont en effet susceptibles de rendre compte des modulations comportementales exercées par les avatars. Ainsi, plusieurs auteurs interprètent ces modulations comme relevant de l'action du processus d'amorçage [27, 28] tandis que d'autres mobilisent l'effet Proteus et donc la théorie de l'auto-perception [40, 42]. La différence est pourtant importante car dans le cas de l'auto-perception, ces effets seraient spécifiques aux avatars (représentations de soi, impliquant un mécanisme d'incarnation), alors que dans le cas de l'amorçage ils seraient transférables aux agents conversationnels animés, autonomes ou pas, ainsi qu'aux robots.

A notre connaissance, seule une étude a tenté de dissocier ces processus [41]. Dans cette expérience, les participants incarnaient un avatar ou observaient ce même personnage sans en avoir le contrôle. Les résultats révèlent que les changements comportementaux observés dans les travaux antérieurs (sur la base de l'attractivité du personnage) sont plus importants lorsque l'individu incarne l'avatar que lors de la présentation « désincarnée » de ce même stimulus. Aussi, si l'amorçage fournit des pistes d'explication convaincantes, l'influence des avatars n'est pas réductible à la seule intervention de ce processus. On peut donc considérer que l'amorçage rend compte de l'influence globale de l'environnement [26] et des situations où l'individu est exposé à un personnage virtuel qu'il n'incarne pas, et que les effets comportementaux sont intensifiés par le mécanisme de l'auto-perception, dans le cas où l'individu incarne le personnage.

5 Conclusion et perspectives

Les environnements virtuels offrent la possibilité de vivre de nouveaux types d'interactions sociales, caractérisées notamment par la déindividuation et la capacité à jouer avec l'image de soi. Cette revue de question visait à fournir une vue d'ensemble de l'influence que l'apparence des personnages virtuels est susceptible d'exercer sur les utilisateurs, en termes comportemental et attitudinal. Les différentes propositions théoriques et les illustrations empiriques présentées permettent d'appréhender la nature multidirectionnelle des liens unissant les personnages virtuels aux utilisateurs : **(1)** l'apparence de l'avatar peut exercer une influence directe sur l'utilisateur via l'effet Proteus, l'auto-perception et/ou le processus d'amorçage, **(2)** elle peut aussi exercer une influence indirecte via les attentes générées chez d'autres utilisateurs (confirmation comportementale et/ou amorçage) et **(3)** même si le personnage virtuel n'est pas incarné par l'utilisateur, le fait d'y être exposé peut activer des concepts susceptibles d'initier des modulations comportementales (confirmation comportementale et/ou amorçage). Notons cependant qu'à l'exception de l'étude de Yee et al. [42], abordant l'influence des avatars dans World of Warcraft, les différentes expériences réalisées ne permettaient pas aux participants de choisir et/ou de configurer leur avatar. En outre, ce domaine de recherche a focalisé son analyse sur les situations de première rencontre entre l'utilisateur et l'avatar / l'agent. L'influence des personnages virtuels lors d'interactions qui s'inscrivent dans la durée et son évolution dans le temps restent donc à étudier.

Au-delà de ces sources de modulation qui sont aujourd'hui étayées d'un point de vue expérimental, il est possible d'envisager plusieurs pistes dont l'intérêt dans le domaine des agents conversationnels ne doit pas en occulter le caractère plus ou moins spéculatif.

Une première piste de recherche consisterait à étudier les effets combinés de l'apparence des représentations de soi (avatars) et d'autrui (agent conversationnel ou autre utilisateur) en situation d'interaction. Selon les cas, l'interaction pourrait aboutir à des effets additifs ou à des interférences. Par exemple, les effets d'un avatar et d'un agent attractifs pourraient s'additionner : un

avatar attractif pousserait à un comportement plus intime, et ce d'autant plus que l'interlocuteur serait attractif lui aussi. Ceci étant, il est également envisageable que le fait d'interagir avec un personnage virtuel moins attractif que l'avatar de l'utilisateur produise un effet de contraste propice aux modulations comportementales. On peut aussi se demander si les effets liés à la couleur du costume ou à l'activation d'un stéréotype peuvent être amplifiés ou atténués dans un contexte d'interaction en groupe réunissant des avatars portant le même costume et/ou susceptibles d'activer les mêmes stéréotypes. Ces considérations dépassent le périmètre de cette revue de question puisqu'elles mobilisent d'autres champs comme les théories sociocognitives du groupe en termes d'identité sociale. Il nous faut toutefois noter que plusieurs travaux révèlent que la similarité des avatars peut être un moyen de stimuler l'identification au groupe [20, 21] et donc l'influence potentielle des facteurs groupaux lors d'interactions en environnement virtuel [28].

Par ailleurs, il serait intéressant d'examiner si un agent conversationnel est capable d'initier des effets de confirmation comportementale. Dans un environnement virtuel, un agent conversationnel, jouant le rôle d'observateur par rapport à l'utilisateur cible, pourrait influencer les comportements de celui-ci : par exemple, l'amener à se comporter de manière plus ou moins sympathique, ou l'amener à obtenir de meilleures performances – ce qui reviendrait à reproduire, en environnement virtuel, l'effet Pygmalion [30]. Si la manière dont l'observateur, dans la réalité, influence le comportement de la cible n'est pas modélisée en détail, il serait possible d'approfondir cette question par les méthodes d'analyse de corpus utilisées dans la conception d'agents conversationnels animés [5]. L'objectif serait de repérer des constantes comportementales (e.g., proximité physique, fréquence et durée des contacts visuels, intonation de la voix, vocabulaire employé, etc.) susceptibles d'expliquer ce phénomène, puis d'injecter celles-ci dans le modèle de l'agent pour qu'il puisse devenir un vecteur de confirmation comportementale.

Enfin, il conviendrait de tester les potentialités du processus d'amorçage pour améliorer l'interaction entre agent conversationnel et utilisateur. Cela impliquerait de rechercher en premier lieu un stéréotype positif à activer en fonction de la situation d'interaction et de ses objectifs, puis d'attribuer à l'agent une apparence permettant d'activer ce stéréotype, et enfin d'en mesurer l'effet sur le comportement de l'utilisateur. Le processus d'amorçage pourrait ainsi permettre de progresser dans la recherche de *crédibilité*, problématique majeure du domaine des agents conversationnels [2, 18]. A cet égard, il serait également intéressant de croiser l'effet Proteus ou l'amorçage avec le phénomène d'Uncanny valley [25], afin de vérifier s'il s'agit de processus indépendants. L'Uncanny valley évoque l'influence de l'apparence des personnages virtuels ou des robots sur le sentiment de familiarité ou d'étrangeté éprouvé par l'utilisateur. Ce phénomène impacte principalement les attitudes ; il existe peu de données empiriques attestant de son influence sur les comportements [15]. Pour examiner si l'Uncanny valley interfère avec l'effet Proteus ou l'amorçage, il faudrait utiliser des personnages porteurs d'indices identitaires (e.g., costume, rôle social) mais également pourvus de caractéristiques anormales (e.g., des yeux surdimensionnés ou des yeux humains sur un visage artificiel [33]). Il s'agirait alors d'observer si les modulations comportementales se maintiennent ou sont perturbées par l'étrangeté des représentations digitales. Dans une perspective proche de l'influence des avatars non-attractifs, nous serions tentés de supposer que ces modulations se maintiennent en dépit de l'étrangeté des représentations.

Dans tous les cas, les processus sociocognitifs présentés dans cette revue de question semblent propres à inspirer de nouvelles recherches dans le domaine des agents conversationnels animés, qu'il s'agisse d'en améliorer la conception ou de mieux comprendre le comportement humain face à ceux-ci.

Remerciements

Cette synthèse a été réalisée dans le cadre du projet ANR CREATIVENESS (Creative Activities in Virtual Environmental Spaces, 2013-2016).

Bibliographie

1. Adams, F.M. and Osgood, C.E. *A cross-cultural study of the affective meanings of color.* Journal of Cross-Cultural Psychology, 4, 1973, pp. 135-156.
2. Ball, G. and Breese, J. Emotion and personality in a conversational agent. In J. Cassell, et al. (Ed.) *Embodied conversational agents*, MIT Press. 2000, pp. 189-219.
3. Bargh, J.A., Chen, M., and Burrows, L. *Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action.* Journal of personality and social psychology, 71, 1996, pp. 230-244.
4. Bem, D. Self perception theory. In L. Berkowitz (Ed.) *Advances in experimental social psychology*, Academic Press, New York. 1972, pp. 1-62.
5. Buisine, S., Abrilian, S., Niewiadomski, R., Martin, J.C., Devillers, L., and Pelachaud, C. Perception of blended emotions: From video corpus to expressive agent. In *Proceedings of IVA'06 International Conference on Intelligent Virtual Agents*, Lecture Notes in Computer Science, Springer, 2006, pp. 93-106.
6. Diener, E., Fraser, S.C., Beaman, A.L., and Kelem, R.T. *Effects of deindividuation variables among Halloween trick-or-treaters.* Journal of Personality and Social Psychology, 33, 1976, pp. 178-183.
7. Diener, E. Deindividuation: The absence of self-awareness and self-regulation in group members. In P.B. Paulus (Ed.) *Psychology of group influence*, Erlbaum, Hillsdale, NJ. 1980, pp. 202-242.
8. Dijksterhuis, A. and Van Knippenberg, A. *The relation between perception and behavior, or how to win a game of trivial pursuit.* Journal of personality and social psychology, 74, 1998, pp. 865-877.
9. Dion, K., Berscheid, E., and Walster, E. *What is beautiful is good.* Journal of personality and social psychology, 24, 1972, pp. 285-290.
10. Eastwick, P.W. and Gardner, W.L. *Is it a game? Evidence for social influence in the virtual world.* Social Influence, 4, 2009, pp. 18-32.

11. Festinger, L., Pepitone, A., and Newcomb, T. *Some consequences of de-individuation in a group.* Journal of Abnormal and Social Psychology, 47, 1952, pp. 382-389.
12. Fox, J., Bailenson, J.N., and Tricase, L. *The embodiment of sexualized virtual selves: The Proteus effect and experiences of self-objectification via avatars.* Computers in Human Behavior, 29, 2013, pp. 930-938.
13. Frank, M. and Gilovich, T. *The dark side of self and social perception: Black uniforms and aggression in professional sports.* Journal of Personality and Social Psychology, 54, 1988, pp. 74-85.
14. Gergen, K., Gergen, M., and Barton, W.H. *Deviance in the dark.* Psychology Today, 11, 1973, pp. 129-130.
15. Groom, V., Nass, C., Chen, T., Nielsen, A., Scarborough, J.K., and Robles, E. *Evaluating the effects of behavioral realism in embodied agents.* International Journal of Human-Computer Studies, 67, 2009, pp. 842-849.
16. Guegan, J. and Michinov, E. *Communication via Internet et dynamiques identitaires: une analyse psychosociale.* Psychologie Française, 56, 2011, pp. 223-238.
17. Higgins, E., Rholes, W.S., and Jones, C.R. *Category accessibility and impression formation.* Journal of Experimental Social Psychology, 13, 1977, pp. 141-154.
18. Isbister, K. and Doyle, P. The blind men and the elephant revisited. In Z. Ruttkay and C. Pelachaud (Ed.) *From brows to trust: Evaluating Embodied Conversational Agents.*, Kluwer Academic Publishers. 2004, pp. 3-26.
19. Johnson, R. and Downing, L. *Deindividuation and valence of cues: Effects on prosocial and antisocial behavior.* Journal of Personality and Social Psychology, 37, 1979, pp. 1532-1538.
20. Kim, J. *Two Routes Leading to Conformity Intention in Computer-Mediated Groups: Matching Versus Mismatching Virtual Representations.* Journal of Computer-Mediated Communication, 16, 2011, pp. 271-287.
21. Lee, E.-J. *Effects of Visual Representation on Social Influence in Computer-Mediated Communication.* Human Communication Research, 30, 2004, pp. 234-259.
22. Lord, C.G. *Social psychology*. Fort Worth, TX: Harcourt Brace, 1997.
23. Meier, B.P., Robinson, M.D., and Clore, G.L. *Why Good Guys Wear White Automatic Inferences About Stimulus Valence Based on Brightness.* Psychological Science, 15, 2004, pp. 82-87.
24. Merton, R.K. *The self-fulfilling prophecy.* The Antioch Review, 8, 1948, pp. 193-210.
25. Mori, M. *The uncanny valley.* Energy, 7, 1970, pp. 33-35.
26. Peña, J. and Blackburn, K. *The priming effect of virtual environments on interpersonal perceptions and behaviors.* Journal of Communication, 63, 2013, pp. 703-720.
27. Peña, J., Hancock, J., and Merola, N. *The Priming Effects of Avatars in Virtual Settings.* Communication Research, 36, 2009, pp. 838-856.
28. Peña, J. *Integrating the influence of perceiving and operating avatars under the automaticity model of priming effects.* Communication Theory, 21, 2011, pp. 150-168.
29. Postmes, T. and Spears, R. *Deindividuation and anti-normative behavior: A meta-analysis.* Psychological Bulletin, 123, 1998, pp. 238-259.
30. Rosenthal, R. and Jacobson, E. *Pygmalion à l'école. L'attente du maître et le développement intellectuel des élèves*. Paris: Casterman, 1971.
31. Sabini, J. *Social psychology*. New York: Norton, 1995.
32. Schachter, S. and Singer, J. *Cognitive, social, and physiological determinants of emotional state.* Psychological review, 69, 1962, pp. 379-399.
33. Seyama, J.I. and Nagayama, R.S. *The uncanny valley: Effect of realism on the impression of artificial human faces.* Presence: Teleoperators and Virtual Environments, 16, 2007, pp. 337-351.
34. Singer, J.E., Brush, C.A., and Lublin, S.C. *Some aspects of deindividuation: Identification and conformity.* Journal of Experimental Social Psychology, 1, 1965, pp. 356-378.
35. Snyder, M., Tanke, E.D., and Berscheid, E. *Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes.* Journal of Personality & Social Psychology, 35, 1977, pp. 656-666.
36. Spivey, C.B. and Prentice-Dunn, S. *Assessing the Directionality of Deindividuated behavior: Effects of Deindividuation, Modeling, and Private Self-Consciousness on Aggressive ans Prosocial Responses.* Basic and Applied Social Psychology, 11, 1990, pp. 387-403.
37. Steele, C.M. and Aronson, J. *Stereotype threat and the intellectual test performance of African Americans.* Journal of personality and social psychology, 69, 1995, pp. 797-811.
38. Turkle, S. *Life on the screen: Identity in the age of the internet*. New York: Touchstone, 1995.
39. Valins, S. and. *Cognitive effects of false heart-rate feedback.* Journal of Personality and Social Psychology, 4, 1966, pp. 400-408.
40. Yee, N. and Bailenson, J. *The Proteus effect: The effect of transformed self-representation on behavior.* Human Communication Research, 33, 2007, pp. 271-290.
41. Yee, N. and Bailenson, J.N. *The difference between being and seeing: the relative contribution of self-perception and priming to behavioral changes via digital self-representation.* Media Psychology, 12, 2009, pp. 195-209.
42. Yee, N., Bailenson, J.N., and Ducheneaut, N. *The Proteus effect. Implications of transformed digital self-representation on online and offline behavior.* Communication Research, 36, 2009, pp. 285-312.
43. Young, T.J. and French, L.A. *Height and perceived competence of U.S. Presidents.* Perceptual and Motor Skills, 82, 1996, pp. 1002.
44. Zimbardo, P.G. The human choice: Individuation, reason, and order vs. deindividuation, impulse, and chaos. In W.J. Arnold and D. Levine (Ed.) *Nebraska Symposium on Motivation*, University of Nebraska Press, Lincoln. 1969, pp. 237-307.

Animation Creation For Virtual Agent System

J. Huang[1] C. Pelachaud[1]

[1] CNRS/Telecom ParisTech/LTCI

jing.huang, catherine.pelachaud@telecom-paristech.fr

Abstract

In this work, we present how we create our animation using our key framing system. Hybrid extension of frames generators can be integrated to achieve expressive gesture animations. This system offers more flexibility in the sense of generating expressive animation sequences for key frame system.

Keywords

Virtual Agent, Animation Synthesis.

1 Introduction

Embodied Conversational Agents are virtual human agents that can communicate through voice, facial expressions, emotional gestures, body movements etc. They use their verbal and nonverbal behaviours to convey their intentions and emotional states. It is necessary that ECAs can display a large variety of behaviours. Our system defines a low level parametrization derived from psychology literature. The EMOTE system [1] uses a similar representation of expressive controls, but abstracted from Laban principles (1980). In the realization of animation, both of these works decomposed character skeleton into small parts (the head, the torso, the arms, etc) and solved the system by different controllers acting locally. Such an approach does not allow modeling motion propagations, ie how motion over one modality may affect another one. In our work, we choose to deal with whole body motion with a global view. Realistic real-time animation generation is always a challenge. The process needs to be simple, easy to implement and flexible to control. The multi-modalities (solve each module independently) is a solution to most real-time virtual agent(VA) systems. Kinematics is a general method for manipulating interactively articulated figures and creating postures. Such as inverse kinematics [2] [3] [4], it solves the computation problem for end-side users, graphics designers, psychology scientists. Previous work [5] focuses on the realization of animation for virtual conversational agents. The model takes as input a sequence of multimodal behaviours to generate, with symbolic representations in McNeill's sectors [6], outputs the composed keyframes using normal or lazy approaches. The inverse kinematics [7] is used to convert the spatial parameters from cartesian space into generalized space. Different expressivity parameters are introduced to configure the timing and the physical space for key frame gestures. Expressive variations are also introduced by different easing in/out functions to variate the transitions. The different interpolation functions are also used, such as Kochanek Bartels splines (TCB splines) [8]. Michael Neff [9] [10] [11] also inspires us a lot for generating expressive animation for the transitions. He presented their aesthetic motion generation system. Their model starts from a high level expressive language that is translated into precise semantic units that can be simulated by physical or kinematics methods.

2 Methodology

We present here, our extensible animation pipeline based on the SAIBA [12] framework. Our system is designed as in Figure 1. We take inputs of symbolic representations such as animation references on high level, then spatial positions, touch-points as signals in symbolic key frame level, etc. Some references are used for data driven methods, such as laughter synthesis based on phoneme. The key frame generator synchronizes the various key points of multi-modalities into animation key frames. The inverse kinematics will be applied to computing full body gestures, involving posture expressivity parameters. The transitions of animation is especially important for generating natural motions. We separate into two parts: one is using interpolation functions which is computational cheap; the other is using dynamic motion, involving gravity and external forces, which is underworking now. The routine can be chosen in real-time by our modular system. We use two springs antagonistic formulation to build joint motor [9].

$$\tau = K_L(\theta_L - \theta) + K_H(\theta_H - \theta)$$

where τ is the torque, θ_L and θ_H are the boundary of the joint, K_L and K_H are gains, can be used to model tension through a whole transition. Later, featherstone [13] will be applied to achieve frames updating. All key gesture animations are predefined by symbolic references in our library. We show here our new Gesture Editor which uses a Java Graphics Interface to edit different gestures and save into

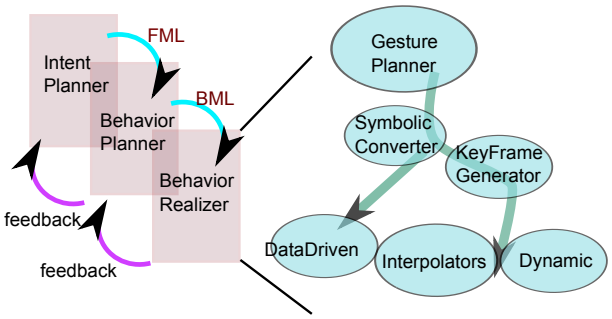


Figure 1: Our system is based on the standard SAIBA framework defines the modularity, functionality and the protocols for ECAs. In the Behaviour Realizer module, the parts of generating gestures correspond to our work in the animation pipeline.

our gestures library. The whole framework is still under working. We will continue to extend our application with more features that supports more in generating realistic motion for communicative VA systems.

References

[1] Chi, D., Costa, M., Zhao, L., Badler, N.: The EMOTE model for effort and shape. In: In Proceedings of SIGGRAPH 2000. SIGGRAPH '00, New York, NY, USA (2000) 173–182

[2] Baerlocher, P., Boulic, R.: An inverse kinematic architecture enforcing an arbitrary number of strict priority levels. The Visual Computer **20** (2004) 402–417

[3] Boulic, R., Thalmann, D.: Combined direct and inverse kinematic control for articulated figure motion editing (1992)

[4] Hecker, C., Raabe, B., Enslow, R.W., DeWeese, J., Maynard, J., van Prooijen, K.: Real-time motion retargeting to highly varied user-created morphologies. ACM Trans. Graph. **27**(3) (August 2008) 27:1–27:11

[5] Huang, J., Pelachaud, C.: Expressive body animation pipeline for virtual agent. In Nakano, Y., Neff, M., Paiva, A., Walker, M., eds.: Intelligent Virtual Agents, 12th International Conference on Intelligent Virtual Agents. Volume 7502 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 355–362

[6] McNeill: Hand and Mind: WHAT GESTURES REVEAL ABOUT THOUGHT. The University of Chicago press, Chicago (1992)

[7] Huang, J., Pelachaud, C.: An efficient energy transfer inverse kinematics solution. In: Proceedings of Motion In Game 2012. Volume 7660., Berlin, Heidelberg, LNCS (2012) 278–289

Figure 2: Gesture Editor can create a geste by defining its gesture phases, arm sptial position, orientation, hand shape, etc, and test with different expressive parameters.

[8] Kochanek, D.H.U., Bartels, R.H.: Interpolating splines with local tension, continuity, and bias control. SIGGRAPH (January 1984)

[9] Neff, M., Fiume, E.: Modeling tension and relaxation for computer animation. In: Proceedings of Symposium on Computer Animation 2002. SCA '02, New York, NY, USA, ACM (2002) 81–88

[10] Neff, M., Fiume, E.: Artistically based computer generation of expressive motion. In: In Proceedings of the Adaptation in Artificial and Biological Systems. (2004) 29–39

[11] Neff, M., Fiume, E.: AER: aesthetic exploration and refinement for expressive character animation. In: Proceedings of Symposium on Computer Animation 2005. SCA '05, New York, NY, USA, ACM (2005) 161–170

[12] Krenn, B., Marsella, S., Marshall, A.N., Pirker, H., ThÂÍÂőrisson, K.R., VilhjÂÍÂćlmsson, H.: Towards a common framework for multimodal generation in ecas: The behavior markup language. In: In Proceedings of the 6th International Conference on Intelligent Virtual Agents, Marina. (2006) 21–23

[13] Featherstone, R., Orin, D.E.: Robot dynamics: Equations and algorithms. In: ICRA. (2000) 826–834

Les Styles pour la Plasticité des Robots Compagnons

Wafa Johal
Univ. Grenoble Alpes, LIG
Wafa.Johal@imag.fr

Sylvie Pesty
Univ. Grenoble Alpes, LIG
Sylvie.Pesty@imag.fr

Gaëlle Calvary
Univ. Grenoble Alpes, LIG
Gaëlle.Calvary@imag.fr

Résumé

Dans ce papier, nous présentons les résultats d'une étude visant l'évaluation de l'expressivité de styles parentaux par des robots compagnons, ainsi que de leur crédibilité par des parents. 93 parents ont participé à cette enquête en ligne, montrant une variabilité dans le choix du style pour le compagnon de leur enfant. Cette variabilité s'avéra indépendante de leur propre style en tant que parents. Cette étude supporte donc l'utilisation des styles pour la personnalisation de l'Interaction Homme-Robot (IHR).
La recherche au sujet de l'acceptabilité des robots sociaux tend vers plus de modèle adaptatif dans le design de l'interaction homme-robot. Les différences individuelles jouent un rôle important dans l'interaction sociale, justifiant les stratégies d'adaptation à l'utilisateur. Nous proposons l'utilisation de styles comme plus value dans la personnalisation de l'interaction avec un robot compagnon Nous nous basons sur des théories de psychologie pour poser les styles comme outils au design des robots compagnons. L'état de l'art sur les styles nous amène à la customisation dans la manière d'effectuer une tâche en utilisant seulement des variations dans l'expression de signaux non-verbaux.

Mots Clef

Style, Plasticité, Interaction Homme Robot

1 Introduction

La recherche en interaction homme-robot se tourne de plus en plus vers l'*affective computing* utilisant les émotions pour rendre les robots domestiques plus engageants. Les robots devenant relativement moins cher, la population étant de plus en plus technophile, et le vieillissement de la population sont parmi les nombreuses raisons qui mènent la recherche à doter ces robots d'une intelligence sociale.
Les robots compagnons sont des artefacts qui ont pour but d'assister l'utilisateur dans sa vie quotidienne. Doter les robots compagnons de rôles, comme coach personnel, ou manager de bureau pourrait nous permettre d'améliorer la qualité de vie de l'utilisateur. Comme soulevé dans [22, 27], l'un des plus gros challenge dans la conception d'un tel compagnon est de le doter de compétences sociales en terme de perception, de raisonnement et d'expression dans l'interaction. D'autres enjeux comme la confiance, la légitimité et la crédibilité du compagnon pèsent également sur son acceptabilité. Certain travaux [16] sur le *framing* des rôles du compagnons ont donné des résultats intéressant en terme de confiance et de conformité. [31, 30]quant à eux proposent l'idée d'avoir un compagnon capable d'endosser de multiples rôles, de changer en fonction des besoins de l'utilisateur et en fonction du contexte. Nous irions donc vers un compagnon que nous qualifions de "polyvalent".
La *Théorie du Compagnon* présentée dans [22] soulève l'importance de différences individuelles influençant l'interaction et la construction d'une relation entre le compagnon et l'utilisateur. [12] propose d'aller vers plus de personnalisation afin d'améliorer l'acceptabilité des compagnons. L'acceptabilité des robots compagnons est lié à son apparence physique, son utilité, sa facilité d'utilisation mais aussi d'autres critères qu'il nous reste à explorer.
Ce travail sur la personnalisation des robots a pour but de créer de la valeur pour les parents d'avoir un compagnon pour leur enfants qui répondrai à leurs attentes et leurs besoins.
Dans une première section de ce papier aborde la définition de la notion de style. Ensuite, nous présentons une première étude sur l'expression de styles parentaux par deux types de robots utilisant des signaux non-verbaux de communication. Nous explorons les travaux connexes dans la littérature en interaction homme-robot. Finalement nous proposons nos perspectives de recherches découlant de cette étude.

2 Le *Style* : Personnalité dans l'action

2.1 Concept de *Style*

Depuis de nombreuses années, les industriels ont utilisé des tests basés sur la notion de style pour améliorer la collaboration dans les tâches en équipe. Les enseignants ont aussi adaptés leurs méthodes et leurs styles pour mieux convenir aux différents styles d'apprentissage des étudiants [32, 18].
Le *Style* est en psychologie une notion qui représente des catégories de comportement adoptés par des individu dans un rôle social spécifique. Il y a différents types de styles correspondant à chaque rôle social. Par exemple, il y a des styles d'apprentissage observés chez les étudiants et apprentis. De même, l'enseignant, le manager, le parent ont leurs styles associés. Ces styles permettent de catégoriser les individus dans différentes situations [18]. Nous voyons ici un lien fort entre le *style* et la notion de *persona* de

[22, 31, 30].
Comme expliqué par N. Darling [11], les valeurs et buts des parents envers le développement de leur enfant influencent leur comportement en tant que parent. Dès lors, leurs attentes en termes d'éducation et de développement pour leur enfant affectent leur style parental. À travers le modèle conceptuel proposé par Darling, nous pouvons entrevoir comment le concept de style parental pourrait être utilisé pour adapter le comportement d'un robot compagnon afin d'être en meilleure adéquation aux buts et valeurs des parents. Ainsi, dans ce papier nous proposons de personnaliser le comportement d'un robot compagnon pour enfant en utilisant les modèles de styles parentaux issus de travaux en psychologie et d'évaluer la réaction des parents en termes de perceptibilité, crédibilité et acceptabilité.

2.2 Styles Parentaux

N. Darling [10] décrit les 2 composants du comportement parental : les pratiques parentales d'une part et les styles parentaux d'autres part. Les pratiques parentales sont des actions régies par un but. Elles sont spécifiques au domaine (aller aux rencontres parent-prof, utiliser la fessé,...). Les styles parentaux sont des groupes d'attitudes dans l'interaction parent-enfant qui portent des signaux socio-affectif (tons de la voix, language corporel,...) mais ne sont pas liés au but. Contrairement aux pratiques parentales qui sont spécifiques au domaine, les styles parentaux sont construits au travers d'interaction multiple par l'utilisation de signaux sociaux. Nous choisissons d'utiliser les styles parentaux car ils influencent le comportement non pas dans la tâche en elle-même, mais plus l'attitude communicative.
Une typologie des styles parentaux a été proposée par Baumrind dans les années 70. De nombreuses études ont validé la robustesse de cette typologie et ont proposé des questionnaires pour les mesurer. En plus des 3 styles parentaux proposés par Baumrind (*Permissif*, *Autoritaire*, *Exigeant/Chaleureux*), en 1983 Baumrind, Maccoby & Martin ajoutèrent un quatrième style : le style *Négligent* [33].

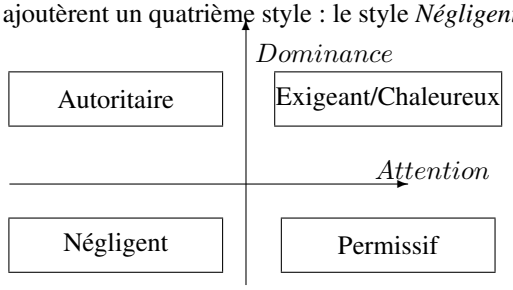


FIGURE 1 – 4 Styles Parentaux arrangés sur 2 dimensions - dominance et attention

La Figure 1 montre les 4 types de styles parentaux arrangés selon deux dimensions orthogonales correspondant au niveau de dominance et d'attention portée à l'enfant. L'échelle de *dominance* (aussi appelé contrôle) évalue le degré d'exigences, de convenir de règles de conduite, de fixer des limites et d'appliquer des sanctions en cas de transgression des règles.

La seconde dimension, l'*attachement* fait référence à la capacité à saisir les demandes et les besoins de l'enfant et d'y répondre en offrant du support émotionnel. Comme le contexte de ces travaux est un robot compagnon pour enfant, il nous semble indispensable que le compagnon soit sensible et attentionné envers l'enfant. Nous évaluons donc seulement les deux styles avec le plus haut taux d'attachement : *exigeant/chaleureux*(Authoritative en anglais) et *permissif* (Permissive en anglais).
L'expression par les robots d'émotions et de signaux sociaux est souvent difficile à cause de leur limitations motrices. Afin d'évaluer leurs capacités à exprimer des styles, nous choisissons de tester les modalités faciale et corporelle de façon indépendante. La figure 2 montre les deux robots utilisés, Reeti [2] (à gauche) et Nao[1] (à droite) exprimant les styles.

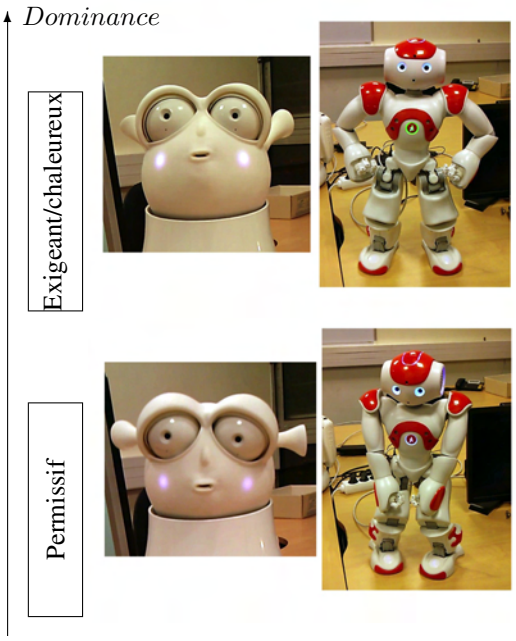


FIGURE 2 – Les deux robots (Reeti sur la gauche, Nao sur la droite), exprimant les deux styles, exigeant/chaleureux (haut) et permissif (bas)

3 Expérimentation

3.1 Hypothèses expérimentales

Afin de tester si les styles parentaux seraient un bon outil d'adaptation des comportements des robots compagnons en fonction des attentes des parents, nous faisons les hypothèses suivantes :

- H1 Les styles parentaux sont perceptibles et reconnaissables par les parents quand exprimés par des signaux non-verbaux faciaux ou corporels.
- H2 Tous les parents ne choisissent pas le même style pour leur enfant.
- H3 Il existe une corrélation entre le choix des parents

dans le style du compagnon et leur propre style en tant que parent :

H3.0 Leur choix est le même que leur propre style en tant que parent.

H3.1 Leur choix est l'opposé que leur propre style en tant que parent.

H4 Dans cette situation (donner une instruction), la crédibilité du robot varie en fonction de son style. Le style exigent/dominant est plus crédible que le style permissif.

H5 Dans cette situation (donner une instruction), la crédibilité du robot varie en fonction de l'age de son interlocuteur

3.2 Méthode

Afin de faire varier le comportement du robot en fonction des styles, nous avons utilisé des paramètres de la littérature [14, 26, 39, 38, 37]. Nous avons utilisé des variables spatiale comme, l'occupation de l'espace, la direction et l'amplitude des gestes. D'autres paramètres ont été utilisés pour la dynamique des mouvements : répétition, vitesse des gestes, vitesse de retour en position neutre, fluidité et rigidité du mouvement.

Afin de construire les styles exigeant/chaleureux et permissif pour les 2 robots nous utilisons la description des signaux non-verbaux de dominance de Hall[17].

Nous avons enregistré des vidéos montrant les robots avec des comportements des deux styles dans une même situation, demander à un enfant d'aller faire ses devoirs. Un questionnaire en ligne nous a permis de recueillir l'opinion de 93 parents. Il est certain qu'une interaction réelle aurait été préférable mais le questionnaire en ligne nous a permis de réunie un échantillon de taille correcte avec une population de parents variés. De plus, comme mentionné dans [35], les critères individuels pour la personnalisation peuvent être extraits d'enquêtes (i.e. *valeur individuelle* vs *valeur partagée*).

Un questionnaire basé sur [4, 19] était associé aux vidéos des robots. Chaque participant a vu un seul des robots jouant successivement les deux styles dans un ordre aléatoire. 44 participants étaient ont visionné le robot Reeti (expressions faciales) et 49 participants ont visionné le robot Nao (expressions corporelles).

Le questionnaire présentait plusieurs parties comme le détaille les paragraphes suivants.

La première partie du questionnaire concernait l'usage des nouvelles technologies. Cette partie avait pour but de détecter des signes de "technophobie" qui auraient pu biaiser notre étude. Comme présenté dans [25], certaines attitudes négatives envers les robots peuvent mener à de l'anxiété et à leur rejet.

Nous avons également questionné les parents au sujet de la crédibilité et de l'efficacité des robots donnant une instructions. Ainsi nous avons pu évaluer la compétence perçue du robot [5].

Dans la dernière partie du questionnaire, nous avons utilisé le questionnaire de style parentaux de [28] afin d'évaluer le style du participant avec son enfant. Nous avons utilisé les items pour les styles permissif et exigeant/chaleureux (les deux styles joués par les robots).

3.3 Résultats & Discussion

Parmi les parents interrogés, le ratio homme-femme était de 30 pour 63. Les participants ont été recrutés via la liste de diffusion du RISC, ainsi nous avons eu un certaine variabilité dans l'âge et la situation socio-professionelle des participants.

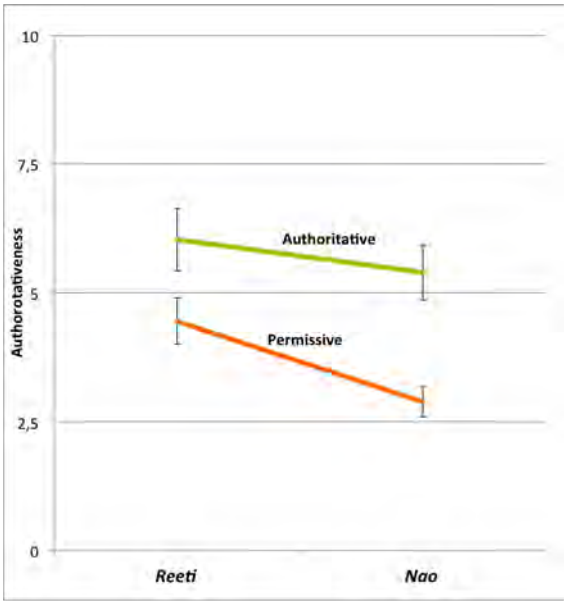


FIGURE 3 – Autorité perçue des robots Nao et Reeti exprimant les styles parentaux

Il a été demandé aux parents de noter le comportement des robots en terme de directivité : *" A votre avis, le comportement de ce robot est-il directif ? Vous noterez de 0 à 10 le caractère directif de ce robot : 0 pas du tout directif à 10 très directif. "*. La moyenne de directivité pour chaque robot est présentée sur la figure 3. Ce graphique montre une interaction entre les deux modalités de styles et de robots en termes de perception de directivité. En effet, il y a un effet du style sur la perception de directivité. Lorsque chaque robot exprime un comportement dominant il est perçut comme plus exigent que la condition avec le style permissif. Ce résultat valide notre première hypothèse (H1) telle que les robots peuvent exprimer des styles en utilisant seulement des signaux de communications non-verbaux et que l'utilisateur peut les reconnaître. Cependant, nous observons également un effet de l'apparence du robot et de la modalité de communication (faciale vs corporelle). En effet, le style permissif est mieux reconnu lorsque exprimé par Nao. De façon surprenante, Reeti est toujours perçut comme plus exigeant que Nao alors qu'il ne s'exprime qu'à travers la modalité faciale. Une explication de ce phénomène pourrait être liée au genre attribué aux robots ; Reeti

(figure 4) est plus souvent perçut comme un robot mâle que Nao.

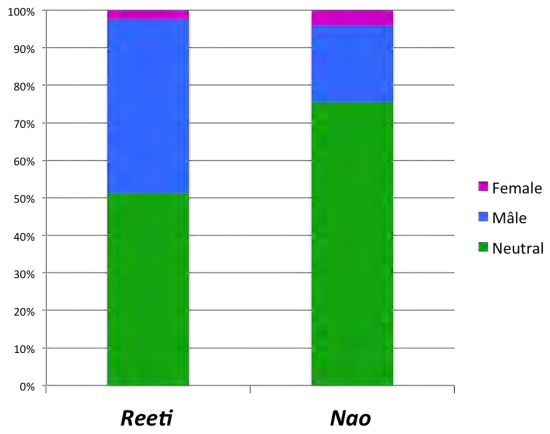


FIGURE 4 – Genre perçut pour chaque robot

Nous avons demandé aux participants de choisir un style exigeant/chaleureux, permissif ou aucun pour leur enfant. Parmi les parents qui ont choisi un style (environ 2/3 des

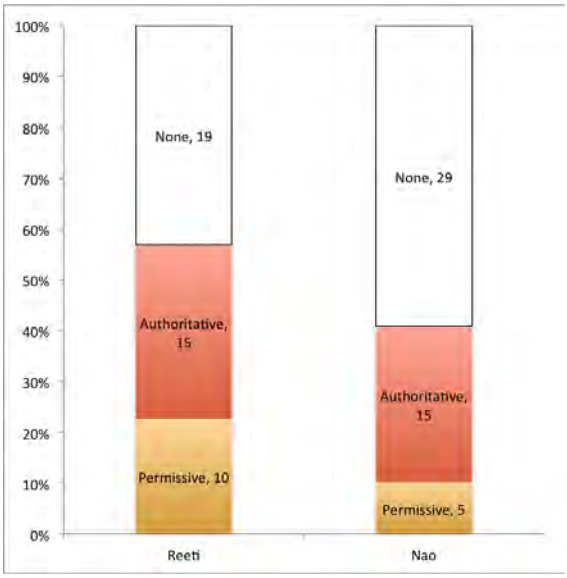


FIGURE 5 – Distribution des choix entre les styles parentaux pour les deux robots

parents), nous observons sur la figure 5 que la proportion de personnes choisissant le robot exigeant/chaleureux est supérieur au robot permissif, et cela pour les deux robots. Pour les personnes qui ont vu Reeti, environ 60% ont exprimé le désir d'utiliser le robot pour leur enfant. 35% des parents de la condition Reeti ont choisi le robot de style exigeant/chaleureux pour leur enfant. Pour Nao, moins de personnes l'ont considéré comme un possible compagnon pour leur enfant (environ 40%), mais lorsqu'il l'on fait, 75% ont opté pour le style exigeant/chaleureux. Ces résultats confirment l'hypothèse H2, à savoir qu'il existe une variabilité significative en terme d'acceptabilité selon l'utilisateur pour les styles exprimés par les robots.
Parmi les parents qui ont répondu au questionnaire sur les styles parentaux, seulement trois se sont révélé permissifs (sur 93), et ceux-ci ne corrèlent pas avec les parents qui ont choisi un robot permissif pour leur enfant.

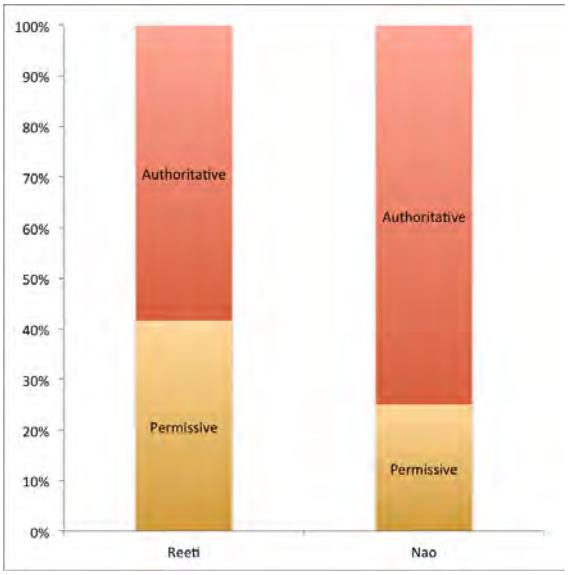


FIGURE 6 – Distribution des styles préférés par les parents exigeant/chaleureux

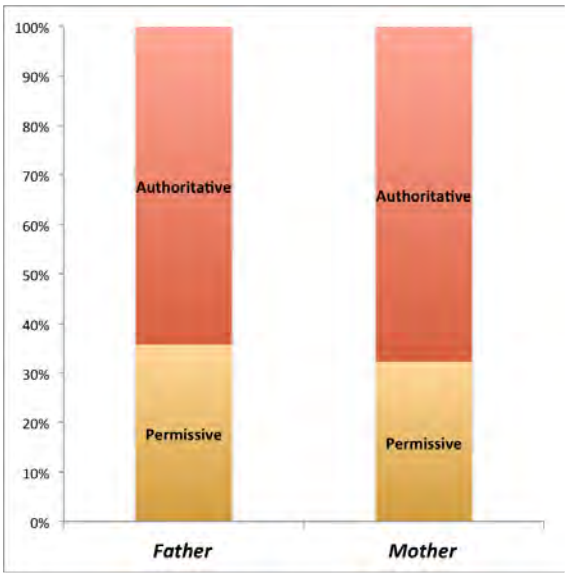


FIGURE 7 – Distribution des styles préférés en fonction du genre des parents

Les résultats présentés figure 6 montrent la distribution des préférences en termes de styles par les parents qui sont eux-même exigeants/chaleureux. Pour les deux robots, nous pouvons voir qu'il n'y a pas de consensus et pas de corrélation apparente entre le fait que ces parents soient exi-

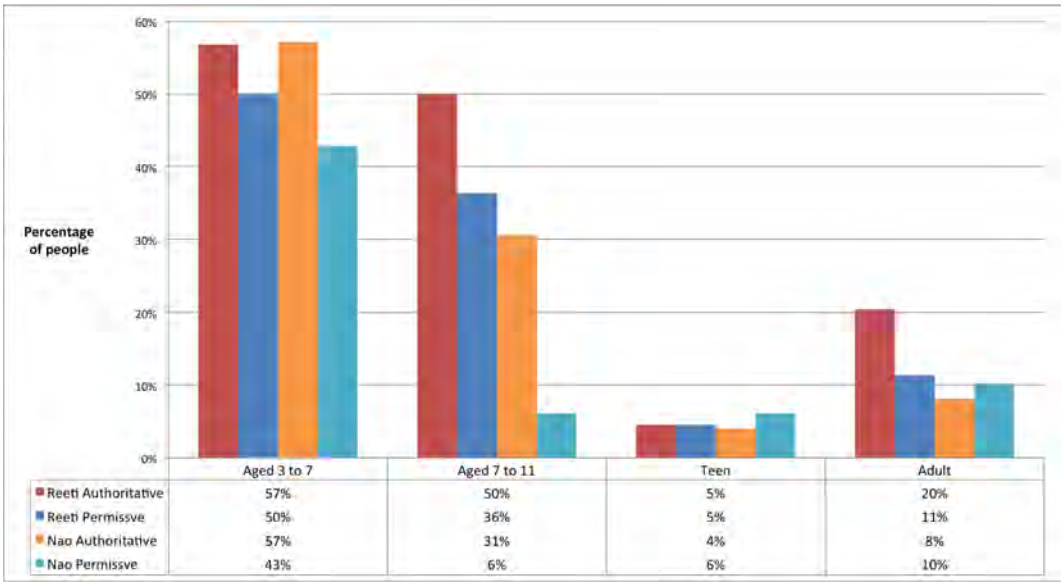


FIGURE 8 – Crédibilité des robots dans ce rôle avec les deux styles en fonction de l'age.

geants/chaleureux et leur choix de style pour le robot de leur enfant, invalidant ainsi nos hypothèses H3 (H3.0 et H3.1).
Les résultats présentés figure 8 montrent la crédibilité des robots dans chacun des styles et pour différentes tranches d'ages (H5 non-rejetée). Nous voyons qu'au dessus de 11 ans la crédibilité du robot comme donneur d'instruction chute. Nous notons également que pour chaque robot le style exigeant/chaleureux est toujours perçu comme légèrement plus crédible que le style permissif, supportant notre hypothèse H4.

4 Travaux connexes

4.1 Motivation & Persuasion

Les questions de persuasion et de motivation sont souvent abordées dans les recherches sur l'interaction sociale humain-robot.
Les travaux de [36, 15] sur le concept de motivation par des robots assistants ont montré que la personnalisation de l'interaction humain-robot avait un impact positif sur la santé. Dans ces travaux, le but était d'influencer la "motivation intrinsèque" de l'utilisateur afin de maintenir l'engagement grâce à la personnalisation de l'interaction sociale avec le robot. Les résultats ont montré une meilleure confiance de l'utilisateur dans le système lors d'une interaction de coaching lorsqu'il était personnalisé.
D'autres travaux adaptant les conseils d'un coach ont montré que cette personnalisation améliorait les performances de l'utilisateur [23].
Certains travaux ont aussi utilisé des robots persuasifs et ont pu évaluer l'influence des signaux sociaux sur la persuasion [29]. Cette étude a montré le risque lié à l'utilisation de signaux sociaux dans une tâche de persuasion. En effet, plus le système montrait des signaux sociaux plus l'utilisateur le "rejetait".
Persuasion et motivation sont définitivement des applications avec un gros enjeux pour la recherche en interaction homme-robot Le contexte de nos travaux est proche de ces questions de persuasion et de motivation, le scénario applicatif de notre recherche étant celui d'un robot compagnon pour un enfant seul à son domicile. Le robot compagnon gère l'emploi du temps de l'enfant et lui demande par exemple, comme nous l'avons vu précédemment, d'aller faire ses devoirs lorsque nécessaire. Ceci nous a amené à nous questionner sur les attentes et besoins des enfants et de leurs parents lors de l'accomplissement d'une telle tâche. Il peut y avoir plusieurs façon de faire cette tâche, mais laquelle est la meilleure, la plus efficace, la plus crédible et la plus acceptable pour l'enfant et aussi pour ces parents ?

Afin d'adapter l'interaction à l'utilisateur, certains travaux on proposés de tester l'expression de traits de personnalité par le robot. Contrairement aux travaux de personnalisation où le robot est une extension de l'utilisateur (je crée mon compagnon à mon image, il reflète mes goûts, etc), les travaux sur la personnalité vise d'avantage à donner une identité propre au compagnon qui peut être compatible avec celle de l'utilisateur.

4.2 Des robots qui ont de la personnalité

Dans [24], les auteurs proposent un moyen pour la conception d'une personnalité en utilisant du "profilage". Meerbeek mentionne l'influence de la tâche et du rôle sur l'expression de cette personnalité sans toutefois proposer de solution pour la modéliser. D'autres travaux [6] sur l'interaction de personnalités compatibles entre enfant et robot n'ont pas montré une amélioration d'acceptabilité. Nous pensons que la limitation de ces travaux est de considérer

la personnalité comme un profil indépendant de la tâche ou du rôle social alors que la personnalité humaine est un ensemble de traits dynamiques et multi-facettes.
Dans [36], deux niveaux de personnalisation ont été testés : personnalité-matching et une adaptation aux performances de l'utilisateur dans une tâche. L'adaptation comportementale du système a été faite en imitant les traits de personnalité de l'utilisateur et en utilisant diverses méthodes thérapeutiques en fonction des performances du participants. Ce papier contribue également à l'hypothèse que la personnalisation mène à un plus grand engagement de l'utilisateur, augmentant sa motivation et ses performances à accomplir une tâche
Dans [15], les auteurs montrent que des signaux relationnels pourraient augmenter la sensation de compagnonnage mais aussi accentueraient la sensation d'utilité dans une situation d'interaction avec un robot coach. Ainsi, l'utilisation d'une interface sociale, pour une même tâche augmente de façon significative la motivation et la valeur de l'interaction avec le robot. Dans ce même article, les auteurs suggèrent qu'il serait bon d'adapter l'interaction en fonction des préférences utilisateur.
Nous appuyant sur ces travaux, nous proposons d'ancrer l'utilisation de styles comme outil de personnalisation par l'utilisateur de son compagnon en fonction de son rôle sociale.

5 Perspectives de recherche

Heerink [20] propose un modèle représentant l'impact de différentes variables affectives et sociales sur l'intention d'utilisation d'un robot compagnon par des personnes agées. La recherche se penche maintenant au delà des questions d'évaluation fonctionnelle (utilité et facilité d'utilisation) en intégrant des notions de sociabilité dans les modèles d'acceptabilité. De récents travaux ont aussi mis en cause l'impact de l'utilité versus sociabilité sur l'acceptabilité et la perception par l'utilisateur. Dans [3], même si les participants étaient engagés dans des "tâches utiles" avec le robot Nao, le robot était principalement perçut comme plus social que utile.
Les travaux de G. Cockton [8, 9] traitent du sujet du Design Centré Valeur. La valeur consiste à une motivation ressenti par l'utilisateur qui le pousse à l'interaction, à l'utilisation, mais aussi à l'achat, à la recommandation d'un système. Nous souhaitons utiliser cette approche afin d'expliciter les motivations de l'utilisateur à interagir avec un robot compagnon.
En psychologie, les valeurs humaines sont décrites comme des critères d'évaluation individuels qui influencent sur nos préférences et nos actions[34]. Dans [35, 7], les auteurs remarquent que les valeurs sont contexte-dépendantes et qu'elles devraient être opérationalisées en fonction du contexte. Ainsi, afin que les utilisateurs trouvent de la valeur dans le compagnon, nous ne devons pas introduire de valeurs cachées mais essayer au contraire d'expliciter les valeurs en fonction de l'utilisateur afin de satisfaire au mieux ses attentes.
Ruckert [31, 30] suggère de concevoir plusieurs personas pour un compagnon unique. Comme le rôle du compagnon sera définit par le contexte, les contraintes de plateformes et les besoin de l'utilisateur, on peut imaginer que ce robot endosserai différents rôles. Comme Ruckert propose, il devrait y avoir une harmonie entre ces différents ?êtres ?. Une idée proche fut développée par Kramer [22] : "Nous devons aller au delà de l'imitation de rôles humains uniques vers une véritable identité pour le compagnon - qui est une collection de différentes identités." Kramer suggère de ne pas créer un *persona* unique et parfait mais de laisser à l'utilisateur une chance d'assigner les rôles et personnalité à son propre compagnon.
Dans le cadre de nos travaux de recherche, nous choisissons de nous rattacher à la notion de *styles* pour faire référence à ces différents *personas* qui reflètent les manières d'accomplir différents rôles. Un robot compagnon polyvalent pourra ainsi accomplir plusieurs rôles avec des styles correspondant aux souhaits de son utilisateur. Nous espérons via cette forme de personnalisation augmenter l'acceptabilité des robots compagnons tout en gardant une flexibilité entre les rôles endossés par ceux-ci.

6 Conclusion

L'étude présentée dans ce papier nous a permis d'explorer la notion de style, dérivée de la psychologie, comme outil de personnalisation du comportement du robot par leur parent. Il s'agit de la première étude sur les styles parentaux pour des robots compagnons. Nous pensons que ceci peut ouvrir à de nombreuses voies de recherche (d'autres styles pour d'autres rôles). Notre expérience n'a pas montré de corrélation entre style des parents et style attendu pour le robot. Cependant, nous sommes conscients des limites dues à l'apparence des robots. Une étude sur l'expression de style par des compagnon non-anthropomorphique pourrait permettre d'évaluer l'impact de l'apparence dans l'expression du style.
Nous allons conduire une étude avec des enfants en interaction avec les robots. Nous envisageons de prendre en compte le contexte [13, 21] dans la prise de décision sur le rôle à adopter. Nous souhaitons demander au parent de personnaliser au préalable le robot pour leur enfant. Nous évaluerons alors l'adaptabilité du système ainsi que le ressenti des parents et des enfants en interaction avec les robots adoptant des styles variés en contextes variés.

Références

[1] Aldebaran nao robot, 2013.

[2] Robopec reeti robot, 2013.

[3] Ritta Baddoura and Gentiane Venture. Social vs. Useful HRI : Experiencing the Familiar, Perceiving the Robot as a Sociable Partner and Responding to Its Actions. *International Journal of Social Robotics*, 5(4) :529–547, September 2013.

[4] Christoph Bartneck, Elizabeth Croft, and Dana Kulic. Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. *Metrics for HRI Workshop, Technical Report*, 2008.

[5] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1(1) :71–81, November 2008.

[6] Tony Belpaeme, Paul Baxter, Robin Read, Rachel Wood, Heriberto Cuay, Bernd Kiefer, Stefania Racioppa, Deutsches Forschungszentrum, Georgios Athanasopoulos, Valentin Enescu, Rosemarijn Looije, Mark Neerincx, Yiannis Demiris, Raquel Ros-espinoza, Aryel Beck, Lola Ca, Antione Hiolle, Matthew Lewis, Ilaria Baroni, Marco Nalin, Fondazione Centro, San Raffaele, Piero Cosi, Giulio Paci, Fabio Tesser, Giacomo Sommavilla, and Remi Humbert. Multimodal Child-Robot Interaction : Building Social Bonds. *Journal of Human-Robot Interaction*, 1(2) :33–53, 2012.

[7] Sunil Choenni. Embedding Human Values into Information System Engineering Methodologies. *...Conference on Information ...*, 2010.

[8] Gilbert Cockton. From quality in use to value in the world. *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04*, page 1287, 2004.

[9] Gilbert Cockton, S Kujala, P Nurkka, and T Hölttä. Supporting worth mapping with sentence completion. *Human-Computer InteractionâĂŞINTERACT*, (August) :24–28, 2009.

[10] N Darling. Parenting Style and Its Correlates. ERIC Digest. pages 1–7, 1999.

[11] N Darling and L Steinberg. Parenting Style as Context : An Integrative Model. *Psychological bulletin*, 1993.

[12] K. Dautenhahn. Robots we like to live with ? ! - a developmental perspective on a personalized, life-long robot companion. *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, pages 17–22, 2004.

[13] Anind K. Dey. Understanding and Using Context. *Personal and Ubiquitous Computing*, 5(1) :4–7, February 2001.

[14] Stephanie Embgen, Matthias Luber, Christian Becker-Asano, Marco Ragni, Vanessa Evers, and Kai O. Arras. Robot-specific social cues in emotional body language. *2012 IEEE RO-MAN : The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 1019–1025, September 2012.

[15] Juan Fasola and Maja J Mataric. Using Socially Assistive HumanâĂŞRobot Interaction to Motivate Physical Exercise for Older Adults. *Proceedings of the IEEE*, 100(8) :2512–2526, August 2012.

[16] Victoria Groom, Vasant Srinivasan, Cindy L. Bethel, Robin Murphy, Lorin Dole, and Clifford Nass. Responses to Robot Social Roles and Social Role Framing. *2011 International Conference on Collaboration Technologies and Systems (CTS)*, pages 194–203, May 2011.

[17] Judith a Hall, Erik J Coats, and Lavonia Smith LeBeau. Nonverbal behavior and the vertical dimension of social relations : a meta-analysis. *Psychological bulletin*, 131(6) :898–924, November 2005.

[18] John Hayes and Christopher Allinson. The Cognitive Style Index : Technical Manual and User Guide.

[19] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. Relating conversational expressiveness to social presence and acceptance of an assistive social robot. *Virtual Reality*, 14(1) :77–84, November 2009.

[20] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. Assessing Acceptance of Assistive Social Agent Technology byÂăOlder Adults : the Almere Model. *International Journal of Social Robotics*, 2(4) :361–375, September 2010.

[21] Frank Honold and F Schussel. Context Models for Adaptive Dialogs and Multimodal Interaction. *Intelligent Environments (IE), 2013 9th International Conference*, pages 57–64, 2013.

[22] Nicole Krämer, Sabrina Eimler, Astrid Von Der Pütten, and Sabine Payr. âĂIJTheory of companionsâĂİ What can theoretical models contribute to applications and understanding of human-robot interaction ? *Journal of Applied Artificial Intelligence*, (231868), 2011.

[23] Daniel Leyzberg, Samuel Spaulding, and Brian Scassellati. Personalizing robot tutors to individuals' learning differences. *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14*, pages 423–430, 2014.

[24] Bernt Meerbeek, Martin Saerbeck, and Christoph Bartneck. Iterative design process for robots with personality. *AISB2009 Symposium on New Frontiers in Human-Robot Interaction. SSAISB*, 2009.

[25] Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Kensuke Kato. Prediction of Human Behavior in Human–Robot Interaction Using Psychological Scales for Anxiety and Negative Attitudes Toward Robots. *IEEE Transactions on Robotics*, 24(2) :442–451, April 2008.

[26] Catherine Pelachaud. Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7) :630–639, July 2009.

[27] S Pesty and Dominique Duhaut. Artificial Companion : building a impacting relation. *Robotics and Biomimetics (ROBIO), 2011 ...*, 2011.

[28] David Reitman and PC Rhode. Development and validation of the parental authority : QuestionnaireâĂŞRevised. *Journal of Psychopathology and Behavioral Assessment*, 24(2), 2002.

[29] Maaike Roubroeks, Jaap Ham, and Cees Midden. When Artificial Social Agents Try to Persuade People : The Role of Social Agency on the Occurrence of Psychological Reactance. *International Journal of Social Robotics*, 3(2) :155–165, January 2011.

[30] JH Ruckert. Unity in multiplicity : Searching for complexity of persona in HRI. *Proceedings of the 6th international conference on Human-Robot Interaction*, pages 237–238, 2011.

[31] JH Ruckert, PH Kahn Jr, and Takayuki Kanda. Designing for sociality in HRI by means of multiple personas in robots. *Proceedings of the 8th ...*, pages 217–218, 2013.

[32] E Sadler-Smith and Richard Riding. Cognitive style and instructional preferences. *Instructional science*, 162691 :355–371, 1999.

[33] Leslie Schoonheere. L'Influence de Styles Parentaux sur la Consommation d'Alcool des Adolescents, 2005.

[34] Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus, and Mark Konty. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4) :663–88, October 2012.

[35] Katie Shilton, Jes Koepfler, and Kenneth Fleischmann. How to see values in social computing : methods for studying values dimensions. *Computer-Supported Cooperative Work and Social Computing*, 2014.

[36] Adriana Tapus, C ÅćÄČpuÅ§, and MJ Matarić. UserâĂŤrobot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intelligent Service Robotics*, 2 :169–183, 2008.

[37] HG Wallbott. Bodily expression of emotion. *European journal of social psychology*, 896(November 1997), 1998.

[38] Junchao Xu and Joost Broekens. Mood expression through parameterized functional behavior of robots. *RO-MAN, 2013*, 2013.

[39] Junchao Xu, Joost Broekens, Koen Hindriks, and MA Neerincx. Bodily Mood Expression : Recognize Moods from Functional Behaviors of Humanoid Robots. *Social Robotics*, 2013.

Strategic Intentions based on an Affective Model and a simple Theory of Mind

Hazaël Jones[1] Nicolas Sabouret[2] Atef Ben Youssef[2]

[1] UMR ITAP, SupAgro, Montpellier
[2] LIMSI-CNRS, UPR 3251, Orsay

Résumé

Cet article présente un modèle informatique pour raisonner sur les émotions de l'interlocuteur, en utilisant un paradigme de théorie de l'esprit (Theory of Mind, ou ToM, en anglais). Le système manipule des représentations sur des croyances à propos des émotions, des préférences et des buts de l'interlocuteur. Notre modèle affectif est conçu dans le contexte de la simulation d'entretien d'embauche mais il n'est pas lié à un ensemble d'affects spécifique. Il s'appuie sur des règles simples pour sélectionner les types de question lors de l'entretien en fonction de la personnalité de l'agent. Nous l'avons implémenté en utilisant une représentation de type OCC des émotions et un modèle dimensionnel PAD pour les humeurs.

Mots Clef

Théorie de l'Esprit, intentions stratégiques, modèles affectifs, entretiens d'embauche.

Abstract

This paper presents a computational model for reasoning about affects of the interlocutor, using a Theory of Mind (ToM) paradigm: the system manipulates representations of beliefs about the interlocutor's affects, preferences and goals. Our affective model is designed for the context of job interview simulation, but it does not depend on a specific set of affects. It relies on simple rules for selecting topics depending on the virtual agent's personality. We have implemented it using an OCC-based representation of emotions and a PAD model for moods.

Keywords

Theory of Mind, Strategic intentions, Affective model, Job interview.

1 Introduction

In order to build a credible interaction between a human and a virtual character, affective computing [Picard, 1995] proposes to simulate human affects in virtual agents, making them more realistic and engaging for interactions. In this context, one main challenge for Artificial Intelligence researchers is to make the virtual character adapt its behaviour to the perceived user's affective state, which will lead to a more natural and credible interaction for the user. To this purpose, we claim that virtual characters must not only use reactive behaviour in answer to a wide range of affects (emotions, moods, social attitudes...) such as in [Marsella et al., 2004, Kriegel and Aylett, 2008, Jones and Sabouret, 2013, Schröder, 2010]. It must also use *strategic intentions* about the human it interacts with. Strategic intentions can be seen as long term goals [Haddadi, 1996] for an agent. Indeed, in an interaction, people have intentions about the goal of a conversation, such as obtaining an information, finding an agreement, changing the interlocutor's point of view or having a fun and relaxing conversation. This paper proposes to analyse these strategic intentions and to use them in the reasoning model of an affective agent. To this purpose, we define a general model that can be adapted to different context. In our work, we apply this general model to a specific case of a formal interaction: job interviews in which the goal of the recruiter is to obtain concrete information about the interlocutor's social and technical skills, so as to select the best candidate.

Our general model is based on logical rules and is inspired by the theory of mind [Baron-Cohen, 1995] paradigm. Based on affect perception from Social Signal Interpretation (SSI), our virtual agent's model derives beliefs about the interlocutor's self-estimation in the interview (job skills, importance of the salary, etc). These informations are confronted to the agent's goals so as to select the next course of actions in the interaction (in our case, to conduct the job interview).

This paper is organized as follows. Section 2 makes a brief state of the art on theory of mind and shows how this has been used in the context of a virtual agent's reasonner. Section 3 briefly presents the job interview context and its specific features. Section 4 presents our architecture. In Section 5, we present in details our affective model that integrates the theory of mind for reasoning about affects in the context of interactions. The rules of this general model are illustrated on examples from the job interview context. The last section concludes on the model and its application to the job interview situation.

2 Related work on Theory of Mind

Theory of mind [Leslie, 1994, Baron-Cohen, 1995], or *ToM*, is the ability to attribute mental states (beliefs, intentions, desires, affects, ...) to others. The literature reports

numerous ToM studies and implementations in agent-agent interaction [Bosse et al., 2007, Dastani and Lorini, 2012]. In our work, we want to model the reasoning process of an agent that reason about the reasoning process of a human (the applicant). This particular configuration raises additional difficulties and leads to a original model for our representation of the ToM. For example, in [Pynadath and Marsella, 2005], an agent has beliefs about others in a subjective way. Agent A has belief about agent B following the real structure of agent B beliefs. However, in our work, since agent B is the human applicant, we do not have any information about its belief structure. We must guess them from the outputs of the affect recognition module.

Nevertheless, this model of influence and belief change [Pynadath and Marsella, 2005] is based on work in psychology: the authors use influence psychological factors in their simulation framework: consistency, self-interest, speaker's self-interest and trust (or affinity). We believe that similar high-level reasoning structures must be proposed in reasoning models, to complement low-level reasoning on perceptions such as what is done in [Scassellati, 2002, Peters, 2006]. These papers focus on the perception aspects of ToM such as the desire of engagement, and are tailored for signal interpretation, not for the cognitive model of the virtual agent.

Several other applications have been studied with a Theory of Mind approach. For instance, [Bosse et al., 2007] proposes a reasoner for task avoidance, the agent can change its behaviour in order to alter the other agent's desires, intentions and *in fine*, actions to occur. This work has been extended in a more generic version [Bosse et al., 2011] that proposes a two-level BDI agent model: the first level is the agent's reasoner and the second one computes the ToM. Following a different approach, [Dastani and Lorini, 2012] also propose a model based on modal logics that extend the BDI paradigm. Each agent has a set of actions and a set of formulas that represent the agent's mental state. A formula has a degree of desirability for the agent and a degree of plausibility. The use of modal logic allows researchers to model the recruiter beliefs, desires and intentions, but it seems difficult to represent a real humans' mental states based only on perceptions.

This is the reason why we propose a model based on general rules that takes as inputs recognised affects from the interlocutor and strategic intentions for the virtual agents, and combines them in the ToM-based affective model. The goal of our model is to represent the reasoning process of an agent that reason about the reasoning process of a human. Our model will be applied and illustrated in the context of the TARDIS project[1] that considers a job interview simulation as an interaction.

[1]TARDIS stands for Training young Adult's Regulation of emotions and Development of social Interaction Skills. url: `http://www.tardis-project.eu/`

3 Job interview context

In job interviews, the recruiter needs to reason about the potential behaviour of the applicant in front of him. This evaluation is done by selecting different questions in order to provoke particular reactions on the interviewee [Rynes and Connerley, 1993]. For example, to test the applicant's capability to manage his or her stress, the recruiter can be voluntarily aggressive during the interview. Our goal in the TARDIS project is to model that kind of recruiter strategic intentions. To this purpose, we propose in the next section a formalism to represent these strategies and to reason about the applicant's state of mind, based on perceived social and emotional signals.

It is important to note that job interview simulation is an interesting situation for studying multimodal affective interaction with a virtual agent. The literature shows that expressed emotions and non-verbal behaviour play a key role in job interview: 1) it is used by the recruiter for evaluating the candidate [Sieverding, 2009] and 2) it influences the applicant's behaviour [Sieverding, 2009]. In addition, [Gatewood et al., 2010] shows that the recruiter uses different questions to assess the applicant's work performance, individual quality and specific regarding the job.

In our work, we take into account these three aspects. First, the applicant's non-verbal performance is evaluated by comparing expected social cues to detected ones, using the SSI system. Second, we compute an affective reaction (see section 5.1) for the recruiter (that should influence the candidate's behaviour). Third, we select the next topic of interest for the recruiter that tries to evaluate the applicant's competencies (see section 5.2). We propose several strategies for a recruiter (in section 5.6), from provocative to helpful, which change this topic selection (and its affective reaction) so as to influence the applicant's behaviour.

4 TARDIS general architecture

Figure 1 shows our global architecture. The TARDIS architecture considers four main components:

- The SSI component provides the affective model with information about the applicant's affects and social attitudes that are detected by the system.
- The Interview Scenario component tells the virtual recruiter the expectation in terms of emotions and attitudes, depending on the interview progress. In TARDIS, the agent has no understanding of the applicant's actual answers to the questions. It follows a scenario, that can be influenced by the recruiter perception and internal states and focus on the affective recognition and adaptation.
- The Animation component is responsible for expressing the virtual recruiter's affective state through its behaviour and expressions.
- The virtual recruiter component which is composed of two modules:

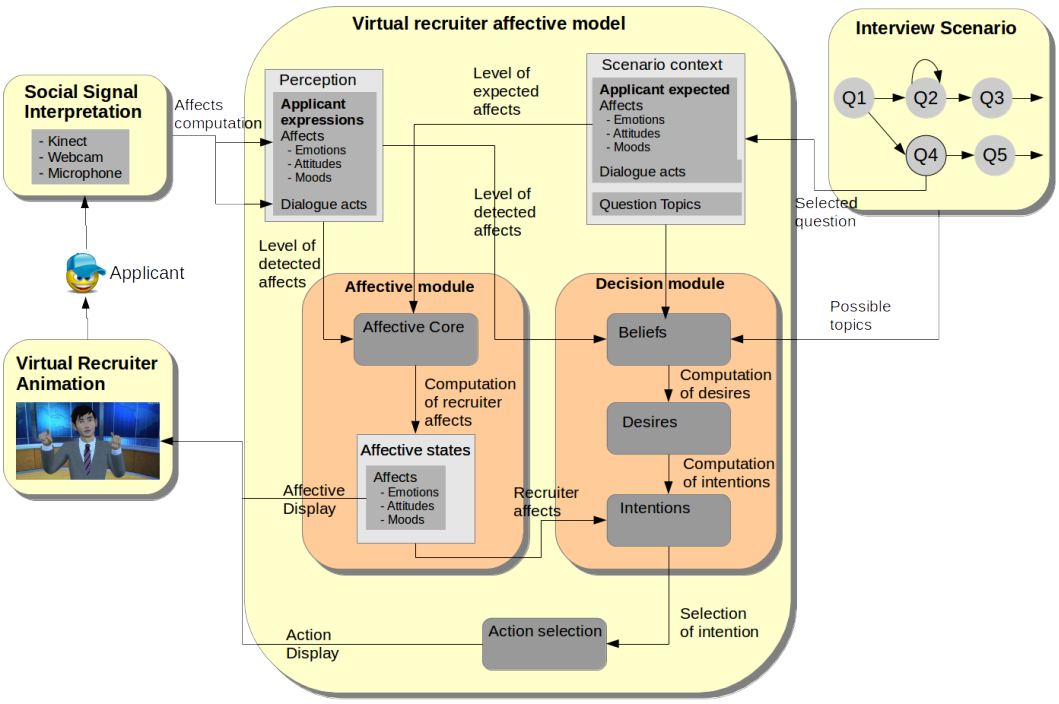


Figure 1: Global architecture for a recruiter in a job interview - Affective and Decision modules

- The Affective module, which is detailed in [Jones and Sabouret, 2012]. It provides a reactive model based on expectations from the recruiter and SSI of the applicant's affects expression [Jones and Sabouret, 2013]. It allows a computation of the recruiter emotions, moods and social attitudes.
- The Decision module, that is the focus of this document. The goal of this module is to build a theory of mind for a cognitive agent in the context of a job interview. Our agent (the recruiter) will deduce intentions of the applicant considering its answers (based on SSI) in a particular context (the question that has just been asked by the recruiter). This model will also influence new questions.

5 A ToM-based model for a cognitive virtual agent

In this section, we will present our model, and then shows its application in the TARDIS project.

5.1 General model for theory of mind

Our main objective is to draw beliefs about the interlocutor's mental states, preferences and understanding of the situation in the course of human-agent interaction, based on the user's reaction in terms of non-verbal behaviour (and social signal interpretation). In human-agent interaction, the agent has to select a new question at each turn-taking. To choose each new question, it is interesting that the agent has an idea of interlocutor's mental state regarding the past questions. This can be used in a wide spectrum of domains in which a human interacts with a virtual avatar in an interview simulation, such as teaching, training, ... The common aspect of these simulations is the use of questions by the avatar. Our model considers the use of question in order to manage the context of the answers of the person in interaction with the simulation.
To summarize, our theory of mind model has three main properties:

- It is about a real person who interacts with the system,
- It is centred on the person's preferences, expectations and interest for the job,
- It uses the context of questions and the affective behaviour to analyse user responses.

Our ToM however does not include any memory, other that the immediate response of the person and the previous ToM. We do not represent the knowledge that the interlocutor might have acquired during the interaction: we focus on its (estimated) preferences and expectations.

5.2 Context Management

With a view to manage the context, labels are given to the questions/sentences of the virtual character in order to interpret the answer/reaction of the human in term of beliefs on some topics. A list of topics can be done for each specific application. The set of topics set_{topic} contains N topics: $\{topic_1, topic_2, \ldots, topic_N\}$. Each subject is applica-

tion dependent and based on the domain of the simulation. A question is concerned by 0 to n topics.

List of topics. In order to manage the context, some labels are given to the questions of the recruiter in order to interpret the answer of the applicant in term of beliefs on certain subject. Here is the set of topics set_{topic} that can be tagged for a job interview:

- $topic_{applicant}$: questions about the applicant (general questions),
- $topic_{job}$: questions about the job,
- $topic_{salary}$: questions about the salary,
- $topic_{hours}$: questions about the working hours,
- $topic_{skill}$: questions about the competencies, the skills of the applicant regarding the job,
- $topic_{socialSkill}$: questions about the general social skills of the applicant,

A question is concerned by 0 to n topics. For example, the question "In what position will you like to work in our enterprise?" can be tagged by two different topics: $topic_{skill}$ and $topic_{job}$ because it tells about the applicant skills (the position he thinks he can apply for in this job) and its knowledge of the job (organisation of the enterprise).

5.3 Beliefs build

In order to build beliefs about the human who interacts with the system, we consider the questions/sentences that were just expressed by the virtual agent (identified by labels about topics) and the quality of the answer of the human from an affective point of view (which is obtained by Social Signal Interpretation, or SSI). Based on that, the agent will update its beliefs about the human on a particular subject. We denote the beliefs of the agent about the human $B_{Human}(topic_i)$ for i in $\{1, \ldots, N\}$.
According to the topic(s) raised by the question/remark of the agent, beliefs will be updated. In pursuance of building the beliefs of the human, we consider its answer (perceived via SSI) and decide if the answer is rather positive, negative or neutral. In order to determine if the global answer is positive or not, we use a performance index that compares the expected social cues (such as smile, large gestures, body movement, directed gaze...) with the detected ones. Expectations can be expressed as positive (signals that should be detected) or negative (behaviours that the user should avoid). This method in presented in [Jones et al., 2014]. The performance index pi is valued in the interval $[0, 1]$.
Based on this value and the topic tags of the question/remarks just done by the agent, the beliefs can be computed. Updates of each belief are done with the following formula for each topic:

$$B_{Human}(topic_i) \leftarrow B_{Human}(topic_i) + \alpha \times pi$$

with $\alpha \in [0, 1]$ a value that can be altered if we want the recruiter beliefs about the human to evolve quickly ($\alpha = 1$) or not (α near of 0). It can rely on the personality of the agent. An impulsive agent has an α near of 1 and a moderate one near of 0.
For instance, after a question about the job, with $\alpha = 1$ (impulsive recruiter) and an actual belief value $B_{Human}(job) = 0.5$, $B_{Human}(job)$ will become -0.3 for an $AverageAns$ of -0.8 making the recruiter beliefs about the applicant change sign in one answer. A moderate recruiter ($\alpha = 0.2$) will obtain a belief $B_{Human}(job) = 0.34$ which will change the dynamic of our simulation. An impulsive recruiter will cause strong dynamics and a moderate one smooth ones, which is the expected behaviour.

List of beliefs. The list of beliefs we consider in the context of job interviews is the following:

- self-confidence $B_{Human}(young)$,
- knowledge about the job $B_{Human}(job)$,
- importance of the salary for the applicant $B_{Human}(salary)$,
- importance of scheduling and working hours for the applicant $B_{Human}(hours)$,
- qualities of job skills $B_{Human}(skills)$,
- qualities of social skills $B_{Human}(socialSkills)$,

According to the topic raised by the question of the recruiter, beliefs can mean different things. For instance $B(young)$ is the belief of the applicant in himself. For a belief equal to 1, the applicant is very confident, and for -1, he has an important lack of self-esteem. Here the value quantifies the confidence of the applicant. If we look at $B(salary)$, the signification is different, it is about the importance that the applicant put in the salary when applying for this job.

5.4 Desires and goals

The desires are used to define the strategic intentions of the agent. We organize our desires in two categories: the high-level ones and the more specific ones. The high level intentions are directly linked to social attitudes. Attitudes can be initialized with personality and can evolve during the simulation but with a dynamics slower than the emotional one which is quite reactive. For more detail about the computation of social attitudes, refer to [Jones and Sabouret, 2012].
The high-level intentions are about the general intentions of the agent for the interaction, the specific ones are about specific beliefs about the human that interest the agent during the interaction.
The high-level desires are denoted: $D(Attitude)$. The specific desires are denoted: $D(B_{Human}(topic_i))$ because specific desires in an interaction are about beliefs of the human on a particular topic. For instance $D(B_{John}(football)$ is the desire of the agent to know if John has knowledge in the $football$ topic.

List of desires and goals. For a job-interview simulation, the recruiter will have a limited set of high-level intentions (provocative, pugnacious, friendly and helpful) and only one of them will be triggered in the same time.
The specific intentions are about subjects that the recruiter want to favour during the interview. The level of each subject will be adapted in function of the high-level intentions and will also consider the beliefs about the applicant. Actually, these specific goals for the recruiter are about the knowledge of applicant's beliefs on certain subjects. For instance, a question about the job will be associated to the goal $G(B_{Human}(job))$ because this question will give more information to the recruiter about the belief $B_{Human}(job)$.

5.5 Dynamics of goals

The high-level desires evolve in function of the social attitude of the agent. Social attitudes used can be defined on Leary circumplex [Leary, 1996]. According to the application, some attitudes will be relevant and some not. As shown by Leary, attitudes can be separated in two categories, the positive ones (friendly, cooperative, extroverted, ...) and the negative ones (hostile, critical, ...).
Based on these two kind of attitudes, we define algorithm 1 in order to update the desires of the agent.

Algorithm 1 Desires computation

if $(Attitude \in set(attitude_{-}))$ **then**
  for $B_{Human}(topic) \in set_{topic}$ **do**
    if $(AverageAns < 0)$ **then**
      $D(topic) \leftarrow D(topic) + \alpha \times |AverageAns|$
    else
      $D(topic) \leftarrow D(topic) - \alpha \times |AverageAns|$
if $(Attitude \in set(attitude_{+}))$ **then**
  for $B_{Human}(topic) \in set_{topic}$ **do**
    if $(AverageAns < 0)$ **then**
      $D(topic) \leftarrow D(topic) - \alpha \times |AverageAns|$
    else
      $D(topic) \leftarrow D(topic) + \alpha \times |AverageAns|$

This algorithm works as follows: if the agent has a negative attitude, he intends to select topics with a negative answer for the human. On the contrary, if the agent has a positive attitude, its desires are about topics with a positive answer from the human.

5.6 Goal selection

Several strategies can be defined for the selection of one desire in the list of possible desires. The most natural one is to select the desire with the maximum value in the available desires. At one moment of the dialogue, every possibilities (topics) cannot be approached in order to conserve the logical sequence of the conversation (a scenario for instance).

Goal selection based on recruiter's high level intentions. The high level goals can be defined directly in the scenario or be computed on the personality of the recruiter. We define 4 main strategies (2 for the positive attitudes and 2 for the negative attitudes). A recruiter with positive attitudes will have positive desires on topics where the applicant has positive average answers. On the contrary, a recruiter with negative attitude will have positive desires on topics where the applicant has negative average answers. Here are some strategies that we use for the virtual recruiter:

- Provocative recruiter: the recruiter will have a negative attitude and will always select the worst topic for the user (the one with the maximum Desire for the negative agent).
  $Intention = max(B_{Human}(subject))$
- Pugnacious recruiter: the recruiter will have a negative attitude but will randomly select one of the worst topic but not always the same.
  $Intention = random(max_n(B_{Human}(subject)))$
  with max_n, the n worst subjects.
- Friendly recruiter: the recruiter will have a positive attitude and will randomly select one of the best topic for the user but not always the same.
  $Intention = random(max_n(B_{Human}(subject)))$.
- Helpful recruiter: the recruiter will have a positive attitude and will always select the best topic (the one with the maximum Desire for the positive agent).
  $Intention = max(B_{Human}(subject))$

These different strategies lead to different goals. At one moment of the interaction, only some subjects can be approached according to the possibilities of the scenario. The recruiter will select the maximum in the available choices.

6 Conclusion

In this article, we propose a theory of mind model for an affective virtual agent. The theory of mind is about a real person in interaction with the system. It is centred on the interpretation of the affective states perceived through Social Signal Interpretation. By building beliefs about the person in interaction with the simulation, we allow an interaction by understanding the subjects where the person is confident or not. Then, according to the virtual agent high level intentions, new questions will be selected in a coherent and credible way regarding the personality of the agent.
This work is actually in the process of integration in the TARDIS platform. After integration, it will be evaluated in order to confirm that our model proposes a credible virtual recruiter for a job interview scenario. The theory of mind should provide coherent actions of the recruiter according to the reactions of the applicant and the personality of the recruiter. This aspect can be evaluated through literature and thanks to the applicants that will interact with the system. One current limit of our model is that it requires manual annotation of the scenario. The definition of an

automated annotation process, based on the utterance's semantic and contextual information, would greatly increase the scalability of the model.

Acknowledgment

This research is funded by the European Union Information Society and Media Seventh Framework Programme FP7-ICT-2011-7 under grant agreement 288578.

References

[Baron-Cohen, 1995] Baron-Cohen, S. (1995). *Mindblindness*. MIT Press, Cambridge, Massachusetts.

[Bosse et al., 2007] Bosse, T., Memon, Z. A., and Treur, J. (2007). A two-level BDI-agent model for theory of mind and its use in social manipulation. In *In AISB 2007 Workshop on Mindful Environments*, pages 335–342.

[Bosse et al., 2011] Bosse, T., Memon, Z. A., and Treur, J. (2011). A recursive BDI agent model for theory of mind and its applications. *Applied Artificial Intelligence*, 25(1):1–44.

[Dastani and Lorini, 2012] Dastani, M. and Lorini, E. (2012). A logic of emotions: from appraisal to coping. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '12, pages 1133–1140, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[Gatewood et al., 2010] Gatewood, R. D., Feild, H. S., and Barrick, M. (2010). *Human resource selection*. South-Western Pub.

[Haddadi, 1996] Haddadi, A. (1996). *Communication and Cooperation in Agent Systems - A Pragmatic Theory*. Springer.

[Jones and Sabouret, 2012] Jones, H. and Sabouret, N. (2012). An affective model for a virtual recruiter in a job interview context. In *4th International Conference on Games and Virtual Worlds for Serious Applications, Genoa, Italy, 29/10/12-31/10/12*, page in press. VS-GAMES'12.

[Jones and Sabouret, 2013] Jones, H. and Sabouret, N. (2013). TARDIS - A simulation platform with an affective virtual recruiter for job interviews. In *IDGEI (Intelligent Digital Games for Empowerment and Inclusion)*.

[Jones et al., 2014] Jones, H., Sabouret, N., Damian, I., Baur, T., André, E., Porayska-Pomsta, K., and Rizzo, P. (2014). Interpreting social cues to generate credible affective reactions of virtual job interviewers. *IDGEI (Intelligent Digital Games for Empowerment and Inclusion)*.

[Kriegel and Aylett, 2008] Kriegel, M. and Aylett, R. (2008). Emergent narrative as a novel framework for massively collaborative authoring. In *Intelligent Virtual Agents*, pages 73–80. Springer.

[Leary, 1996] Leary, T. (1996). Interpersonal circumplex. *Journal of Personality Assessment*, 66(2):301–307.

[Leslie, 1994] Leslie, A. M. (1994). ToMM, ToBy, and agency: Core architecture and domain specificity. In Hirschfeld, L. A. and Gelman, S. A., editors, *Mapping the mind Domain specificity in cognition and culture*, chapter 5, pages 119–148. Cambridge University Press.

[Marsella et al., 2004] Marsella, S. C., Pynadath, D. V., and Read, S. J. (2004). PsychSim: Agent-based modeling of social interactions and influence. In Munro, P., editor, *Proceedings of the International Conference on Cognitive Modeling*, volume 36, pages 243–248. Citeseer.

[Peters, 2006] Peters, C. (2006). A perceptually-based theory of mind for agent interaction initiation. *International Journal of Humanoid Robotics*.

[Picard, 1995] Picard, R. W. (1995). Affective Computing. *Emotion*, TR 221(321):97–97.

[Pynadath and Marsella, 2005] Pynadath, D. V. and Marsella, S. C. (2005). PsychSim : Modeling Theory of Mind with Decision-Theoretic Agents. *Information Sciences*, 19(1):1181–1186.

[Rynes and Connerley, 1993] Rynes, S. and Connerley, M. (1993). Applicant reactions to alternative selection procedures. *Journal of Business and Psychology*, 7(3):261–277.

[Scassellati, 2002] Scassellati, B. (2002). Theory of Mind for a Humanoid Robot. *Autonomous Robots*, 12(1999):13–24.

[Schröder, 2010] Schröder, M. (2010). The SEMAINE API: towards a standards-based framework for building emotion-oriented systems. *Advances in human-computer interaction*, 2010:2.

[Sieverding, 2009] Sieverding, M. (2009). 'Be Cool!': Emotional costs of hiding feelings in a job interview. *International Journal of Selection and Assessment*, 17(4).

Expressing social attitudes in virtual agents for social coaching

Hazaël Jones[1] Mathieu Chollet [2] Magalie Ochs[3]
Nicolas Sabouret[4] Catherine Pelachaud[3]

[1] Montpellier SupAgro; UMR ITAP
[2] Institut Mines-Telecom ; Telecom Paristech
[3] CNRS LTCI ; Telecom ParisTech
[4] LIMSI - CNRS ; Universit Paris-Sud

hazael.jones@supagro.inra.fr, nicolas.sabouret@limsi.fr,
{mathieu.chollet, magalie.ochs, catherine.pelachaud}@telecom-paristech.fr

Résumé

L'utilisation des agents virtuel pour le coaching social a connu une forte croissance ces dernires annes. Dans ce domaine, l'agent virtuel doit tre capable d'exprimer diffrentes attitudes sociales pour permettre l'utilisateur de s'entraner surmonter des situations de la vie relle. Dans cet article, nous proposons un modle d'attitudes sociales qui permet un agent virtuel de raisonner sur l'attitude sociale la plus approprie exprimer au cours d'une interaction avec l'utilisateur, en s'appuyant sur un modle d'agent affectif. L'expression de cette attitude se fait travers le comportement non-verbal.

Mots Clef

Attitudes sociales, motions, informatique affective, agent virtuel, comportement non-verbal.

Abstract

The use of virtual agents in social coaching has increased rapidly in the last decade. In social coaching, the virtual agent should be able to express different social attitudes to train the user in different situations than can occur in real life. In this paper, we propose a model of social attitudes that enables a virtual agent to reason on the appropriate social attitude to express during the interaction with a user given the course of the interaction, but also the emotions, mood and personality of the agent. Moreover, the model enables the virtual agent to display its social attitude through its non-verbal behaviour.

Keywords

Social Attitudes, Emotions, Affective computing, Virtual Agent, Non-verbal behaviour.

1 Introduction

Social coaching workshops constitute a common approach to help people in acquiring and improving their social competencies. The main difficulty with this approach is that it relies on the availability of trained practitioners as well as the willingness of the people to engage in exploring their social strengths and weaknesses in front of their peers and practitioners. For this reason, the use of virtual agents in social coaching has increased rapidly in the last decade [22, 2, 8]. However, most of the proposed models focus on the simulation of emotions and do not take into account the different social roles that the virtual agent may embody. Yet, given its role and the course of the interaction, the virtual agent should be able to express different social attitudes to train the user in different situations that can occur in real life. For this reason, one of the key elements of a virtual agent in the domain of social coaching is its ability to reason on social attitudes and to express them through its behaviour. This is why, in this research work, we propose a model of social attitudes for expressive virtual agents.

Social attitude can be defined as "*an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, colouring the interpersonal exchange in that situation (e.g. being polite, distant, cold, warm, supportive, contemptuous)*" [18]. As highlighted in [23], one's social attitude depends on one's personality but also one's moods that is directly influenced by the course of the interaction. One's social attitude is mainly conveyed by one's non-verbal behaviour [1].

Our aim is to develop a model of social attitudes that can be used by virtual agents to select determine expressive behaviours given its simulated affective state and the course of the interaction. The model we propose in this paper considers user's emotions, mood and

personality and compute and display agent's appropriate social attitudes, based on classical work from the literature in affective computing [7, 14]. Moreover, it enables the virtual agent to display its social attitude through its non-verbal behaviour, based on models of social attitude expression for virtual agents such as [3]. Our model has been developed in the context of job interview (JI) simulation. This context is interesting for several reasons: 1) social attitude plays a key role in JI: the applicant tries to fit the social norm and 2) the recruiter is in a role-play game, which can be simulated with a virtual agent. Moreover, job interview is a type of social coaching situation with high social impact. The methodology used to develop such a model combined a theoretical and an empirical approach. Indeed, the model is based both on the literature in Human and Social Sciences on social attitudes but also on the analysis of an audiovisual corpus of job interviews and on post-hoc interviews with the recruiters on their expressed attitudes during the job interview.

2 General Architecture

Our interview simulation for social coaching involves two main actors, namely the participant (*i.e.* the person that is training on the system) and the interlocutor (*i.e.* person or virtual agent that respond to the trainee). In our platform, the interlocutor is replaced by a virtual agent. Although the model presented here is general and can be applied to different interaction situations, our corpus and the derived cognitive architecture and non-verbal behaviour are designed in the context of job interview simulations when the agent acts as recruiter.
Fig. 1 presents our general architecture.

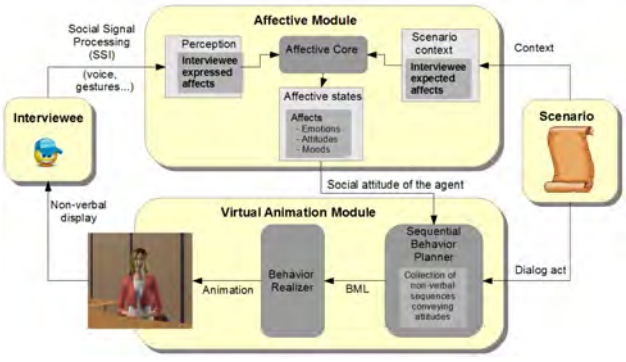


Figure 1: General architecture

This architecture is organized as follows. First, a *scenario* describes the context of the interaction and the expectations at the different stages of the interview in terms of affects expression[1]. For example, after a tough question, the interviewer will expect negative affects from the interviewee (distress, agitation). The *affective module* (described in ection 4) takes as inputs the scenario information and the detected social cues. It computes a social attitude for the virtual agent (the recruiter in our case). These attitudes are turned into non-verbal behaviours in the virtual agent *animation module* (Section 5) and used in the next interaction step. This module is based on the Greta platform [15] and is composed of an *intent planner* that generates the communicative intentions (what the agent intents to communicate) based on the scenario, and a *behaviour planner* that transforms the communicative intentions into a set of signals (e.g. speech, gestures, facial expressions) based on the agent's attitude and mental states.
In the following sections, we first present the corpus that was used to build our social attitude model. We explain how the corpus was collected, based on mock job-interviews, and it was annotated. We then give details of the Affective Module and the Animation Module and we show their interrelation.

3 Corpus

We have collected a corpus of real interpersonal interaction between professional recruiters and job seekers interviewees. The recordings consisted in creating a situation of job interviews between 5 recruiters and 9 interviewees. The setting was the same in all videos. The recruiter and the interviewee sat on each side of a table. A single camera embracing the whole scene recorded the dyad from the side. This resulted in a corpus of 9 videos of job interview lasting between 15 and 20 minutes each. We discarded 4 videos as the recruiter was not visible due to bad position of the camera. Out of the 5 remaining videos, we have so far annotated 3, for a total of 50 minutes and 10 seconds of video.
The non-verbal behaviours of the recruiters during the job interview have been annotated. We also annotate information on the interaction: the *turn taking* (i.e. who is speaking), the *topic of the discussion* (i.e. document related or general), the perceived affects of the interviewee (e.g. embarassed, relieved, bored...), and the attitude of the recruiter (i.e. the level of perceived dominance and friendliness of the recruiter). Different modalities of the recruiter's non-verbal behaviour have been annotated (e.g. the gaze behaviour, the gesture, the posture, the head movements, etc). The coding scheme, the resulting annotations, and the inter-annotators agreements are described in more details in [4]. These annotations have been used to construct the animation module for the virtual agent's expression of social attitudes (Section 4). Moreover, the annotation on the recruiter's social attitude has been used

[1] In our work, we use the SSI framework that recognizes the user's emotions expressed through his voice and his facial expressions.

for the evaluation of the affective model to check that it produces outputs that correspond to the human behaviour (Section 6).
In addition to these videos, post-hoc interviews with the recruiters were used to elicit knowledge about the expectations and mental states during the interview, following the methodology proposed by [16]. This knowledge was used in the affective module to select the relevant social attitudes and to set up the rules to give the capability to the virtual agent to select the appropriate social attitude to express.

4 Reasoning on social attitudes

The Affective Module is based on a set of rules that compute categories of emotions, moods and attitudes for the virtual recruiter, based on the contextual information given by the scenario and the detected affects (emotions, moods and attitudes) of the participant (Section 3). The computation of the virtual agent's emotions is based on the OCC model [14] and the computation of the agent's moods in the PAD space [13] is based on the ALMA model [7]. The personality is represenly by a vector in the OCEAN space [5]. The details of the computation of emotions and moods will not be presented in this paper; it can be found in [10].

4.1 Virtual recruiter's social attitudes

Several research has shown that one's social attitude is influenced by one's affective state (e.g. his emotions and moods [6]) and the course of the interaction (e.g. the affective reaction of the other [23]). For instance in [23], Wegener et al. show that a positive mood can help influence a change of attitude in the interlocutor, and that people tend to feel a higher likelyhood toward interlocutors that are in a positive mood.
Relations between attitudes and personality and moods and attitudes [23] has been exhibit in literature. Although we cannot give all the details of these papers here, their results show that one's mood has an influence not only on the interlocutor's attitude, but also on one's own reaction to events. This knowledge has been turned into expert rules that compute values for the attitude of the virtual agent.

Computation of attitudes. The way we compute attitudes follow this principle: an agent can adopt an attitude according to its personality [20] and to its actual mood [23]. For example, an agent with a non-aggressive personality may still show an aggressive attitude if its mood becomes very hostile. The mood compensates the personality and vice versa. For this reason, we use a logical-OR as condition on these two dimensions. As a consequence, in our model, the attitude can be triggered by one of these two dimensions. Then, the maximum value (mood or personality) is kept to compute the corresponding attitude, as is classically done in Fuzzy logics. We also use a threshold θ to define the minimum value for a trait to have an influence on the attitude.
As an example, the definition of the attitude "friendly" is based on Agreeableness (as was shown by Costa [5]) and positive and arroused moods (as proposed by Isbister [9]), i.e. exhuberance in the PAD space: If $(P_a > \theta) \vee (M_p > \theta \wedge M_a > \theta)$, then:

$$val(friendly) = max(\frac{M_p + M_a}{2}, P_a)$$

with P_a the value of the Agreeable personality trait in the OCEAN model, M_p and M_a the values of pleasure and arrousal in the PAD space.
Using similar rules, we compute 7 categories of attitudes: friendly, aggressive, dominant, supportive, inattentive, attentive and gossip. Details on that computation can be found in [17].

4.2 Interpersonal circumplex

The non-verbal behaviour model of our agent, presented in the next section, does not work directly with the categories that were identified in the Knowledge elicitation phase of the corpus collection. It makes use of continuous values that relies on the annotation of corpus which uses the Friendly and Dominant dimensions of the interpersonal circumplex [9]. For example, the *aggressive* attitude is defined on the circumplex as a vector $(-0,5,0,5)$ (friendly at -0.5 and dominant at 0.5).
To convert the attitudes represented by categories into continuous values of dimensions, we rely on the work by Isbister [9].When several attitudes are triggered at the same time, we compute the global attitude (that is the attitude that emerges) as the average of the associated vectors of these attitudes. The vectors' magnitudes influence this average giving more importance to an attitude with a large magnitude (*i.e.* intensity)). For n attitudes $a_i \in a_1, ..., a_n$:

$$\overrightarrow{attitude} = \frac{1}{n} \sum_{i=1...n} \overrightarrow{a_i}$$

with a_i the attitudes in the circumplex.

4.3 Evaluation of the affective model

To evaluate the affective module, we compare the affects that are computed from the affective module with the manual annotation of emotions from the job interview corpus. To perform such an evaluation, we chose arbitrarily one video of this corpus. The affective states of the interviewee and the social attitudes of the recruiter have been manually annotated. Our evaluation consists in comparing if, given the affective states of the interviewee, the affective module computes social attitudes for the recruiter that are identical (i.e. similar in terms of dimensions representation) as the manually annotated social attitude of the recruiter in the video.

The interview in the selected video is composed of 8 speaker turns. In the first 4 questions/answers turns, the interviewee is confident and gives good answers to the recruiter (positive detected affects). The interviewer expectations has been annotated positively during this first sequence. Then, the 4 following questions appear more difficult for the interviewee; he shows expression of negative affects. On the other hand, expectations of the recruiter were annotated positive for questions 5 and 6 and negative for questions 7 and 8.

Fig. 2 shows the recruiter affects. Fig. 2(up) shows the output for the recruiter's affects after each of its question. We can notice that positive affects (triangles) are positively correlated with supportive attitude while negative affects (squares) are positively correlated with aggressive attitude. Fig. 2(bottom) displays the annotated social attitudes of the recruiter as it evolves during the interaction. Positive values mean that the recruiter is perceived supportive by the user while negative values as aggressive. We also indicate where the 8 questions happen in the course of the interaction.

Comparing both data, the outputs of the affective module and the manual annotation, we can remark that the results are not really comparable in term of intensity of social attitudes. However the attitudes computed by the affective module coincide with the manually annotated attitudes. The variation of attitudes from supportive to aggressive happens in both cases. This example shows that our affective module computes, for a given input (the affect states of the interviewee), similar attitudes for the recruiter of those that are perceived in real human-human job interview.

5 Expression of attitudes

Research has shown that interpersonal attitude are conveyed through non-verbal behaviours (see Section 1). However it is insufficient to look at signals independently of the other surrounding signals: a smile is a sign of friendliness, but a smile preceded by head and gaze aversion conveys submissiveness [11]. In our work, to give the capability to a virtual agent to convey attitudes, we choose a corpus-based approach to find how attitudes are expressed through signal sequences. We then use this knowledge to generate new sequences of non-verbal behavior for animating our ECA. In the following section, we present an algorithm that extract sequences of non-verbal signal.

5.1 Frequent sequences extraction

In order to extract significant sequences of non-verbal signals conveying interpersonal attitudes from our corpus, we used a *sequence mining* technique. Using the attitude annotation files described in section 3, we segmented our corpus in time intervals preceding attitude variations. With a clustering technique, we regrouped together these attitude variations, such as small (resp. large) increases (resp. decreases) of friendliness (resp. dominance). The next step consisted of applying the Generalized Sequence Pattern (GSP) frequent sequence mining algorithm described in [21]. This nets us a set of frequent sequences for each type of attitude variation, we can characterize each of these sequences with several *quality measures*: *Support*, *i.e.* how many times the sequence appears in the data ($[0; \infty] \in \mathbb{N}$) ; *Confidence*, *i.e.* the proportion of a sequence's occurrences happening before one type of attitude variation ($[0; 1] \in \mathbb{R}$). Keeping sequences with a support of at least 10, we extracted a set of 879 sequences for dominance variations and 329 for friendliness variations. For instance, the $HeadNod \rightarrow Smile$ sequence was found frequently before large friendliness increases ($Support = 32$, $Confidence = 0.59$). In the following section, we describe the algorithm for generating non-verbal signals sequences conveying attitudes.

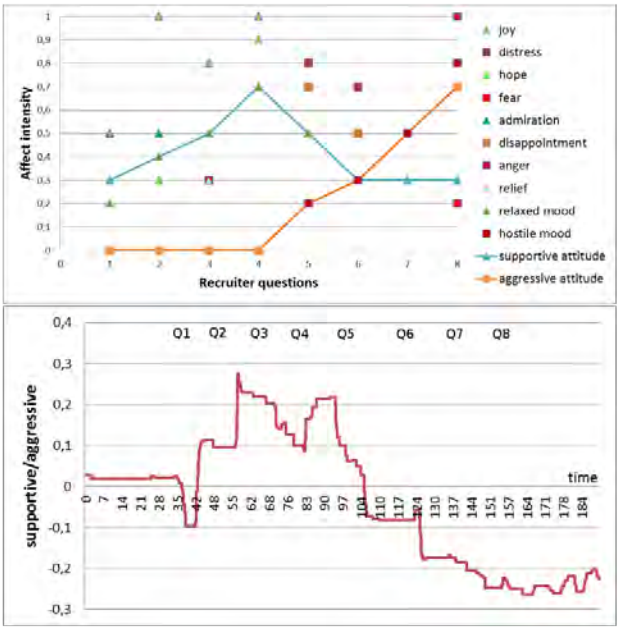


Figure 2: Computed recruiter affects after each question (up). Annotated recruiter state in real time (bottom)

5.2 Sequence generation

Given an input attitude that an ECA should express and an input utterance tagged with communicative intentions defined in the Functional Markup Language (FML) [12], the objective of our model is to generate a sequence of non-verbal signals that conveys the appropriate attitude. Our algorithm follows three steps:

Building minimal sequences - In a conversation, communicative intentions can be expressed through non-verbal behavior as well as through speech. For

instance, in Western culture, it is possible to convey uncertainty by performing a particular hesitation gesture. We specified behavior sets for every possible input communicative intention, *i.e.* the different non-verbal signals that can be displayed by an ECA to express the intention. The first step in our algorithm builds non-verbal signals sequences expressing an input message by selecting one signal in the behavior set of each communicative intention of the input message. Such a resulting sequence is called a *minimal sequence*.

Generating candidate sequences - For each minimal sequence obtained in the previous step, we retrieve all the time intervals where it is possible to insert other signals. For instance, if there is enough time between two head signals, we might insert a head nod or a head shake. For this purpose, we represent the extracted frequent sequences (Section 5.1) with a Bayesian Network (BN). This enables us to represent the causal and non deterministic relation of the attitudes on the signals (*e.g.* there might be more smiles for friendliness increases) and the sequences of signals (*e.g.* hands rest pose changes appear after gestures). An interesting feature of this model is that non-verbal signals sequences that did not occur in our data can still be generated, and their likelihood can be evaluated. Starting with the minimal sequences obtained after the previous step, we use the BN to add new signals in the available intervals, pruning out sequences that are too unlikely. The remaining sequences are called *candidate sequences*.

Selecting the final sequence - For selecting the final sequence, we defined a score variable of a candidate sequence s as $Sc(s) = P(s) * Conf(s)$, where $P(s)$ is the probability of s computed by the appropriate Bayesian Network (BN for dominance or friendliness depending on the input attitude), and $Conf(s)$ is equal to the sequence's confidence. The sequence with the highest score Sc is selected. In the next section, we present an study we realized to evaluate our model.

5.3 Evaluation

We evaluated our model with an online study realized using Adobe Flash technology. Participants were asked to compare 8 pairs of videos of a virtual character acting as a job recruiter expressing non-verbal signals when speaking (see Figure 3). For every pair of videos, the virtual recruiter said a different job interview question of the scenario . The character's speech was identical in both videos, however, the non-verbal behavior of the recruiter was different. The left video was generated with Greta's existing *Behavior Planner*, which does not consider attitudes and therefore considered neutral, while the right video was generated with our model with one of the 8 following attitudes *high dominance, low dominance, high low submissiveness, high submissiveness, high friendliness, low friendliness, low hostility, high hostility.*

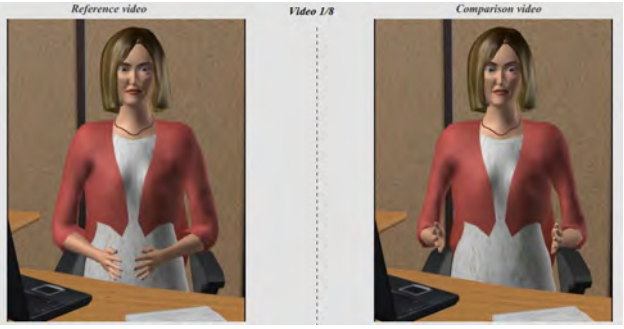


Figure 3: The main screen of the online study.

For every pair of videos, the participants answered were asked (*Q1*) how the right video compared to the left in terms of attitude (*e.g. much more dominant*) and how they would rate the attitude's intensity (*Q2*). Eighty-one participants took part in our study. The results of the study validate partially our model: the results of *Q1* showed that the expression of dominance, friendliness and hostility were recognized, however submissiveness was not recognized. All four results were statistically significant. For *Q2*, the only significant difference was found between intensity of large and small decreases in friendliness, however the participants identified large decreases as small decreases, and vice-versa. Therefore, it seems that our model cannot simulate attitudes of different intensities.

6 Concluding remarks

This paper proposes a new architecture for expressive agents that can reason about and display social behaviours. Unlike classical reactive models, our approach combines an affective reasoner that generates affects for the virtual character with a sequence selection mechanism based on a domain corpus annotated on two dimensions: dominance and friendliness. The methodology we propose contains several stages, from corpus collection and annotation, knowledge elicitation with experts for the definition of rules, implementation of behaviours corresponding to sequences and sequence selection based on the generated internal affects.

This architecture, whose components have been tested separately, has been integrated using the SEMAINE platform [19] and is currently being tested with real users. This will allow us to validate the global behaviour of our platform in the context of social coaching. However, several components can still be improved. One first limit of our model is that we assume exact inputs from the perception module. In addition, we intend to provide the affective reasoner with a representation of the interaction from the recruiter's point of view. We believe that allowing the recruiter to reason about the actual and potential behaviour of the applicant, following a Theory of Mind paradigm,

will allow a more credible decision process. We will also take the sequence extraction procedure further to take into account the user's non-verbal signals. This will require the sequence selection model to plan for and react to user signals.

References

[1] M. Argyle. *Bodily Communication*. University paperbacks. Methuen, 1988.

[2] R. Aylett, A. Paiva, J. Dias, L. Hall, and S. Woods. Affective agents for education against bullying. In *Affective Information Processing*, pages 75–90. Springer, 2009.

[3] D. Ballin, M. Gillies, and B. Crabtree. A framework for interpersonal attitude and non-verbal communication in improvisational visual media production. In *Proc. CVMP*, 2004.

[4] M. Chollet, M. Ochs, and C. Pelachaud. A multimodal corpus for the study of non-verbal behavior expressing interpersonal stances. In *Proc. IVA'13 Workshop Multimodal Corpora*, 2013.

[5] P. T. Costa and R. R. MacCrae. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional Manual*. Psychological Assessment Resources, 1992.

[6] Joseph P. F., Gordon H. B., and Susan E. K. The influence of mood on perceptions of social interactions. *Journal of Experimental Social Psychology*, 20(6):497–513, 1984.

[7] P. Gebhard. ALMA - A Layered Model of Affect. In *Proc. AAMAS*, pages 29–36, 2005.

[8] M. Hoque, M. Courgeon, J.-C. Martin, Bilge M., and Rosalind P. MACH: My Automated Conversation coacH. In *Proc. UbiComp*. ACM Press, 2013.

[9] K. Isbister. *Better Game Characters by Design: A Psychological Approach*. Morgan Kaufmann Publishers Inc., 2006.

[10] H. Jones and N. Sabouret. TARDIS - A simulation platform with an affective virtual recruiter for job interviews. In *Proc. IDGEI*, 2013.

[11] D. Keltner. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 68:441–454, 1995.

[12] M. Mancini and C. Pelachaud. The FML - APML language. In *Proc. AAMAS'08 FML Workshop*, Estoril, Portugal, May 2008.

[13] A. Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual Differences in Temperament. *Current Psychology*, 14(4):261, 1996.

[14] A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, July 1988.

[15] I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, and B. De Carolis. GRETA: a believable embodied conversational agent. In *Multimodal intelligent information presentation*, pages 3–25. Springer, 2005.

[16] K. Porayska-Pomsta, M. Mavrikis, S. D'Mello, C. Conati, and R. Baker. Knowledge elicitation methods for affect modelling in education. *International Journal of Artificial Intelligence in Education*, 2013.

[17] N. Sabouret, H. Jones, M. Ochs, M. Chollet, and C. Pelachaud. Expressing social attitudes in virtual agents for social training games. *Proc. IDGEI*, 2014.

[18] K. R Scherer. Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, pages 92–120, 2001.

[19] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, G. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. F. Valstar, and M. Wöllmer. Building autonomous sensitive artificial listeners. *Trans. Affective Computing*, 3(2):165–183, 2012.

[20] M. Snyder. The influence of individuals on situations: Implications for understanding the links between personality and social behavior. *Journal of Personality*, 51(3):497–516, 1983.

[21] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In P. Apers, M. Bouzeghoub, and G. Gardarin, editors, *Advances in Database Technology*, volume 1057 of *LNCS*, pages 1–17. Springer Berlin Heidelberg, 1996.

[22] A. Tartaro and J. Cassell. Playing with virtual peers: bootstrapping contingent discourse in children with autism. In *Proc. ICLS*, pages 382–389, 2008.

[23] D. T. Wegener, R. E. Petty, and D. J. Klein. Effects of mood on high elaboration attitude change: The mediating role of likelihood judgments. *European Journal of Social Psychology*, 24(1):25–43, 1994.

Etablissement de relations entre émotions, couleurs subjectives et couleurs objectives à partir d'annotations spontanées

Marie-Jeanne Lesot[1,2] Marcin Detyniecki[1,2,3]

[1] Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, Paris, France
[2] CNRS, UMR 7606, LIP6, Paris, France
[3] Académie des Sciences de Pologne, IBS PAN, Varsovie, Pologne
{marie-jeanne.lesot,marcin.detyniecki}@lip6.fr

Résumé

Cet article étudie les relations entre couleurs et émotions à partir d'annotations spontanées, collectées sur un site web de partage d'images, en considérant différents niveaux de subjectivité de description chromatique : il examine des couleurs objectives, interprétées et subjectives. Les résultats obtenus montrent que le double passage entre les 2 domaines, chromatique et émotionnel, et les 2 niveaux, objectif et subjectif, est difficile et soulignent la pertinence de représentations enrichies, subjectives mais non émotionnelles, pour faciliter la reconnaissance d'émotions.

Mots Clef

représentation des couleurs, perception subjective, apprentissage automatique, classification, caractérisation

Abstract

This paper studies the relationships between colours and emotions from pictures collected from an image sharing website, considering several subjectivity levels for the chromatic description : it examines objective, interpreted and subjective colours. The obtained results show that the double gap between the 2 domains, chromatic and emotional, and the 2 levels, objective and subjective, is difficult to bridge and they highlight the relevance of enriched representations, subjective but not emotional.

Keywords

colour representation, subjective perception, machine learning, classification, characterization

1 Introduction

La reconnaissance d'émotions vise à prédire les émotions des utilisateurs, qu'elles soient ressenties, quand les données décrivent les utilisateurs eux-mêmes, par exemple leurs expressions faciales, ou provoquées, quand les données décrivent les stimulus qui influencent les utilisateurs, par exemple des images. Cet article se place dans le second cas en considérant les émotions suscitées par les couleurs.
Les couleurs jouent en effet un rôle important et leurs relations avec les émotions sont exploitées dans de nombreux domaines, comme le design ou la publicité [8, 2]. Elles ont été largement étudiées, en psychologie [18, 11, 7] ou dans le domaine de l'art [12].

La reconnaissance d'émotions peut être formulée comme une tâche de classification automatique, où les classes correspondent aux émotions. La spécificité de ces dernières, due à leur subjectivité, transforme le problème traditionnel du fossé sémantique : celui-ci désigne l'écart qui existe entre la description, bas niveau et dépouillée de signification, d'une donnée, et le sens qu'on lui associe, qui constitue une représentation haut niveau dans un domaine d'interprétation distinct. Dans le cas de la détection d'émotions, ce passage entre les domaines bas et haut niveau se double d'un passage entre les niveaux objectif et subjectif : il ne s'agit pas uniquement d'extraire la signification des descriptions bas niveau objectives, mais de reconnaître le ressenti qu'elles suscitent, à un niveau subjectif.

Dans le cas des relations entre couleurs et émotions, il faut donc effectuer un double passage, entre les domaines chromatique et émotionnel d'une part, entre les niveaux objectif et subjectif d'autre part. Nous étudions la possibilité d'établir ces relations en exploitant des niveaux intermédiaires, enrichissant le niveau objectif par l'intégration d'une interprétation qui guide la mise en correspondance de la description numérique avec la perception émotionnelle : nous introduisons des représentations chromatiques interprétées ainsi qu'une notion de couleur subjective.

L'article est organisé de la façon suivante : la section 2 rappelle les travaux existants sur les relations entre émotions et couleurs. La section 3 est consacrée à la problématique de la constitution d'un corpus permettant de réaliser l'étude proposée, que nous basons sur l'exploitation d'un site de partage d'images. La section 4 discute la représentation des couleurs, en considérant le niveau objectif, son enrichissement par intégration d'une interprétation, ainsi que le niveau subjectif. La section 5 présente les méthodes mises en œuvre et les résultats obtenus pour la mise en correspondance de ces représentations chromatiques avec les émotions ; elle comporte également une étude sur la mise en relation des couleurs objectives et subjectives, étudiant le passage entre ces 2 niveaux au sein d'un même domaine.

2 Contexte

Les relations entre couleurs et émotions ont été largement étudiées en psychologie : des approches reposant sur des modèles dimensionnels des émotions ont par exemple établi des équations liant la valence ou l'activation de l'émotion ressentie face à une couleur, en fonction de la saturation et de la clarté de cette dernière [18]. Des approches catégorielles ont par exemple relié le rouge à l'amour, la colère et la passion, ou l'orange et le jaune à la joie [7]. Plus généralement, les couleurs claires (resp. sombres) sont associées à des émotions positives (resp. négatives) [11]. Pour ces études, les couleurs sont présentées sous la forme d'échantillons homogènes [11] ou évoquées par des étiquettes verbales, afin de laisser plus de liberté dans l'interprétation [7]. Ces études ne s'appliquent donc pas à des combinaison de couleurs, ni à des images complexes.

De nombreux travaux ont porté, de façon plus générale, sur les relations entre images et émotions, combinant divers indices parmi lesquels la couleur constitue un descripteur important, combiné avec la texture ou la forme par exemple [5, 6, 15]. D'autres travaux sont basés sur la prise en compte du contenu sémantique des images, pour lesquelles la charge émotionnelle est liée à ce qu'elles représentent, comme dans le corpus IAPS [13].

L'approche considérée dans cet article se situe à mi-chemin de ces 2 types d'approches : elle ne porte pas sur des couleurs isolées ou juxtaposées, mais sur des images complexes, pour lesquelles les répartitions des couleurs sont naturelles et jouent un rôle. Toutefois, elle est focalisée sur l'effet du seul attribut chromatique.

3 Acquisition de données

3.1 Sélection des données

L'acquisition d'un corpus représentatif constitue une étape essentielle de toute tâche d'apprentissage automatique. Elle vise ici à collecter des images mettant en relation des couleurs et des émotions.

Exploitation de sites de partage d'images Nous proposons d'exploiter le site Flickr (http ://www.flickr.com) qui permet de charger des photos et de les annoter par des mots-clé et qui présente de nombreux avantages pour la construction de corpus d'images annotées : d'abord, son usage très répandu implique qu'il contient de *grandes quantités* de données. De plus, ces images sont *générales* et non spécifiques à un contexte donné. En outre, la *validité* et la *spontanéité* des données récoltées sont garanties : les utilisateurs eux-mêmes choisissent les images qu'ils souhaitent charger, ainsi que le nombre de mots-clé et les mots-clé eux-mêmes ; les annotations reflètent donc leur perception sans contrainte. Il faut souligner que certaines étiquettes peuvent néanmoins être bruitées : les utilisateurs annotent parfois un ensemble d'images en appliquant des mots-clé à tout un répertoire et non à chaque image individuellement. La spontanéité des annotations constitue une différence importante par rapport à de nombreuses études sur les relations entre émotions et couleurs, dans lesquelles les images sont sélectionnées par les expérimentateurs et les annotations limitées à une liste prédéfinie. Enfin Flickr impose que toutes les annotations associées à une image soient fournies par un même utilisateur. Cette contrainte garantit leur *cohérence* pour une image donnée : elles reflètent toutes la même subjectivité.

Acquisition des annotations subjectives L'utilisation de Flickr conduit à une forme particulière d'annotation : au lieu de demander à des sujets d'étiqueter des images choisies au préalable, on collecte les images dont les mots-clé contiennent les annotations recherchées. La question se pose alors des termes émotionnels à rechercher. Parmi les nombreux modèles catégoriels des émotions, nous utilisons le niveau intermédiaire du modèle de Plutchik [17], considérant les émotions indiquées dans le tableau 1.

Sélection selon les termes chromatiques Afin de sélectionner les images pour lesquelles la charge émotionnelle est portée par le contenu chromatique et non le contenu sémantique, nous proposons de considérer les images pour lesquelles l'utilisateur a jugé pertinent d'inclure un nom de couleur (au moins) dans leurs annotations : nous considérons qu'il souligne par là que la couleur joue un rôle important pour cette image. Ainsi, nous faisons l'hypothèse qu'une image dont les annotations comportent au moins un terme chromatique et un terme émotionnel constitue un exemple permettant d'extraire des connaissances sur la mise en relation entre couleurs et émotions.

En ce qui concerne le choix des termes de couleurs à considérer, nous proposons d'utiliser les 11 termes chromatiques universels classiques [4] indiqués dans le tableau 1.

3.2 Corpus constitué

On collecte alors les images en utilisant une requête du type "couleurs ET émotions", où les parties chromatiques et émotionnelles sont respectivement les disjonctions des termes indiqués précédemment. Pour la période du 01.01.2006 au 01.07.2009 on obtient de la sorte une base contenant 24 896 images dont 4 exemples sont donnés dans la figure 1, avec leurs annotations.

Analyse préliminaire On observe un déséquilibre dans les distributions des couleurs et émotions, indiquées dans le tableau 1 (la somme est supérieure au nombre total d'images car certaines images sont associées à plusieurs étiquettes) : 5 couleurs présentent moins de 3 000 occurrences. Le déséquilibre est plus grand encore pour les étiquettes émotionnelles : *joy* est présent dans plus de la moitié des images, alors que *disgust* et *anticipation* correspondent à moins de 2% des images. Cette répartition peut être expliquée par un biais lié aux sites de partage d'images : on peut faire l'hypothèse que les utilisateurs sont plus enclins à partager des images gaies que des images liées au dégoût.

On peut également noter que 62% des images sont étiquetées avec un unique mot-clé de couleurs : cela signifie que

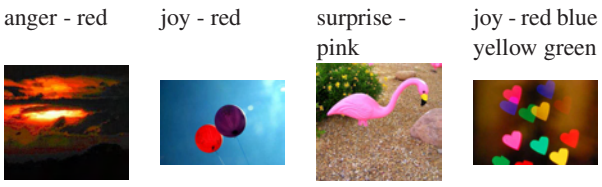


FIGURE 1 – Exemples d'images et leurs annotations.

TABLE 1 – Distribution des étiquettes chromatiques et émotionnelles, par fréquence décroissante

white	black	red	blue	green	
6 374	6 190	5 889	5 800	4 486	
yellow	pink	orange	purple	brown	grey
3 179	2 676	1 768	1 191	981	314
joy	sadness	fear	surprise	trust	
11 557	3 723	3 426	2 738	1 752	
anger	anticipation	disgust			
1 731	383	91			

la couleur subjective ressentie par un être humain est une simplification de la complexité chromatique de l'image entière. Cet effet est illustré sur les 2 premières images de la figure 1 : bien qu'elles contiennent beaucoup de noir (resp. de bleu), elles sont uniquement étiquetées comme *red*.
Les images présentant plusieurs mots-clé émotionnels constituent seulement 2% du corpus (494 images). Ceci signifie que la signification émotionnelle des images partagées est principalement considérée comme non ambiguë par rapport aux termes émotionnels étudiés.

Filtrage et corpus final Le corpus final est obtenu en éliminant les images dont les seuls mots-clé chromatiques sont {*black, white*}. En effet, certaines corrélations avec les émotions peuvent être étudiées pour les images noir et blanc, mais elles n'entrent pas tout à fait dans le cadre de l'étude des relations entre couleurs et émotions. Les images associées à plusieurs étiquettes émotionnelles étant ambiguës et largement minoritaires, elles sont également supprimées. Nous excluons aussi les émotions rares *disgust* et *anticipation*, qui ne possèdent pas suffisamment d'exemples d'apprentissage. Les émotions étudiées sont donc *joy*, *anger*, *fear*, *sadness*, *surprise*, *trust*. Après exclusion des images non libres de droit, le corpus final contient 21 749 images.

4 Représentations des couleurs

Plusieurs niveaux de subjectivité peuvent être envisagés pour la description des couleurs d'une image, cette section présente les couleurs subjectives et objectives puis l'intégration d'une interprétation dans ces dernières.

4.1 Couleurs subjectives

Les couleurs subjectives sont définies comme le résultat de la perception du contenu chromatique d'une image, exprimée par les termes chromatiques présents parmi les annotations associés à l'image. Ceux-ci expriment en effet l'interprétation de l'être humain qui regarde l'image et sélectionne celles qui lui semblent significatives.
La représentation de la couleur subjective est donc un vecteur binaire à 11 composantes, indiquant respectivement la présence ou l'absence des 11 termes chromatiques considérés dans les annotations de l'image encodée.

4.2 Couleurs objectives

Les couleurs objectives correspondent à une information chromatique bas niveau, extraite du contenu des images à l'échelle de leurs pixels : une image est représentée par un histogramme, basé sur une décomposition d'un espace de couleurs, tel que RGB ou HSV. Chaque composante de l'histogramme indique la proportion de pixels dont la couleur est située dans le sous-espace correspondant.
Le choix de la représentation de l'image dépend alors de la définition des sous-espaces. Une représentation objective est obtenue par une décomposition uniforme : ainsi, on peut décomposer les teintes en 8 régions égales et la saturation et la valeur en 4 régions chacune. La représentation numérique d'une image est alors un vecteur de 128 composantes, correspondant aux 128 sous-espaces ainsi définis. Dans la suite, cette représentation est appelée HSV128.

4.3 Couleurs interprétées par décomposition de l'espace

Le principe précédent offre différents niveaux d'objectivité de la description numérique des images : les sous-espaces peuvent aussi être définis de façon à regrouper des régions de l'espace des couleurs perceptuellement homogènes, la décomposition de l'espace étant orientée par l'interprétation perceptuelle, conduisant à une représentation objective enrichie, qualifiée d'interprétée.
Nous utilisons la représentation qui décompose les teintes en 7 régions non-uniformes et l'espace saturation-valeur en 6 régions dont certaines sont fusionnées, conduisant à 36 sous-espaces [19]. Elle est appelée HSV36 dans la suite.

4.4 Couleurs interprétées par dictionnaire chromatique

Nous proposons également une décomposition de l'espace des couleurs basée sur la définition de points de référence : les sous-espaces ne sont pas définis par leurs frontières, mais comme les ensembles de points plus proches d'un point de référence fixé que des autres points de référence.
Cette approche permet de définir des frontières non linéaires, rendant la définition des sous-espaces plus flexibles. De plus, elle permet de calculer la distance entre points dans tout espace de couleur, alors qu'une approche par seuils impose des comparaisons dans l'espace considéré, le plus souvent HSV. En particulier, il est possible d'utiliser l'espace CIE Lab, précisément défini de telle sorte que la distance euclidienne dans cet espace soit corrélée à la distance perceptuelle entre couleurs. En outre, il est plus naturel et intuitif de définir des références de couleurs

que des seuils, qui sont les zones ambiguës de transition.
La méthode proposée pour constituer l'ensemble des points de référence, appelé dictionnaire chromatique, consiste à échantillonner régulièrement 500 points de l'espace chromatique, puis à associer chaque point à une étiquette chromatique. L'association a été obtenue dans le cadre d'une expérimentation impliquant une dizaine de volontaires. Cette décomposition définit alors 11 sous-espaces en 2 étapes : un pixel est d'abord associé au point de référence dont il est le plus proche, la distance étant calculée dans l'espace CIE Lab. Il est ensuite affecté au sous-espace défini par l'étiquette associée au point de référence.
La représentation obtenue des images, appelée D500 dans la suite en référence au dictionnaire à 500 entrées, est alors, comme HSV36, objective interprétée : elle exploite uniquement le contenu chromatique objectif de l'image, défini par ses pixels, mais repose sur une interprétation de la décomposition de l'espace des couleurs.

5 Relations entre émotions, couleurs objectives et subjectives

Trois types de mise en correspondance sont alors considérés entre les émotions et les espaces de représentation chromatiques, aux niveaux objectif, interprété et subjectif.

5.1 Caractérisation d'émotions à partir de couleurs objectives et interprétées

La caractérisation des émotions à partir de couleurs objectives, éventuellement interprétées, peut être formulée comme une tâche de classification [9]. Nous mettons en œuvre des arbres de décision construits par C4.5. Cet algorithme requérant des données équilibrées, 10 bases de données sont construites par tirage aléatoire avec remise ; une validation croisée 10-folds est effectuée sur chaque base. Pour des contraintes de place, l'évaluation des résultats est restreinte à la moyenne et l'écart-type du taux de bonne classification (tbc).

Constitution d'une référence Afin de comparer les problématiques des fossés sémantique et émotionnel, nous considérons une tâche pour laquelle le contenu chromatique des images joue un rôle plus évident que pour les émotions. Dans ce but, nous constituons un corpus contenant des images dont les mots-clé comportent *night, forest, lavender, desert, cloud* ou *sunset*, en appliquant le même protocole que pour la constitution du corpus émotionnel.
La prédiction de ces 6 classes à partir de la description HSV36 par C4.5 conduit à un taux de bonne classification de 46%. Cette valeur est supérieure à celle obtenue par un classifieur aléatoire, qui est de 17%, mais elle reste faible.

Caractérisation de 6 émotions Les taux de bonne classification moyens obtenus pour les 6 classes émotionnelles considérées sont 26.8% $\pm$ 0.5 pour HSV128, 27.8% $\pm$ 0.6 pour HSV36 et 26.5% $\pm$ 0.7 pour D500.
Les valeurs sont supérieures à celles obtenues par un classifieur aléatoire, mais plus faibles que celles obtenues sur la tâche de référence : comme attendu, la prédiction des concepts subjectifs que constituent les émotions est plus difficile encore que la prédiction des concepts haut niveau. On observe que HSV36 donne des résultats meilleurs que HSV128 et D500 qui sont comparables. On peut expliquer les performances faibles de HSV128 par son caractère pleinement objectif qui ne comporte pas d'interprétation des couleurs. Les résultats de D500, qui constitue une représentation interprétée comme HSV36, peuvent être dus à sa faible dimensionnalité : il conduit à une description des images dans un espace à 11 dimensions seulement, ce qui peut être inadapté pour l'algorithme C4.5.
L'étude des résultats par émotion pour HSV36, non détaillés ici, montre que certaines des émotions sont plus difficiles à prédire que d'autres : les plus faciles, correspondant à un tbc de 32%, sont *joy* et *trust*, la plus difficile est *surprise* qui présente un taux de 20%.

Caractérisation de paires d'émotions Afin de diminuer la complexité de la tâche de prédiction, nous considérons des tâches bi-classes, définies pour les 15 paires d'émotions, pour la représentation HSV36. Les résultats, non détaillés ici, montrent que les taux de bonne classification varient de 66.7% pour la paire (*fear*, *joy*) à 55.1% pour (*sadness*, *fear*). On constate de plus que, de façon générale, les émotions de même valence apparaissent comme plus difficiles à distinguer que les émotions de valences opposées. On observe également que toutes les paires d'émotions comportant *surprise* obtiennent de faibles taux de bonne classification, tous inférieurs à 62.7%, au contraire de *trust* pour laquelle ils sont toujours supérieurs à 60.7%, cette valeur minimale étant obtenue par opposition à *surprise*.

Caractérisation par valence Ayant observé que la valence des émotions joue un rôle, nous regroupons les émotions en 2 catégories, opposant {*joy*, *trust*, *surprise*} à {*anger, sadness*, *fear*}. Il faut noter que l'affectation de *surprise* est discutable, cette émotion pouvant avoir une valence positive ou négative. Cette particularité peut expliquer la difficulté à la prédire dans les expériences précédentes. Le tbc est alors de 63.7%, ce qui reste faible.

5.2 Caractérisation d'émotions à partir de couleurs subjectives

Afin d'examiner plus en détail l'hypothèse selon laquelle la difficulté de la reconnaissance des émotions est due au double passage entre domaines, chromatique et émotionnel, et entre niveaux, objectif et subjectif, nous étudions la relation couleur – émotions au niveau subjectif exclusivement, en considérant les couleurs subjectives.
Une formulation de ce problème comme une tâche de classification ne semble pas appropriée, car, comme rappelé à la section 4.1, la représentation subjective des couleurs est binaire dans un espace de faible dimension : elle ne peut représenter que $2^{11} = 2048$ vecteurs distincts. Nous proposons d'extraire des règles d'association pour identifier des relations de la forme « couleur et ... et couleur $\Rightarrow$ émotion », en appliquant l'algorithme Apriori classique.

Extraction de présences et d'absences significatives Nous évaluons la qualité des règles par le critère du risque relatif RR [14], calculé comme le quotient entre la fréquence d'occurrence de l'émotion parmi les images étiquetées par les couleurs indiquées dans la prémisse de la règle et sa fréquence parmi les images ne possédant pas ces annotations. Une valeur supérieure à 1 indique une présence significative, une valeur inférieure à 1 indique que l'émotion est plus fréquente en l'absence de ces couleurs qu'en leur présence, ce qui correspond à une règle dite d'absence, de la forme « absence de couleurs ⇒ émotion ».

Résultats expérimentaux Les résultats obtenus montrent que toutes les règles ont un support faible, *joy* est la seule émotion suffisamment fréquente pour que ses co-occurrences avec plusieurs termes chromatiques aient un support supérieur au seuil fixé [10]. Le tableau 2 résume les règles obtenues telles que $RR > 1.4$, en ne conservant qu'une seule règle concluant à *joy*, ainsi que les règles telles que $RR < 0.7$.

Les résultats sont compatibles avec un savoir commun, montrant que les données spontanées collectées à partir de Flickr reflètent des relations couleurs–émotions couramment admises, par exemple entre *black* et *fear*, *red* et *anger* ou *green* et *trust*. Des règles moins attendues établissent des associations intéressantes, par exemple entre *pink* et *surprise*. Les règles d'absence sont de même compatibles avec l'intuition, caractérisant par exemple *joy* par l'absence de *black*, et contiennent des règles moins attendues, associant par exemple l'absence de *blue* et *anger*.

La combinaison des règles d'absence et présence permet d'identifier des relations fortes. Ainsi *green* apparaît comme hautement typique de *trust* : sa présence caractérise exclusivement cette émotion, qui n'est associée à aucune autre couleur. Au contraire, la présence et l'absence de *black* sont liées à plusieurs émotions, ce qui peut être interprété comme une richesse émotionnelle de la couleur. Enfin, il est intéressant de noter que toutes les couleurs n'apparaissent pas : certaines, comme *grey*, sont trop rares et leurs combinaisons ne dépassent jamais le seuil de support fixé. Pour d'autres, comme *white*, leur absence signifie qu'elles n'ont pas d'association privilégiée avec quelque émotion que ce soit. Ce résultat en négatif constitue également une caractérisation intéressante de la couleur.

TABLE 2 – Risque relatif de quelques règles de présence (notées P) « couleur et ... et couleur ⇒ émotion » ou d'absence (notées A) « absence de couleur ⇒ émotion ». « *rgyb* » désigne {*red green yellow blue*}

RR	*sadness*	*anger*	*fear*	*surprise*	*trust*	*joy*
red	-	P2.00	-	-	-	-
green	A0.65	-	-	-	P1.51	-
yellow	A0.63	-	A0.70	-	-	-
black	P1.57	P1.61	P2.46	A0.69	-	A0.64
blue	-	A0.64	-	-	-	-
pink	-	-	-	P1.47	-	-
rgyb	-	-	-	-	-	P1.72

5.3 Établissement de relations entre couleurs objectives et subjectives

La section précédente a examiné les relations entre les 2 domaines, chromatique et émotionnel, au même niveau, subjectif. Cette section s'intéresse aux 2 niveaux, objectif et subjectif, dans un même domaine, chromatique, pour lier la couleur subjective à la couleur objective.

Cette tâche est liée au nommage de couleurs [19, 1, 3] qui vise à déterminer le terme linguistique approprié pour décrire une couleur. Toutefois elle considère des images réelles contenant plusieurs couleurs, de distributions et de positions variables, et non des échantillons de couleurs homogènes. Elle peut également être formulée comme l'identification de couleur saillante, définie ici comme la couleur perçue comme si significative qu'elle est incluse dans les annotations de l'image. La saillance est une notion exploitée pour la description des images, par exemple dans la définition de *sift*, ou au niveau des objets figurant sur l'image [16]. A notre connaissance, la saillance chromatique est un domaine moins exploré.

Protocole expérimental Afin de caractériser les couleurs subjectives, nous considérons les images associées à une unique étiquette chromatique, qui constituent environ 62% du corpus initial et excluons les couleurs associées à moins de 1 000 images. Le corpus contient donc 13 590 images pour une tâche de classification en 7 classes. Le même protocole d'équilibrage et de validation croisée que précédemment est mis en œuvre.

Nous étudions également le rôle de la position des couleurs dans l'image, pour tester l'hypothèse selon laquelle la couleur subjective dépend de la zone dans laquelle elle apparaît, et non seulement de sa fréquence : outre la représentation classique calculée à partir de l'image entière, nous considérons la zone centrale, définie comme le rectangle central dont l'aire est la moitié de l'aire de l'image. Elle correspond à l'hypothèse selon laquelle les bords de l'image correspondent à l'arrière-plan, dont les couleurs n'influencent pas les annotations chromatiques. Une troisième représentation considère que la couleur de l'arrière-plan produit du bruit, et qu'elle ne doit pas seulement être ignorée, mais utilisée pour corriger la représentation. Elle est définie comme la différence entre les histogrammes respectivement construits à partir de la zone centrale et des bords de l'image.

Résultats On observe, dans le tableau 3, que les tbc sont significativement supérieurs à celui d'un classifieur aléatoire, qui obtiendrait 14%. De plus, ils sont plus élevés que pour la prédiction des émotions, bien que le nombre de classes soit supérieur : ceci montre que, comme attendu, les relations entre les couleurs objective et subjective sont plus fortes que celles entre couleurs et émotions. On peut également noter que, quelle que soit la représentation des couleurs utilisée, la zone centrale obtient les meilleurs résultats : c'est elle qui, le plus souvent, est responsable de

TABLE 3 – Moyenne et écart-type du taux de bonne classification pour la prédiction de la couleur subjective.

	image entière	centre	différence
HSV128	46.0 ± 0.7	49.9 ± 0.4	49.0 ± 0.5
HSV36	46.8 ± 0.7	51.5 ± 0.4	50.7 ± 0.2
D500	46.0 ± 0.7	51.3 ± 0.4	50.0 ± 0.6

l'étiquette chromatique. On peut envisager d'expliquer les mauvais résultats de l'approche par différence par une inadéquation du mode de correction considéré.
Pour la zone centrale, les 2 représentations interprétées, HSV136 et D500, obtiennent des résultats comparables, alors que HSV128 est significativement moins performant. Ce résultat semble naturel puisque l'interprétation vise précisément à favoriser la corrélation entre la décomposition de l'espace chromatique et la couleur perçue. Ces résultats diffèrent de ceux obtenus pour les émotions : la couleur subjective est, comme attendu, plus liée à l'interprétation de la représentation chromatique que les émotions.

6 Conclusions et perspectives

Nous avons étudié les relations entre couleurs et émotions à partir d'un corpus basé sur des annotations spontanées et montré que le double passage entre les 2 domaines, chromatique et émotionnel, et les 2 niveaux, objectif et subjectif, est difficile. Les résultats sont meilleurs entre les 2 domaines pour un même niveau et plus encore dans un même domaine entre les 2 niveaux. Cette étude encourage la définition de représentations enrichies, subjectives mais non émotionnelles, pour faciliter la reconnaissance d'émotions. Une perspective vise à étudier un autre type d'enrichissement, basé sur des descripteurs chromatiques plus expressifs, notamment pour tenir compte d'harmonies ou de contraste de couleurs, tels que définis dans des études artistiques par exemple [12]. Une autre perspective vise à inclure la prise en compte de des contextes spécifiques, en particulier l'influence d'effets culturels, par exemple par le biais de la localisation géographique des utilisateurs lors de la construction du corpus. Une contextualisation sémantique, tenant compte des autres annotations des images, semble également pertinente, par exemple pour établir une relation entre *white* et *joy* en présence de *wedding*.

Références

[1] A. Aït Younes, I. Truck, and H. Akdag. Image retrieval using fuzzy representation of colors. *Soft Computing*, 11 :287–298, 2006.

[2] B. J. Babin, D. M. Hardesty, and T. A. Suter. Color and shopping intentions : The intervening effect of price fairness and perceived affect. *Journal of Business Research*, 56 :541–551, 2003.

[3] R. Benavente, M. Vanrell, and R. Baldrich. A data set for fuzzy colour naming. *Color Research and Application*, 31(1) :48–56, 2006.

[4] B. Berlin and P. Kay. *Basic color terms : their universality and evolution*. Center for the Study of Language and Information, 1969.

[5] N. Bianchi-Berthouze and T. Kato. K-DIME : an adaptive system to retrieve images from the web using subjective criteria. In *Proc. of DNIS*, 2000.

[6] C.-H. Chen, M.-F. Weng, S.-K. Jeng, and Y.-Y. Chuang. Emotion-based music visualization using photos. In *Proc. of the Advances in Multimedia Modeling*, pages 358–368. Springer, 2008.

[7] T. Clarke and A. Costall. The emotional connotations of color : a qualitative investigation. *Color Research and Application*, 33(5) :406–410, 2008.

[8] C. Colombo, A. Del Bimbo, and P. Pala. Semantics in visual information retrieval. *IEEE Multimedia*, 6(3) :38–52, 1999.

[9] H. Feng, M.-J. Lesot, and M. Detyniecki. Associating color with emotions based on social tagging. In *Proc. of KEER*, 2010.

[10] H. Feng, M.-J. Lesot, and M. Detyniecki. Using association rules to discover color-emotion relationships based on social tagging. In *Proc. of KES*, 2010.

[11] M. Hemphill. A note on adults' color-emotion associations. *Journal of Genetic Psychology*, 54 :275–281, 1996.

[12] J. Itten. *The art of color : the subjective experience and objective rationale of color*. John Wiley & Sons, 1997.

[13] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (IAPS) : Affective ratings of pictures and instruction manual. Technical Report A-8, University of Florida, 2008.

[14] P. Lenca, B. Vaillant, P. Meyer, and S. Lallich. Association rule interestingness measures : Experimental and theoretical studies. *Quality Measures in Data Mining*, 43 :51 – 76, 2007.

[15] Q. Li, S. Luo, and Z. Shi. Fuzzy aesthetic semantics description and extraction for art image retrieval. *Computers and Mathematics with Applications*, 2008.

[16] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters : Contrast based filtering for salient region detection. In *Proc. of CVPR*, 2012.

[17] R. Plutchik. *The psychology and biology of emotion*. HarperCollins College, 1994.

[18] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of experimental psychology : general*, 123(4) :394–409, 1994.

[19] L. Zhang, F. Lin, and B. Zhang. A CBIR method based on color-spatial feature. In *IEEE Region 10 Annual International Conference*, 1999.

Modeling sensory-motor behaviors for social robots

Alaeddine Mihoub [1, 2], Gérard Bailly [1], Christian Wolf [2]

[1]GIPSA-Lab, Speech & Cognition department, Grenoble, France
[2]LIRIS, Lyon, France

Abstract

Modeling multimodal perception-action loops in face-to-face interactions is a crucial step in the process of building sensory-motor behaviors for social robots or users-aware Embodied Conversational Agents (ECA). In this paper, we compare trainable behavioral models based on sequential models (HMMs) and classifiers (SVMs and Decision Trees) inherently inappropriate to model sequential aspects. These models aim at giving pertinent perception/action skills for robots in order to generate optimal actions given the perceived actions of others and joint goals. We applied these models to parallel speech and gaze data collected from interacting dyads. The challenge was to predict the gaze of one subject given the gaze of the interlocutor and the voice activity of both. We show that Incremental Discrete HMM (IDHMM) generally outperforms classifiers and that injecting input context in the modeling process significantly improves the performances of all algorithms.

Keywords

Social behavior model, HMMs, SVMs, cognitive state recognition, gaze generation

1 Introduction

The design of social robots/agents able to engage efficient and believable face-to-face conversations with human partners is still an open issue. Although this kind of communication is considered as one of the most basic and classic forms of communication in our daily life [1], it is a complex and sophisticated bi-directional multimodal phenomenon in which partners continually convey, perceive, interpret and react to the other person's verbal and co-verbal displays and signals [2]. Studies on human behavior has confirmed for instance that co-verbal features – such as body posture, arm/hand gestures , head movement, facial expressions, eye gaze– strongly participate in the encoding and decoding of linguistic, paralinguistic and non-linguistic information. Several researchers have notably claimed that these features are largely involved in maintaining mutual attention and social glue [3].

Human interactions are paced by multi-level perception-action loops [4]. Thus, social robots/agents aiming at monitoring a multimodal and natural communication should mimic the very aspects of this complex close-loop system. In concrete terms, the robot has to couple two principal tasks: (1) scene analysis and (2) behavior generation. A multimodal behavioral model is responsible for computing behavior generation given the scene analysis and the intended goals of the conversation.

Our goal is to train statistical multimodal behavioral model that learns by observation of human-human interactions i.e. that maps perception to action. In this context, we present and compare three different candidate models: the first one is based on Hidden Markov Models (HMMs) and models the evolution of joint perception/action features over time. The two others are standard classifiers (Support Vector Machines and Decision Trees) that perform direct mapping without any explicit sequential modeling.

The paper is organized as follows: The next section reviews the state-of-the art of trainable multimodal generation systems. The three models are introduced in section 3. Section 4 illustrates the application of our models on data collected in a previous experiment [5]. We analyze the impact of contextual data in section 5. Finally, we conclude in section 6.

2 Related Work

The analysis of multi-party interaction is an interdisciplinary domain spanning research not only in signal and image processing but also in social and human science involving sociology, psychology and anthropology [6]. In recent years, it is becoming an attractive research area and there is an increasing awareness about its technological and scientific challenges. Actually, automatic conversation scene analysis copes with several issues, including turn taking, addressing, activity recognition, roles detection, degree of engagement or interest, state of mind, personal traits and dominance. Several computational models have been proposed to predict or generate observed multimodal human behavior.

For instance, Otsuka et al. [7] proposed a Dynamic Bayesian Network (DBN) to estimate addressing and turn taking ("who responds to whom and when?"). The DBN framework is composed of three layers. The first one perceives speech and head gestures; the second layer generates gaze patterns while the third one estimates conversations regimes. While the first layer is observable,

the others are latent and should be estimated. In order to recognize individual and group actions, Zhang et al. [8] suggested a two layered HMM. The first layer estimates personal actions taking as input raw audio-visual data. The second one infers group actions taking into account the estimations of the first layer. A Decision Tree is used in [9] for automatic role detection in multiparty conversations. Based mostly on acoustic features, the classifier assigns roles to each participant including effective participator, presenter, current information provider, and information consumer. In [10], Support Vectors Machines have been used to rate each person's dominance in multiparty interactions. The results showed that, while audio modality remains the most relevant, visual cues contribute in improving the discriminative power of the classifier. More complete reviews on models and issues related to nonverbal analysis of social interaction can be found in [11] [12][13].

For multimodal behavior generation, several platforms have been proposed for virtual agents and humanoid robots. Cassel et al. [14] notably developed the BEAT system ("Behavior Expression Animation Toolkit") which processes textual input and generates convenient and synchronized behaviors with speech such as intonation, eye gaze and iconic gestures. The synthesized nonverbal behavior is assigned on the basis of a contextual and linguistic analysis that relies on a set of rules inspired from research on conversational social human behavior. Later, Krenn [15] introduced the NECA project ("Net Environment for Embodied Emotional Conversational Agents") which aims to develop a platform for the implementation and the animation of conversational emotional agents for Web-based applications. This system hosts a complete scene generator and has the advantage of providing an ECA with communicative attitudes (e.g. head nodes, eye brow raising) as well as non communicative attitudes (e.g. moving/walking in the scene, physiological breathing). Another major contribution of the NECA platform is Gesticon [16]. It consists of repository of predefined co-verbal animations and gestures that can drive both virtual and physical agents. "MAX", the "Multimodal Assembly eXpert" developed by Kopp and colleagues [17], interacts with humans in a virtual reality environment and collaborates with them in order to achieve some tasks. MAX is able to ensure reactive and deliberative actions via synthetic speech, facial expressions, gaze, and gestures. Most mentioned platforms have many similarities: multimodal actions are selected, scheduled and integrated according to rules-based configurations. The SAIBA framework [18] has been developed to establish a unique platform, unify norms and accelerate advancements in the field. It is organized into three main components: "Intent planning", "Behavior planning" and "Behavior realization". It's worth noticing that SAIBA offers only a general framework for building multimodal behavioral models. In fact, the modeling within each component and its internal processing is treated as a "black box" and it is to researchers to fill the boxes by specifying their own models. One missing aspect of SAIBA is the perception dimension. In [2] a specific representation of perceptual cues was introduced to fill this gap. Many systems have adopted the SAIBA framework, particularly the GRETA platform [19] and the SmartBody system [20].

In the next section we will present our proposed models that, unlike pre-mentioned rule-based models (BEAT, SAIBA, etc), rely on machine learning and statistical modeling to intrinsically associate actions and percepts and to organize sequences of percepts and actions into so-called joint sensory-motor behaviors.

3 Social behavior modeling

This section presents statistical/probabilistic approaches for modeling jointly multimodal sensory-motor behaviors. Thus, these models should enable an artificial agent (1) estimate its cognitive state from perceptual observations (e.g. speech activity/gaze fixations of the partner). This state should reflect the joint behaviors of the conversation partners at that moment; (2) generate suitable actions (e.g. its own gaze fixations) that should reflect its current cognitive state and its current awareness of the evolution of the shared plan.

Each situated conversation is controlled by a specific syntax that defines a particular sequencing of joint cognitive states by a sort of behavioral grammar. As matter of fact, we chose HMM because they have intrinsic sequential and temporal modeling capabilities. We compare here their performance with those of two well-known powerful classifiers (SVMs and Decision Trees).

3.1 HMMs

For each dyad, we model each cognitive state with a single Discrete Hidden Markov Model (DHMM) and the whole interaction with a global HMM, that chains all single models with a task-specific grammar. The hidden states of these HMMs model the perception-action loop by capturing joined behaviors. In fact, the observations vectors are composed by two streams: the first stream contains the perceptual observations and the second stream observes actions. The "hidden" states are then sensory-motor. In the training stage, all data are available while in testing only perceptual observations are available. After training, two sub-models are thus extracted: a recognition model that will be responsible of estimating sensory-motor states from perceptual observations and a generation model that will generate actions from these estimated states. In our model, these two phases of decoding and generation are performed incrementally using a modified version of the Short-Time Viterbi algorithm [21]. Since observations here have

discrete values, we called this model IDHMM (for Incremental Discrete HMM). For more details about the IDHMM model see [22].

3.2 SVMs and Decision Trees

SVMs and Decision Trees are among the most used and powerful classifiers. In our context, we will train two distinct classifiers: the first one will estimate the most likely cognitive state from perceptual observations while the second one will directly determine the most likely actions from perceptual observations.

4 Application to a Face-to-face interaction

4.1 Experimental setting

The dataset used has been collected by Bailly et al. [5]. The setting is shown in Figure 1. It consists of speech and gaze data from dyads playing a speech game via a computer-mediated communication system that enabled eye contact and dual eye tracking. The gaze fixations of each one are estimated by positioning dispersion ellipsis on fixation points gathered for each experiment after compensating for head movements. The speech game involved an instructor who reads and utters a sentence that the other subject (respondent) should repeat immediately in a single attempt. Dyads exchange Semantically Unpredictable Sentences (SUS) that force the respondent to be highly attentive to the audiovisual signals. The experiment was designed to study adaptation: one female main speaker called "LN" interacted with eight subjects (females) both as an instructor for ten sentences and as a respondent for another set of ten sentences.

4.2 Data and models

For each dyad, we have two observations streams: voice activity (*v1/v2* with 2 modalities: on/off) and gaze fixations (*g1/g2* with 5 regions of interest ROI: face/mouth/left eye/right eye/else) of both speakers. Seven cognitive states [23] (CS) have been labeled semi-automatically ('Read', 'Prephon', 'Speak', 'Wait, 'Listen', 'Think' and 'Else'). For SVMs and Decisions Trees, a first classifier is used to estimate the CS of the principal subject LN from (*v1*, *v2*, *g2*). Then a second classifier is used to estimate his gaze (*g1*) from the same data. Similarly for the IDHMM, the recognition model is used to estimate the CS from (*v1*,*v2*,*g2*) and the eye fixations (*g1*) are synthesized using the generation model.

Gaze data have been monitored by two Tobii® eyetrackers operating at 25Hz. Voice activity detection has been sampled at the same rate.

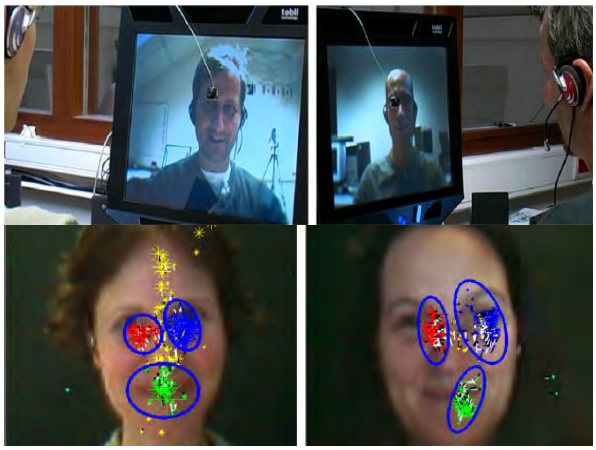


Figure 1: Experimental setting

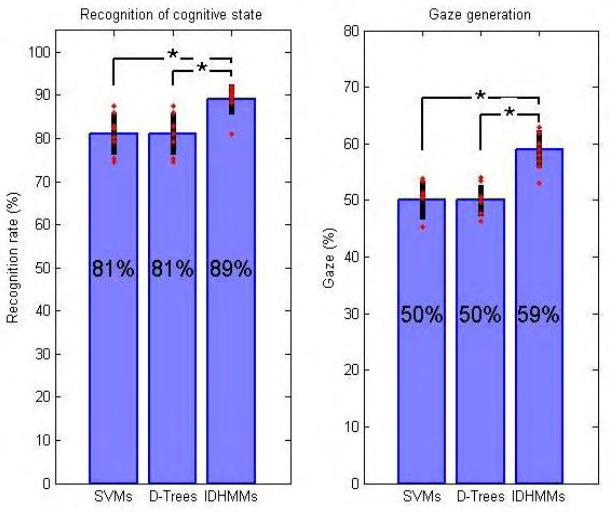


Figure 2: Results of the three models: SVMs, Decision Trees and IDHMMs

4.3 Results and comparison

DHMMs are trained with HTK [24], the IDHMM model was implemented in Matlab using PMTK3 toolkit [25]. For SVMs/Decision Trees, the Weka java package [26] has been used for both training and testing. For all models, 8-fold cross validation was applied: 7 subjects have been used for training while the eighth for testing.

Accuracy rates are used to evaluate cognitive state recognition, where the Levenshtein distance [27] is adopted for the evaluation of gaze generation because it allows more adequate comparison between generated and original signals. In fact, the Levenshtein distance is a metric for measuring the difference between two sequences; it computes the minimum number of elementary operations (insertions, deletions and substitutions) required to change one sequence into the other. From this optimal alignment, recall, precision and their harmonic mean (the F-measure)

can be directly computed. In this paper all generation rates represent F-measures.

Figure 2 clearly shows that there is no significant variation between the two classifiers. However, the IDHMM model outperforms the two classifiers and the improvement provided by this model is quite significant ($p<0.05$). The IDHMM model has a rate of 89% for cognitive state detection and 59% for eye gaze generation. Moreover Figure 5 shows that the IDHMM model is more efficient in detecting the structure of the interaction. We can see that the estimated path of cognitive states reflects correctly the predefined syntax of the task. In comparison, the SVMs has more difficulty in capturing the organization of the real path (see Figure 5) and discards short transition states: we can see that the estimated states are principally « Speak », « Wait » and « Listen ». This is in not in contradiction with the 81% recognition rate because these three cognitive states alone represent 85% of the ground truth. This performance gap is mainly due to the sequential constraints afforded by HMMs. This lack of sequential organization impairs the performance of SVMs and Decision Trees that should exclusively exploit bottom-up information provided by the observations.

5 Models with contextual attributes

5.1 New models

Classifier performance can be improved by adding memory (historical values) to each observation. In fact at a time t, the initial models use only the data of that moment. In the new configuration, we added the same three attributes (*v1*,*v2*,*g2*) but from a previous instant *t-T*, *T* being the size of the memory. Moreover we have varied this sole instant *T* from 1 frame to 80 frames to find the optimal delay.

5.2 Results and comparison

Our tests revealed that there is no significance difference between SVMs and Decision Trees, thus, in the rest we will focus on comparative performance of SVMs vs. IDHMMs. Figure 3 shows that the optimal delay for this task is $T= \sim 55$ frames (~ 2 seconds). We got the same value for Decision Trees. This optimal delay corresponds exactly to [28] in which authors demonstrate that, if a speaker looks at an object, 2 seconds after the listener will most likely be looking at the same object. From Figure 4, we can see that the addition of past observations results in better performance ($p<0.05$) for both SVM recognition (91%) and generation (59%). This memory injection leads also to a better modeling of the interaction structure. In fact, in Figure 5 we can obviously notice the improvement of the recognition of cognitive states.

Likewise, we added this past observation to the sensory stream of the IDHMM. As a result (Figure 4), we also observe a significant improvement in the gaze generation (59% to 63%) while the recognition rate remains the same at a 95% confidence level.

In the initial configuration, we concluded that IDHMM model was the most efficient due to the sequential property of Markov Models. In the second configuration, the results are generally improved; while the IDHMM is still better in gaze generation (63% vs. 59%), the SVM model leads to a higher rate (91% vs. 87%) for a 95% confidence level. Hence, supplying the SVM model with memory has relatively addressed the missing temporal aspect.

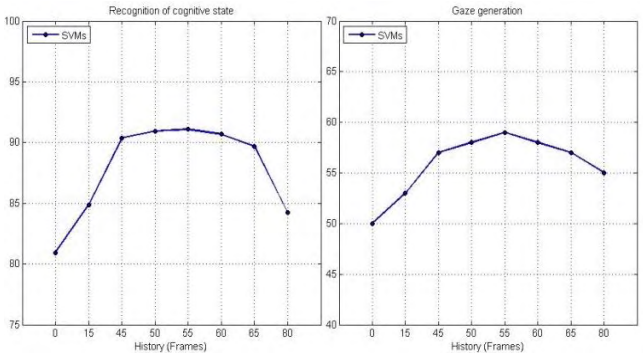


Figure 3: Optimal memory instant for the SVM

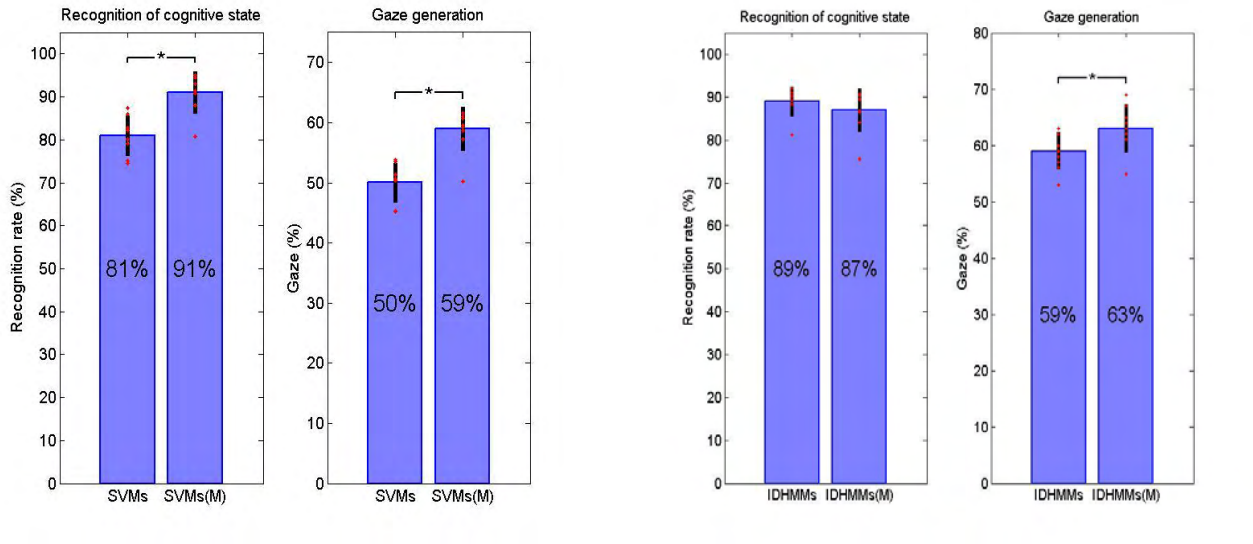
6 Conclusions

In this paper, we presented a comparative study of three behavioral models designed for social robots/agents (SVMs, Decision Trees and IDHMMs). These models have been tested in two different configurations: with & without history features. Comparison results showed that, in both settings, the IDHMM, thanks to its sequential modeling properties, remains a robust model for cognitive state recognition and eye gaze generation, and that classic classifier like SVMs could result in high performance if a certain memory (~2 seconds in our case) was included in the input observations.

Currently, we are studying a new scenario of a face-to face interaction that allows generating not only gaze but also deictic gestures. For the IDHMMs, we are also studying the influence of the number of hidden sensory-motor-states on the performance of each cognitive state and thus the impact on the generation figures.

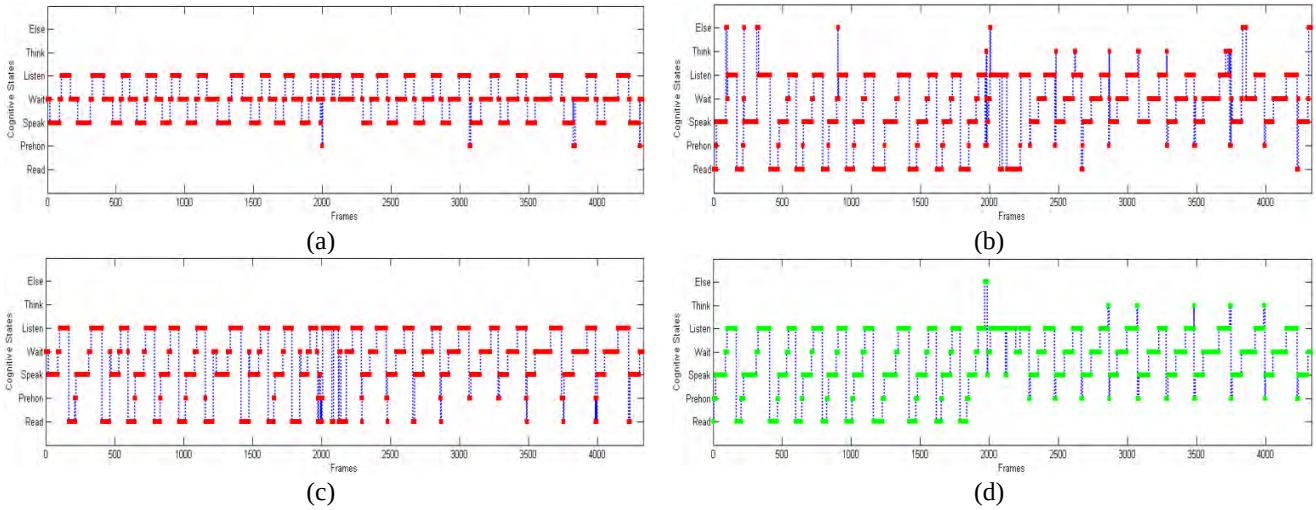
7 Acknowledgments

This research is financed by the Rhône-Alpes ARC6 research council.



(a) (b)

Figure 4: No memory / Memory (M=55) (a) for SVMs (b) for IDHMMs



(a) (b)

(c) (d)

Figure 5: Estimation of the cognitive state (CS) for a specific subject (a) using SVMs (b) using IDHMM (c) using SVMs and memory attributes (d) the real CS path

8 References

[1] K. Otsuka, « Multimodal Conversation Scene Analysis for Understanding People's Communicative Behaviors in Face-to-Face Meetings », p. 171‑179, 2011.

[2] S. Scherer, S. Marsella, G. Stratou, Y. Xu, F. Morbini, A. Egan, et L.-P. Morency, « Perception markup language: towards a standardized representation of perceived nonverbal behaviors », in *Intelligent Virtual Agents*, 2012, p. 455–463.

[3] J. L. Lakin, V. E. Jefferis, C. M. Cheng, et T. L. Chartrand, « The Chameleon Effect as Social Glue: Evidence for the Evolutionary Significance of Nonconscious Mimicry », *J. Nonverbal Behav.*, vol. 27, n° 3, p. 145‑162, sept. 2003.

[4] G. Bailly, « Boucles de perception-action et interaction face-à-face », *Rev. Franccaise Linguist. Appliquée*, vol. 13, n° 2, p. 121–131, 2009.

[5] G. Bailly, S. Raidt, et F. Elisei, « Gaze, conversational agents and face-to-face communication », *Speech Commun.*, vol. 52, n° 6, p. 598‑612, juin 2010.

[6] K. Otsuka, « Conversation Scene Analysis [Social Sciences] », *IEEE Signal Process. Mag.*, vol. 28, n° 4, p. 127‑131, 2011.

[7] K. Otsuka, H. Sawada, et J. Yamato, « Automatic inference of cross-modal nonverbal interactions in multiparty conversations: "who responds to whom, when, and how?" from gaze, head gestures, and utterances », in *Proceedings of the 9th international conference on Multimodal interfaces*, New York, NY, USA, 2007, p. 255–262.

[8] D. Zhang, D. Gatica-Perez, S. Bengio, et I. McCowan, « Modeling individual and group actions in meetings with layered HMMs », *Multimed. IEEE Trans. On*, vol. 8, n° 3, p. 509–520, 2006.

[9] S. Banerjee et A. I. Rudnicky, « Using simple speech–based features to detect the state of a meeting and the roles of the meeting participants », 2004.

[10] D. B. Jayagopi, H. Hung, C. Yeo, et D. Gatica-Perez, « Modeling dominance in group conversations using nonverbal activity cues », *Audio Speech Lang. Process. IEEE Trans. On*, vol. 17, n° 3, p. 501–513, 2009.

[11] D. Gatica-Perez, « Automatic nonverbal analysis of social interaction in small groups: A review », *Image Vis. Comput.*, vol. 27, n° 12, p. 1775–1787, 2009.

[12] D. Gatica-Perez, « Analyzing group interactions in conversations: a review », in *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, 2006, p. 41–46.

[13] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, et M. Schroeder, « Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing », *IEEE Trans. Affect. Comput.*, vol. 3, n° 1, p. 69‑87, 2012.

[14] J. Cassell, H. Vilhjalmsson, et T. Bickmore, *BEAT: the Behavior Expression Animation Toolkit*. 2001.

[15] B. Krenn, « The NECA project: Net environments for embodied emotional conversational agents », in *Proc. of Workshop on emotionally rich virtual worlds with emotion synthesis at the 8th International Conference on 3D Web Technology (Web3D), St. Malo, France*, 2003, vol. 35.

[16] B. Krenn et H. Pirker, « Defining the gesticon: Language and gesture coordination for interacting embodied agents », in *Proc. of the AISB-2004 Symposium on Language, Speech and Gesture for Expressive Characters*, 2004, p. 107–115.

[17] S. Kopp, B. Jung, N. Lessmann, et I. Wachsmuth, « Max - A Multimodal Assistant in Virtual Reality Construction », *KI*, vol. 17, n° 4, p. 11, 2003.

[18] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, et H. Vilhjálmsson, « Towards a Common Framework for Multimodal Generation: The Behavior Markup Language », in *INTERNATIONAL CONFERENCE ON INTELLIGENT VIRTUAL AGENTS*, 2006, p. 21–23.

[19] Q. A. Le et C. Pelachaud, « Generating Co-speech Gestures for the Humanoid Robot NAO through BML », in *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, E. Efthimiou, G. Kouroupetroglou, et S.-E. Fotinea, Éd. Springer Berlin Heidelberg, 2012, p. 228‑237.

[20] M. Thiebaux, S. Marsella, A. N. Marshall, et M. Kallmann, « Smartbody: Behavior realization for embodied conversational agents », in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, 2008, p. 151–158.

[21] J. Bloit et X. Rodet, « Short-time Viterbi for online HMM decoding: Evaluation on a real-time phone recognition task », in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, 2008, p. 2121‑2124.

[22] A. Mihoub, G. Bailly, et C. Wolf, « Social Behavior Modeling Based on Incremental Discrete Hidden Markov Models », in *Human Behavior Understanding*, Springer International Publishing, 2013, p. 172‑183.

[23] S. Baron-Cohen, *Mind Reading: The Interactive Guide to Emotions*, Édition : Cdr. London u.a.: Jessica Kingsley Publishers, 2004.

[24] HTK, The Hidden Markov Model Toolkit, *http://htk.eng.cam.ac.uk/*. .

[25] M. Dunham et K. Murphy, *PMTK3: Probabilistic modeling toolkit for Matlab/Octave, http://code.google.com/p/pmtk3/*. .

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten, « The WEKA data mining software: an update », *SIGKDD Explor Newsl*, vol. 11, n° 1, p. 10–18, nov. 2009.

[27] V. Levenshtein, « Binary Codes Capable of Correcting Deletions, Insertions and Reversals », *Sov. Phys. Dokl.*, vol. 10, n° 8, p. 707‑710, févr. 1966.

[28] D. C. Richardson, R. Dale, et K. Shockley, « Synchrony and swing in conversation: coordination, temporal dynamics, and communication », in *Embodied Communication in Humans and Machines*, I. Wachsmuth, M. Lenzen, et G. Knoblich, Éd. Oxford University Press, 2008, p. 75‑94.

Robots sociaux : design et recherche aux frontières de l'expérimentation

Mme OCNARESCU Ioana[1] Mme PAIN Frédérique [1]

[1]STRATE Ecole de design
i.ocnarescu@stratecollege.fr
f.pain@stratecollege.fr

Domaine principal de recherche: Experience design / UX, Robots sociaux, Design & Management
Papier soumis dans le cadre de la journée commune: NON

Résumé
Le poster décrit le processus d'évaluation, auprès d'une cible dite dépendante (âgées et/ou handicapés), des scénarios d'usage d'un robot social. Nous présentons dans un premier temps l'apport méthodologique du croisement de la méthode de design avec les protocoles expérimentaux de la recherche en robotique. Dans un deuxième temps nous proposons pour l'étape de l'évaluation un outil spécifique issu de la culture design: les « scénarios d'intention ». Ce sont des pré-scénarios d'usage qui utilisent des objets intermédiaires - prototypes non fonctionnels, et un medium propre au designer - la vidéo stop motion.

Mots Clef
Design, Approche design, Méthodologie, Robots sociaux, Scenarios d'usage.

Abstract
This poster describes the evaluation process of different social robots for dependent people (elderly and / or disabled). Firstly we present a crossing methodology that combines a design approach with research protocols and studies mostly coming from psychology and medical care. Secondly we propose a design tool to evaluate use-cases with social robots: the "intentional scenarios". They are pre-scenarios that use intermediate objects - non-functional prototypes, and a specific design medium for presentation - stop motion videos.

Keywords
Design, Design approach, Methodology, Social Robots, Use-cases.

1 Introduction
La recherche menée dans cette étude est le résultat d'une collaboration qui lie les chercheurs de la société Aldebaran, de Strate Ecole de Design, du laboratoire LIMSI et de l'association APPROCHE. Elle étudie les contributions potentielles du robot compagnon Romeo2, dont le postulat de base est d'en faire un robot adaptable aux besoins et attentes de l'utilisateur (ici les personnes dépendantes).

L'objet du projet ROMEO2 est de tester en situation réaliste un robot humanoïde de grande taille pour l'assistance aux personnes dont l'âge ou un handicap physique ou cognitif leur fait perdre une part d'autonomie. Le robot pourra, lors de ces périodes, être une présence permanente pour aider à la réalisation de tâches simples mais aussi à une certaine stimulation cognitive ; voire même jusqu'à parler ici d'intelligence émotionnelle.

De point de vue académique, l'objectif du projet est d'enrichir les connaissances sur les robots sociaux et de tester différentes hypothèses et tendances de la robotique, plus précisément de la robotique humanoïde. Est-ce que le robot doit prendre des initiatives, le contrôle ? Est-ce que la relation homme-robot doit se limiter à certaines activités ? Est-ce que la forme humanoïde est appropriée pour les tâches proposées? Loin d'avoir les réponses à ces questions, le design arrive dans ce projet comme un acteur qui a comme mission de travailler plutôt la forme du robot dans un premier temps. Cependant le rôle du designer va évoluer quand il fera rentrer la recherche en design dans le projet. Etat "un pont entre l'abstraction de la recherche et les besoins concrets de la vie réelle"[1], le design dans le projet Romeo2 va questionner les limites de la robotique humanoïde et sociale et va proposer de construire d'autres alternatives.

2 Court état de l'art sur la robotique sociale pour les personnes dépendantes
La littérature actuelle en robotique sociale et plus précisément dans la robotique dédiée aux personnes dépendantes, propose en 2013 avec la conférence « Social robots – the 5th international conférence on Social Robotics (ICSR2013) » les différents modèles théoriques pour analyses et proposer une relation homme-robot d'*accompagnement,* sinon de *camaraderie* (en anglais *'companionship'*)[2]. Certaines recherches présentent des résultats issus des focus groups, workshops, questionnaires et interviews pour comprendre des thématiques comme *la fonctionnalité, la sécurité, l'opérabilité*, *le soutien mutuel* – une thématique nouvelle et prometteuse, et *l'apparence* [3], d'autres se focalisent sur la *stimulation cognitive*, *la collaboration*[4], les limites et l'évolution de l'engagement dans l'interaction entre les robots et les personnes âgées[5]. L'aide physique pour des tâches de tous les jours (habillage, transfère, transport, etc.) ainsi que les rappels, la stimulation cognitive et la collaboration sont des éléments importants pour créer de l'engagement entre des personnes âgées et leurs robots. Le robot peut être aussi un médiateur social, aider les personnes à être autonomes et leur donner confiance

pour se développer et s'enrichir. Ces études mettent au centre l'utilisateur. Elles créent de la connaissance sur *le pouvoir d'anticipation* et *les attentes* de cette cible vis-à-vis des systèmes complexes comme les robots de compagnie.

La critique que nous pouvons amener est liée au faible nombre d'études qui son faites in situ, en dehors d'un contexte prédéfini et contrôlé de l'expérimentation. Peu d'études étudient l'expérience directe entre des personnes âgées et ce type de systèmes sociaux dans leur cadre de vie. Hélas les normes éthiques et le développement technologique peuvent ralentir cette volonté d'observer des robots sociaux in-situ et dans l'interaction. Cependant les designers et leur méthodologie d'observation et de création de la connaissance peuvent amener des débuts de réponse à cette problématique. Le designer observe le geste, le non-dit, l'imaginaire[1] et dans une « conversation » avec son dessin [6], donne forme et fond à des propositions. Ainsi ces propositions s'intègrent subtilement dans la vie réelle. Cette démarche est décrite dans la section suivante. Nous présentons aussi le processus qui intègre plusieurs acteurs dans la création et la validation des scenarios d'usage pour des robots sociaux (voir **Figure 1**).

3 Le processus de recherche - focus sur l'étape de l'évaluation des scenarios d'usage

I. *Le brief* est le point de départ du projet pour Romeo2. Il a pour objectif de créer dans une approche pluridisciplinaire des scénarios d'usage d'un robot humanoïde au service des personnes âgées.

II. *L'observation* est une étape clé du processus. La création d'un premier 'focus group' amène des notions sur la modélisation des imaginaires des personnes handicapées et de la robotique sociale. Un travail d'observation design complète ces premières notions en leur donnant un sens plus large, en intégrant des valeurs sensibles pour les designers : la dignité humaine, un travail sur l'altérité, le rapport intime entre la personne âgée et son compagnon robot. Ces résultats de terrains ont été confrontés avec les modèles théoriques de la littérature présentés antérieurement. Pour plus de détails sur le processus d'observation des designers, le poster va présenter des exemples de prise de notes et les premières propositions de concept issues de cette phase.

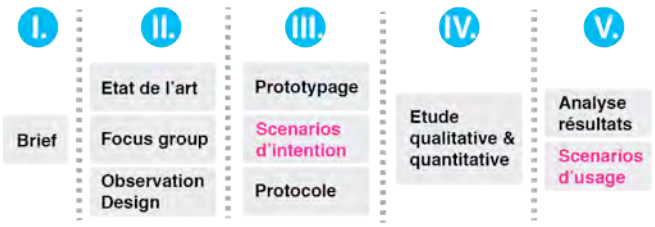


Figure 1 Le processus de définition et évaluation des scenarios d'usage des robots sociaux.

[1] L'imaginaire des « seniors de demain » est une problématique en soi ; cette cible a un rapport différent avec la technologie que les personnes âgées d'aujourd'hui qui participent à des études sur la robotique.

III. La troisième étape consiste dans le regroupement des connaissances pour la création des scenarios d'intention. Dans cette étape le design propose d'autres formes des robots sociaux comme Anubis, un robot construit en utilisant l'impression 3D. Différents objets intermédiaires sont utilisés pour la mise en place des scenarios : Anubis, Romeo2 en carton et Nao non-fonctionnel (voir **Figure 2**). Chaque objet est mis en scène en interaction avec une personne âgée en utilisant le medium de la vidéo en stop motion. Cette technique est choisie pour que les objets intermédiaires ne sont pas encore des prototypes fonctionnels. De plus ce langage cinématographique laisse un espace à l'interprétation et à la discussion qui est aussi le but des scenarios d'intention. Enfin ces scenarios sont une étape pour questionner et discuter des hypothèses recherche avec la population cible.

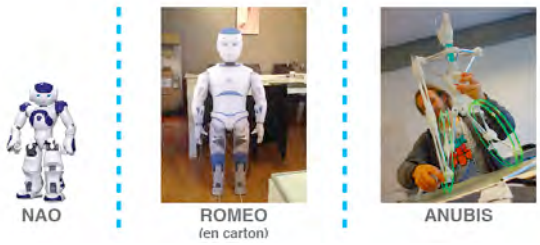


Figure 2 Les trois objets intermédiaires utilisés dans le processus de recherche sur les robots sociaux.

IV. En utilisant les scenarios d'intention comme support de discussion, nous avons effectué une étude qualitative (focus group) et quantitative (questionnaire en suivant la grille développée par [3]) auprès des 72 personnes âgées et/ou en situation de handicap. Nous sommes en train de finaliser cette étape. Les scenarios d'usage de Romeo2 (étape **V.**) vont être finalisés à partir de la synthèse des résultats de cette étude.

Le poster va expliquer le processus ici présent avec des exemples spécifiques à chaque étape ainsi que des premiers résultats issus de l'étude.

Bibliographie

[1] H. Aldersey-Williams, P. Hall, T. Sargent, and P. Antonelli, Design and the Elastic Mind. The Museum of Modern Art, New York, 2008.

[2] "Preface," in Social Robotics - 5th International Conference, ICSR 2013, Bristol, UK, October 27-29, 2013, Proceedings, Springer International Publishing, 2013.

[3] S. Frennert, H. Eftring, and B. Östlund, "What Older People Expect of Robots: A Mixed Methods Approach," in Social Robotics, vol. 8239, Springer International Publishing, 2013.

[4] Y. Li, K. Tee, S. Ge, and H. Li, "Building Companionship through Human-Robot Collaboration," in Social Robotics, vol. 8239, Springer International Publishing, 2013.

[5] S. Frennert, H. Eftring, and B. Östlund, "Older People's Involvement in the Development of a Social Assistive Robot," in Social Robotics, vol. 8239, Springer, International Publishing, 2013.

[6] B. Lawson, How Designers Think. Routledge; 4 edition, 2005.

Modélisation de l'influence de la personnalité d'un compagnon artificiel sur ses attitudes sociales

F. Pecune[1] C. Faur[2] M. Ochs[1] C. Clavel[2] C. Pelachaud[1] J-C. Martin[2]

[1] CNRS LTCI ; Télécom ParisTech
{pecune, ochs, pelachaud}@telecom-paristech.fr
[2] LIMSI CNRS
{faur, clavel, martin}@limsi.fr

Résumé

Les compagnons artificiels visent à être utilisés pour établir et maintenir une relation à long-terme avec leurs utilisateurs. Afin de rendre les interactions avec ces compagnons crédibles, un des éléments clés est de les doter d'une personnalité qui va elle-même influencer leur comportement social, et en particulier leur attitudes sociales. La personnalité peut être décrite comme un ensemble de comportements stables caractérisant un compagnon, et permet donc d'assurer une certaine cohérence lors des interactions. L'attitude sociale peut évoluer dans le temps et permet d'illustrer la relation qui lie le compagnon à l'utilisateur lors d'une interaction. Dans ce papier, nous proposons une modélisation de la personnalité et des attitudes sociales et en particulier de leur influence d'un point de vue cognitif.

Mots Clefs

Compagnons artificiels, personnalité, attitudes sociales, agents virtuels, modèle cognitif.

1 Introduction

Les machines occupent une place de plus en plus importante dans notre vie de tous les jours, et le nombre de travaux sur les compagnons artificiels n'a cessé d'augmenter durant ces dix dernières années. Selon [1], un compagnon peut être défini comme un robot ou un agent conversationnel animé (ACA) doté d'une autonomie, d'un certain niveau d'intelligence et de compétences sociales lui permettant d'établir et de maintenir des relations à long-terme avec les utilisateurs. L'intérêt grandissant pour les compagnons artificiels peut être expliqué par les travaux menés dans le domaine de la psychologie sociale. En effet, de nombreuses théories soulignent les besoins d'affiliation et l'impact des relations sociales sur le bien-être des individus [2]. Si les compagnons virtuels ne sont évidemment pas destinés à se substituer aux compagnons humains, les études menées par Turkle [3] montrent que les utilisateurs deviennent de plus en plus à l'aise à l'idée d'interagir avec ce genre d'interfaces.

Nos travaux de recherche ont pour but de développer un compagnon artificiel doté d'une personnalité propre, mais également d'un modèle cognitif d'attitudes sociales. A travers ce papier, nous étudierons également les liens qui existent entre ces deux notions. La combinaison de la personnalité et des attitudes sociales permettra au compagnon (1) de garder une certaine stabilité et une cohérence de son comportement sur le long-terme tout en (2) adaptant son comportement à la situation et au rôle qui lui est dévolu. En effet, un compagnon jouant le rôle d'un professeur aura un comportement différent, étant donné son attitude sociale, d'un autre compagnon incarnant un camarade de jeu. Par ailleurs, un compagnon doté d'une personnalité étourdie, stricte ou colérique gardera un comportement cohérent quelle que soit la situation. De plus, la personnalité du compagnon, à travers son influence sur l'état mental du compagnon, va fortement influencer l'attitude sociale de ce dernier.

L'article est organisé comme suit. Dans la section 2, nous dressons un état de l'art des différents travaux ayant trait à la personnalité et aux attitudes sociales. Nous présentons notre architecture en section 3 avant de décrire plus en détails les modèles de personnalité et d'attitudes sociales dans les sections 4 et 5. Nous illustrons notre modèle par un exemple décrit en section 6 avant de conclure et discuter de nos travaux futurs dans la section 7.

2 Background théorique et agents virtuels

Personnalité. La psychologie de la personnalité recouvre un large champ d'étude car le concept de personnalité peut être abordé à différents niveaux : le niveau de l'espèce humaine, le niveau des différences interindividuelles et celui des comportements propres à un individu [4]. En s'intéressant aux deux derniers niveaux, deux approches se distinguent : les approches de type "traits" et les ap-

proches dites "sociocognitives". Les traits de personnalité permettent de décrire une personne. Le modèle le plus utilisé dans le domaine des différences individuelles, en psychologie comme en informatique affective, est le Five Factors Model (FFM) [5]. Le modèle FFM se fonde sur cinq dimensions de la personnalité, qui sont aussi appelées les Big Five. Ces cinq dimensions sont : l'Ouverture à l'expérience, la Conscience, l'Extraversion, l'Agréabilité et le Névrotisme. Ces traits, dans leur ensemble ou certains en particulier, sont très souvent utilisés pour modéliser informatiquement des personnalités. Les Big Five sont utilisés pour influencer les motivations et la sélection de buts [6]. Ils sont également utilisés pour influer sur le comportement émotionnel d'entités virtuelles [7] ainsi que les comportements verbaux ou non-verbaux au cours d'une conversation [8].

L'approche socio-cognitive de la personnalité souligne l'importance de la situation lors de l'expression de comportements liés à la personnalité et tente de comprendre les processus cognitifs et sociaux qui conduisent à la personnalité. Mischel, contributeur majeur de cette approche, a conçu avec Shoda le modèle CAPS (Cognitive Affective Processing System) [9]. CAPS est un cadre métathéorique de la personnalité qui définit le système de la personnalité comme étant caractérisé par un réseau stable d'unités cognitivo-affectives, reliées entre elles par des liens d'activation et d'inhibition. Les comportements sont le résultat de la propagation de l'activation engendrée par les caractéristiques situationnelles au sein de ce réseau. En informatique affective, cette approche est moins utilisée mais néanmoins présente. Par exemple, l'approche socio-cognitive a été combinée avec les traits du Big Five pour guider le comportement d'agents virtuels autonomes [10] ou pour déterminer leur état émotionnel [11]. L'approche socio-cognitive se retrouve également dans le framework cognitivo-affectif proposé par Sandercock et al. [12] dans lequel la personnalité se développe en fonction de l'environnement.

Attitudes sociales. Bien qu'il existe différentes approches pour modéliser les attitudes sociales d'un agent virtuel, la méthode la plus répandue consiste à représenter ces dernières selon une ou plusieurs dimensions [13]. Parmi ces dimensions, la *dominance* et l'*appréciation* sont celles qui sont le plus souvent utilisées. L'appréciation peut être définie comme un sentiment général, positif ou négatif, à propos d'une personne [14]. Cette notion est asymétrique [15, 16] -et donc pas nécessairement réciproque. On peut également s'appuyer sur [15] pour définir la dominance comme la capacité d'un agent à influencer le comportement d'un autre agent. Cette influence est elle-même caractérisée par les ressources et les stratégies disponibles pour l'agent [17].

Dans [18], les auteurs dressent un état de l'art sur les agents relationnels et les différents domaines dans lesquels ils sont utilisés. Parmi ces agents, certains modélisent les relations sociales et plus précisément la dynamique de ces relations durant l'interaction. L'une des approches permettant la modélisation de cette dynamique est fondée sur les émotions ressenties par l'agent. Dans SCREAM [16], les émotions ressenties par l'agent jouent un rôle important, changeant la relation en fonction de leur valence et de leur intensité. Une émotion positive déclenchée par un autre agent va ainsi augmenter la valeur appréciation, alors qu'une émotion négative aura l'effet inverse.

D'autres modèles formalisant les relations sociales et leur dynamique selon des concepts logiques ont été proposés. Dans [19], les auteurs essaient de faire collaborer des humains et des agents virtuels en représentant formellement les dimensions de dominance et d'appréciation. L'évolution de ces deux dimensions repose sur le contenu des interactions entre les agents. Dans [20], Castelfranchi formalise les différents types de dépendance qui peuvent apparaître dans une relation. Un agent est dépendant d'un autre si ce dernier peut l'aider à réaliser l'un de ses buts. La valeur de dépendance peut varier si l'agent trouve des solutions alternatives, ou s'il réussit à créer un dépendance mutuelle ou réciproque.

Bien que la plupart de ces modèles se concentrent sur la dynamique des relations sociales, peu d'entre eux proposent d'initialiser ces relations de manière formelle. En effet, les valeurs d'appréciation et de dominance sont généralement fixées intuitivement en fonction du contexte de l'interaction. De plus, la plupart de ces modèles ne différencient pas la relation sociale calculée de l'attitude sociale finalement exprimée par l'agent.

3 État mental de l'agent

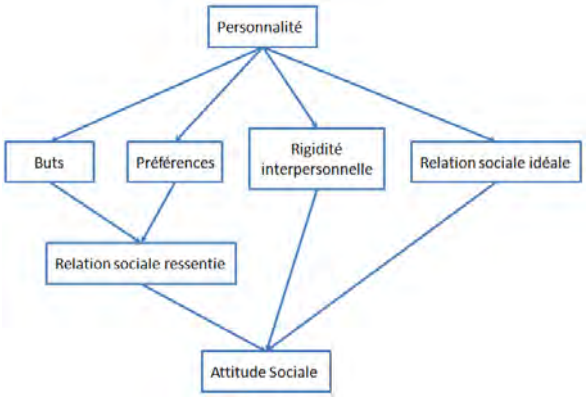


FIGURE 1 – État mental de l'agent décrivant l'influence de la personnalité sur les attitudes sociales de l'agent

Les travaux présentés en [21] mettent en lumière l'influence non-réciproque de la personnalité sur les relations sociales des individus (voir Figure 1). En effet, si la personnalité d'un compagnon peut permettre de prédire certains aspects de ses relations, l'inverse n'est pas vérifié : étudier la relation sociale d'un compagnon ne suffit pas à déterminer sa personnalité.

Dans nos travaux, nous proposons un modèle de personnalité basé sur une approche sociocognitive. Cette personnalité définit ainsi un ensemble d'éléments de l'état mental de l'agent tels que les buts, les préférences ou la rigidité interpersonnelle [22] (voir Section 4), notions étant elles mêmes au coeur du calcul des relations sociales et des attitude sociales d'un agent virtuel (voir Section 5).
Dans nos travaux, nous considérons des agents cognitifs ayant une représentation explicite de leurs buts et de leurs croyances. Comme dans [23], l'agent a un point de vue subjectif de son environnement, et ses croyances incluent des croyances à propos des autres agents, formant ainsi une théorie de l'esprit. Dans notre modèle, l'attitude sociale de l'agent dépend de ses propres buts et croyances. Par conséquent, cette attitude va évoluer dès lors que l'agent mettra à jour son état mental.

4 Modèle de personnalité

Le modèle PERSEED est un modèle de personnalité basé sur le Soi pour des compagnons artificiels. Le modèle PERSEED est basé sur différents travaux en psychologie ayant trait, d'une part, à l'approche sociocognitive de la personnalité et d'autre part, aux modèles de self-regulation. Une description complète du modèle et de ses fondements théoriques peut être trouvée dans Faur et al. [24]. Notre modèle s'articule autour de deux éléments : (1) un réseau de self-images et d'attributs et (2) des règles d'injection.

4.1 Self-images et attributs

Le compagnon dispose d'une collection de d'images de Soi ou self-images. Ces self-images sont liées à différents points de vue : le sien propre et ceux de personnes signifiantes (des personnes qui comptent dans l'environnement du compagnon). Par exemple, dans le cas de compagnons artificiels pour enfants, l'utilisateur principal, c'est-à-dire l'enfant, est signifiant. Nous pouvons également ajouter d'autres acteurs jouant un rôle important dans l'environnement de l'enfant, comme les parents de l'enfant. Pour chaque point de vue (POV), l'agent a un soi idéal (ce qu'il aimerait être) et un soi imposé (ce qu'il devrait être). Les self-images (idéal et imposé) et leur organisation par point de vue ont été directement inspirées par la *self-discrepancy theory* [25]. Ces sois sont reliés à des attributs, inspirés par les composants définis dans le *dynamic self-regulatory processing framework* [26]. Les attributs prennent la forme de (1) *connaissances sur le Soi* : buts et croyances concernant le Soi, (2) *processus de self-régulation intra-personnels* : modes d'évaluation particuliers, filtres perceptifs, règles d'attribution causale ou sensibilité aux émotions, et (3) *stratégies de self-régulation inter-personnelles* : mécanismes de sélection de buts et de planification, schémas spécifiques ou des préférences d'action.
La Figure 2 illustre cette organisation pour un compagnon servant de "copain de jeu". Dans cet exemple, nous postulons une architecture cognitive permettant de réaliser des

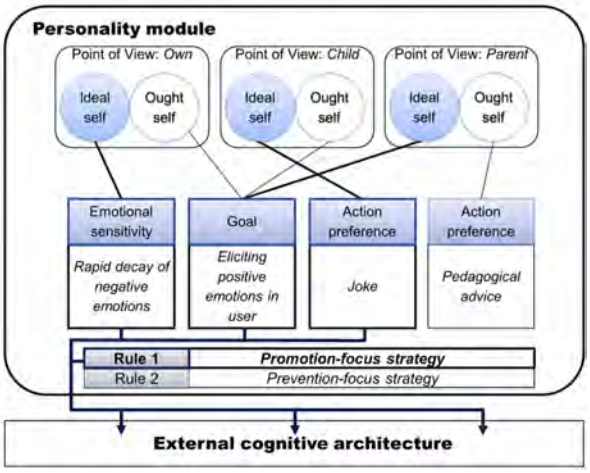


FIGURE 2 – Représentation du module de personnalité PERSEED injectant des attributs dans une architecture cognitive suivant une stratégie promotion-focus

actions et de percevoir l'état émotionnel de l'utilisateur et proposant une sélection d'action orientée buts et un modèle émotionnel pour l'agent. Le modèle PERSEED lui-même ne contient pas de contraintes quant à l'architecture cognitive spécifique du compagnon, c'est pourquoi nous laisserons le terme "architecture cognitive" non spécifié. Dans cet exemple, le compagnon a trois points de vue différents : le sien propre, le point de vue de l'enfant et le point de vue d'un parent de l'enfant. Par rapport à lui-même, l'agent voudrait avoir une faible sensibilité aux émotions négatives ; cette sensibilité est donc liée à son soi idéal. L'agent pense qu'il devrait susciter des émotions positives chez son utilisateur ; cet objectif est donc lié à son soi imposé. Ce dernier but est nécessaire du point de vue de l'enfant (soi imposé) et désirable pour le parent (soi idéal). L'enfant aimerait que l'agent fasse des plaisanteries (soi idéal) alors que pour le parent, l'agent doit profiter du jeu pour donner des conseils d'ordre pédagogique (soi imposé).

4.2 Règles d'injection

L'injection est un processus qui sélectionne les attributs et les transpose au sein de l'architecture cognitive, en altérant alors les fonctionnalités visées par les attributs. L'injection suit des règles définis dans le module de personnalité. Nous proposons d'utiliser dans un premier temps des règles inspirées par la *regulatory-focus theory* [27] : promotion-focus et prévention-focus. Ces règles correspondent aux façons d'utiliser les contenus des sois proposés par Higgins. Les personnalités de type *promotion-focus* utilisent préférentiellement leurs sois idéaux alors que celles de type *prévention-focus* utilisent préférentiellement leurs sois imposés. Le mode d'injection (c.-à-d. l'ensemble des règles utilisées pour l'injection) peut être fixé à l'avance ou défini par l'architecture cognitive sur la base de spécificités contextuelles. L'injection peut avoir lieu une seule fois, à

l'initialisation de l'agent ou être rejouée en fonction des retours fournis par l'architecture cognitive. Un agent présentant l'organisation illustrée par la Fig. 2 et utilisant une stratégie promotion-focus utilisera les attributs liés aux sois idéaux. La fonction de déclin des émotions négatives (du modèle émotionnel) sera remplacée par une nouvelle fonction fournie par le module de personnalité. Un but de haut-niveau concernant l'élicitation d'émotions positives chez l'utilisateur sera ajouté à la liste des buts. L'action "plaisanter" sera pondérée de façon à être plus facilement sélectionnée si l'occasion se présente.

5 Modèle d'attitudes sociales

Dans sa définition des attitudes sociales, Scherer dénote des origines pouvant être à la fois stratégiques et spontanées [28]. Nous nous appuyons donc sur cette définition pour modéliser les attitudes sociales de l'agent comme une combinaison de deux notions : la *relation ressentie* et la *relation idéale*. Dans cette section, nous expliquons en détail comment chaque dimension de l'attitude sociale (appréciation et dominance) est cognitivement représentée par un ensemble de croyances de l'agent, et comment la dynamique de cette attitude est modélisée.

5.1 Calcul de la relation ressentie de l'agent

Afin de représenter la dynamique de la relation ressentie de l'agent, une représentation à partir d'un réseau de neurones est proposée (voir Figure 3). Les noeuds du réseau correspondent aux différents buts et croyances de l'agent. Les liens permettent de représenter l'influence des différents noeuds entre eux.

Appréciation. Nous nous appuyons sur la *Balance Theory* de Heider [29] pour représenter formellement la dimension d'appréciation. Cette théorie peut être représentée comme un schéma triangulaire entre un agent A, un autre agent B et une entité C pouvant être un objet, un concept ou un troisième agent. Les arêtes du triangle décrivent l'appréciation de A envers B, l'appréciation de A envers l'entité C et l'appréciation de B envers cette même entité C. Selon [30] l'état entre les trois entités A, B et C est dit équilibré si les trois relations d'appréciation sont positives, ou si deux sont négatives et une positive. Par exemple, si un agent A apprécie un agent B et que tout deux apprécient le même concept C, l'état est équilibré. La *Balance Theory* indiquant que les gens ont tendance à vouloir atteindre des états équilibrés, on peut définir deux scénarios : (1) si A croit qu'il partage la même appréciation que B à propos d'une entité C, la valeur d'appréciation de A envers B va augmenter. (2) Au contraire, si A croit que leurs opinions divergent (l'un apprécie l'entité C et l'autre ne l'apprécie pas) alors la valeur d'appréciation de A envers B va diminuer. Pour modéliser l'influence de l'accord ou du désaccord sur la valeur d'appréciation, nous introduisons la notion d'*importance* accordée par l'agent A au concept C. Plus précisément, nous distinguons l'importance qu'accorde A au fait d'être d'accord avec B à propos d'un concept C, et l'importance accordée au désaccord avec B. Ainsi, si A croit qu'il est en désaccord avec B à propos d'un concept qu'il considère comme étant très important, la valeur d'appréciation de A envers B va grandement diminuer. Au contraire, si A croit qu'il partage la même opinion que B sur un concept qu'il considère peu important, la valeur d'appréciation de A envers B va légèrement augmenter.

Dominance. Notre modèle de dominance est fondé sur les travaux d'Emerson [31] et plus particulièrement sur sa définition de la *dépendance*. Pour Emerson, la dépendance d'un agent A à l'égard d'un autre agent B est (1) directement proportionnelle à l'importance accordée aux buts de A pouvant être influencés par B et (2) inversement proportionnelle au nombre d'agents K ayant une influence positive sur ces mêmes buts. Par exemple, l'agent A a pour but d'assister à un concert. Comme il n'a aucun moyen de s'y rendre. l'agent B propose à A de l'y conduire. Le niveau de dépendance de A à l'égard de B dépend ici de l'importance qu'accorde A au fait d'assister au concert. Cependant, si A connait un autre agent K pouvant lui prêter sa voiture, A aura une solution alternative, et sera donc moins dépendant envers B. Un agent pouvant avoir de multiples buts, son niveau de dépendance global correspond dans notre modèle à la valeur individuelle maximale de dépendance. Dans ses travaux, Emerson [31] définit également le pouvoir d'un agent A sur un autre agent B comme l'influence potentielle relative à la dépendance de A envers B mais également à la dépendance de B envers A. Dans notre modèle, un agent A est dominant par rapport à un agent B s'il croit que B est plus dépendant envers lui qu'il ne l'est lui-même envers B.

5.2 De la relation ressentie à l'attitude exprimée

Comme expliqué dans [28], l'attitude sociale exprimée par le compagnon est un mélange de spontanéité et de stratégie. Si l'aspect spontané est représenté par la relation ressentie du compagnon décrite ci-dessus, nous pouvons nous appuyer les différentes images de soi décrites en section 4.1 pour définir la *relation idéale* du compagnon. La relation idéale représente en effet la relation que l'agent souhaiterait idéalement exprimer dans une situation particulière (ex : un enseignant voudra montrer une forte dominance à ses élèves pendant la rentrée des classes). Dans notre modèle, l'attitude sociale exprimée par l'agent sera donc une combinaison entre la relation ressentie et cette relation idéale.

Afin d'obtenir l'attitude sociale finalement exprimée par l'agent, nous introduisons la notion de rigidité interpersonnelle [22], caractéristique découlant elle aussi de la personnalité. Selon cette théorie, les personnes faisant preuve d'une grande rigidité ont du mal à adapter leur comportement et leur attitude en fonction des situations. Dans notre modèle, cela se traduit par une plus forte influence de la relation ressentie sur l'attitude sociale exprimée, au détriment de la relation idéale. Par exemple, un compagnon hostile doté d'une rigidité interpersonnelle élevée ne montrera

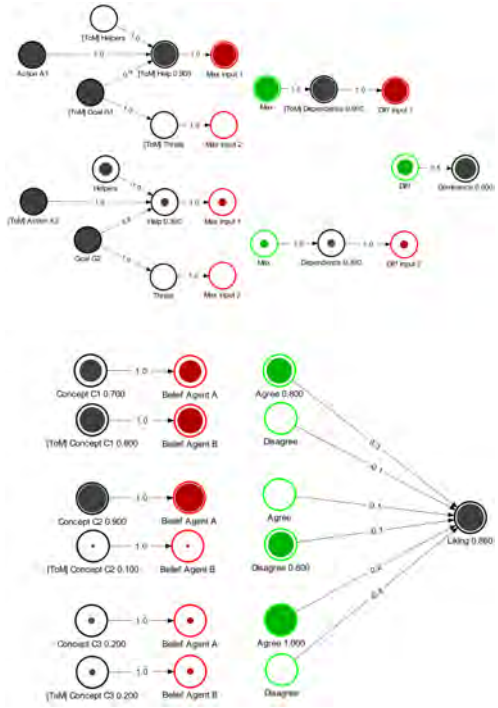


FIGURE 3 – Représentation des dimensions d'appréciation et de dominance selon les croyances et les buts du compagnon

pas de compassion ni de sympathie. Par contre, ce même agent doté d'une faible rigidité saura se montrer plus diplomate si le besoin s'en fait sentir.

6 Exemple de scénario

Afin d'illustrer l'influence de la personnalité sur les attitudes sociales, nous considérons l'exemple suivant en nous concentrant sur la dimension de dominance. Un compagnon C joue le rôle de tuteur pédagogique pour un enfant E qui doit faire ses devoirs de mathématiques en attendant le retour de ses parents. Le compagnon C a deux buts : le but que l'enfant acquière de nouvelles connaissances (B1) et le but que l'enfant termine ses devoirs (B2). Le compagnon croit que si l'enfant refuse d'apprendre quelque chose de nouveau, il ne peut pas l'y obliger. C a donc la croyance que E est le seul agent a pouvoir réaliser B1. C est donc dépendant de E pour la réalisation de B1. Le compagnon croit également que l'enfant a des difficultés en mathématiques et qu'en l'absence des parents de l'enfant, il est le seul à pouvoir l'aider à finir ses devoirs. C a donc la croyance que E est dépendant de lui pour réaliser B2.
Le modèle de personnalité du compagnon C comprend l'organisation suivante (pour son propre point de vue) : idéalement, le compagnon souhaite que l'enfant acquiert de nouvelles connaissances et il pense qu'il a l'obligation de faire en sorte que l'enfant finisse son travail. Le but B1 est donc relié au soi idéal de C et le but B2 est relié au soi imposé de C. Le compagnon pense également qu'un professeur veut et se doit de montrer de la dominance. La relation idéale (montrer de la dominance) de C est reliée à son soi idéal et à son soi imposé. Si le compagnon C a une personnalité promotion-focus, alors C va accorder une importance plus haute au but B1 (par rapport à B2). C a pour croyance qu'il est dépendant de E pour réaliser B1, sa relation ressentie a une valeur de dominance faible. Sa relation idéale étant de montrer une forte dominance, C va donc exprimer un dominance modérée. Si le compagnon C a une personnalité prevention-focus, aors C va accorder une haute importance au but B2 (par rapport à B1). C a pour croyance que E est dépendant de lui pour réaliser B2, sa relation ressentie a donc une valeur de dominance forte. Sa relation idéale étant également de montrer une forte dominance, C va exprimer une forte dominance.

7 Conclusion et Travaux futurs

Dans cet article, nous avons introduit un modèle représentant l'influence de la personnalité sur les attitudes sociales d'un compagnon artificiel. L'approche socio-cognitive de la personnalité permet de définir des caractéristiques spécifiques telles que les buts ou les préférences du compagnon. Notre modèle de personnalité permet également de définir deux différents types de règles venant influencer la prise de décision de l'agent. Les personnalités de type promotion-focus favorisent leurs sois idéaux là où les prévention-focus favorisent leurs sois imposés.
Les caractéristiques définies par la personnalité sont également utilisées pour représenter formellement les attitudes sociales du compagnon, modélisant l'influence de la personnalité sur ces mêmes attitudes. Ainsi, en fonction de ses préférences et de l'importance qu'il leur accorde, le niveau d'appréciation du compagnon envers l'utilisateur variera. Les buts de l'agent et l'importance qui leur est accordée permettent de définir le niveau de dominance de l'agent. Le calcul de ces deux dimensions passe par ailleurs par une représentation subjective des croyances de l'utilisateur, formant ainsi une théorie de l'esprit.
La prochaine étape consiste à modéliser de manière formelle l'influence de la personnalité sur les stratégies mises en place par le compagnon pour modifier son attitude sociale. Par exemple, lorsque le compagnon et l'utilisateur ne partagent pas le même avis sur un concept, le compagnon pourra, suivant sa personnalité (1) changer son propre avis sur le concept en question ou (2) essayer de faire changer l'avis de l'utilisateur afin d'obtenir un état équilibré. De même, un agent désirant augmenter sa dominance pourra (1) modifier ses propres croyances afin d'être moins dépendant ou (2) tenter de changer les croyances de l'utilisateur afin de le rendre plus dépendant.

Références

[1] Lim, M. : Memory models for intelligent social companions. Human-Computer Interaction : The Agency Perspective Studies in Computational Intelligence **396** (2012) 241–262

[2] Berscheid, E., Peplau, L. : The Emerging Science of Relationships. In : Close Relationships. W.H. Freeman, New York (1983) 1–19

[3] Turkle, S. : In Good Company : On the treshold of robotic companions. In : Close Engagements with Artificial Companions : Key social, psychological, ethical and design issues. John Benjamins (2010) 1–10

[4] Revelle, W. : Personality processes. Annual Review of Psychology **46** (1995) 295–328

[5] Costa, P.T., McCrae, R.R. : Four ways five factors are basic. Personality and Individual Differences **13**(6) (1992) 653–665

[6] Lim, M.Y., Dias, J., Aylett, R., Paiva, A. : Creating adaptive affective autonomous NPCs. Autonomous Agents and Multi-Agent Systems (2012) 1–25

[7] Signoretti, A., Feitosa, A., Campos, A.M., Canuto, A.M., Fialho, S.V. : Increasing the efficiency of NPCs using a focus of attention based on emotions and personality. In : Proceedings of the Brazilian Symposium on Games and Digital Entertainment (SBGAMES'10). (2010) 171–181

[8] Liu, K., Tolins, J., Tree, J.E.F., Walker, M., Neff, M. : Judging iva personality using an open-ended question. In : Intelligent Virtual Agents, Springer (2013) 396–405

[9] Mischel, W., Shoda, Y., Smith, R. : Introduction to Personality : Towards an Integration. 7 edn. John Wiley & Sons (2004)

[10] Poznanski, M., Thagard, P. : Changing personalities : towards realistic virtual characters. Journal of Experimental & Theoretical Artificial Intelligence **17**(3) (2005) 221–241

[11] Romano, D.M., Wong, A.K.L. : Personality model of a social character. In : 18th British HCI Group Annual Conference. (2004)

[12] Sandercock, J. : Using Adaptation and Goal Context to Automatically Generate Individual Personalities for Virtual Characters. PhD thesis, RMIT University, Melbourne (2009)

[13] Bickmore, T., Picard, R. : Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer-Human Interaction (TOCHI) **12**(2) (2005) 293–327

[14] Moshkina, L., Arkin, R. : On tameing robots. IEEE International Conference on Systems, Man and Cybernetics **4** (2003) 3949–3959

[15] Prada, R., Paiva, A. : Social Intelligence in Virtual Groups. In : New Advances in Virtual Humans. Volume 140. Springer (2008) 113–132

[16] Prendinger, H., Descamps, S., Ishizuka, M. : Scripting affective communication with life-like characters in web-based interaction systems. Applied Artificial Intelligence : An International Journal **16**(7-8) (2002) 519–553

[17] Raven, B. : The bases of power and the power/interaction model of interpersonal influence. Analyses of Social Issues and Public Policy **8**(1) (2008) 1–22

[18] Campbell, R., Grimshaw, M., Green, G. : Relational agents : A critical review. The Open Virtual Reality Journal **1** (2009) 1–7

[19] Prada, R., Paiva, A. : Teaming up humans with autonomous synthetic characters. Artificial Intelligence **173**(1) (2009) 80–103

[20] Castlefranchi, C., Miceli, M., Cesta, A. : Dependence relations among autonomous agents. ACM SIGOIS Bulletin **13**(3) (1992) 14

[21] Asendorpf, J.B., Wilpers, S. : Personality effects on social relationships. Journal of Personality and Social Psychology **74**(6) (1998) 1531

[22] Tracey, T.J. : Interpersonal rigidity and complementarity. Journal of Research in Personality **39**(6) (2005) 592–614

[23] Marsella, S.C., Pynadath, D.V., Read, S.J. : Psychsim : Agent-based modeling of social interactions and influence. In : Proceedings of the international conference on cognitive modeling, Citeseer (2004) 243–248

[24] Faur, C., Clavel, C., Pesty, S., Martin, J.C. : Perseed : A self-based model of personality for virtual agents inspired by socio-cognitive theories. In : Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, IEEE (2013) 467–472

[25] Higgins, E.T. : Self-discrepancy : A theory relating self and affect. Psychological Review **94**(3) (1987) 319–340

[26] Morf, C.C. : Personality reflected in a coherent idiosyncratic interplay of intra-and interpersonal self-regulatory processes. Journal of Personality **74**(6) (2006) 1527–1556

[27] Higgins, E.T. : Beyond pleasure and pain. American psychologist **52**(12) (1997) 1280–1300

[28] Scherer, K.R. : What are emotions ? and how can they be measured ? Social science information **44**(4) (2005) 695–729

[29] Heider, F. : The Psychology of Interpersonal Relations. Lawrence Erlbaum Associates Inc (1958)

[30] Zajonc, R.B. : The concepts of balance, congruity, and dissonance. Public Opinion Quarterly **24**(2) (1960) 280–296

[31] Emerson, R. : Power-dependence relations. American Sociological Review **27**(1) (1962) 31–41

Virtual Interactive Behavior : une architecture modulaire pour ACA

André-Marie Pez[1] Pierre Philippe[1] Brice Donval[2] Catherine Pelachaud[2]

[1]Institut Mines-Telecom
[2]LTCI – CNRS

andre-marie.pez@telecom-paristech.fr, pierre.philippe@telecom-paristech.fr,
brice.donval@telecom-paristech.fr, catherine.pelachaud@telecom-paristech.fr

1 Introduction

Après plus de dix ans de développement de Greta [1], des besoins nouveaux et la volonté de rendre son architecture plus adaptable ont imposé une restructuration profonde de la plateforme.

Aussi notre nouvelle approche est d'avoir une plateforme générique, que nous avons nommée Virtual Interactive Behavior (VIB), commune à toutes les technologies et utilisant une architecture composée de plusieurs modules apportant chacun une fonctionnalité particulière. Un Agent Conversationnel Animé (ACA) est ainsi défini par l'adjonction de différents modules.

Plusieurs de ces modules reprennent les algorithmes principaux de l'ancienne version de la plateforme Greta.

L'architecture VIB est compatible avec le standard SAIBA [2].

2 Architecture

Les actions que doivent effectuer les ACA sont décrites par différents formats allant d'un niveau cognitif (par exemple intentions communicatives) à un niveau physique (animation du squelette).

Ces formats de descriptions permettent de définir les entrées et les sorties des modules et sont transmis sous forme d'événements. Chaque module de VIB traite automatiquement les événements reçus ou en émet de nouveaux.

Outil « Modular »

Modular est un outil graphique facilitant l'instanciation et la connexion des modules dynamiquement.

Il permet de visualiser les entrées et sorties autour de l'agent : envoi de FML/BML, envoi de l'animation sur différents *players*, manipulation des données, fichiers log, etc. (figure 1).

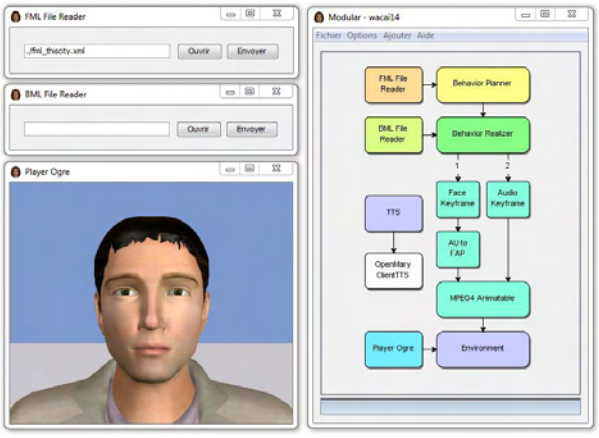


Figure 1. Interface graphique de Modular

3 Modules

Sont présentés ici les modules de base et certains modules que nous utilisons couramment.

3.1 Modules de base

Ces modules assurent une compatibilité avec le standard SAIBA, par le *Behavior Planner* et le *Behavior Realizer*, et gèrent ainsi les fichiers FML et BML.

En sortie du *Behavior Realizer*, des *keyframes performers* permettent de calculer en temps réel l'animation du corps et du visage.

Plusieurs modules sont utilisés pour la connexion à différents TTS (Text To Speech), comme OpenMary [3], Acapela [4], Cereproc [5].

La gestion de l'environnement virtuel est prise en compte et est découplée du *player*. Ce découplage permet d'avoir un *player* intégré utilisant Ogre3D et un *player* externe Unity3D.

3.2 Editeurs

Les bibliothèques de mouvements utilisées ne sont pas figées et peuvent être mises à jour en permanence. Pour cela, des modules « éditeurs » peuvent être employés. La plateforme possède des éditeurs pour définir les expressions du visage par Unités d'Action (AU) [6], les gestes [7][8], le mappage entre AU et FAP [9], la forme des mains, etc.

3.3 Communications réseau

Des modules de communications réseau ont été développés afin de communiquer avec des logiciels externes à la plateforme ou d'échanger les événements entre différentes instances de la plateforme réparties sur plusieurs machines. Actuellement, différentes API sont utilisées telles que ActiveMQ et Thrift [10].

3.4 Entrées utilisateurs

Comme entrée, outre les fichiers standards de SAIBA, nous nous servons également du logiciel SHORE [11] qui, par le biais d'une communication Thrift, nous permet de connaître les émotions exprimées par l'utilisateur en utilisant les entrées vidéo.

La reconnaissance vocale est une fonctionnalité de VIB qui est rendue possible grâce à l'API Web Speech de Google. Des modules additionnels pour VIB ont ainsi été développés afin de récupérer l'audio depuis le microphone et d'envoyer le fichier à l'API Web Speech. D'autres modules peuvent ensuite récupérer le texte reconnu afin de déclencher des réactions au sein de l'agent.

Plus qu'une simple entrée textuelle, Disco [12] est un gestionnaire de dialogue qui ajoute à l'interaction un arbre de dialogue combiné à des tâches hiérarchisées. Ce logiciel open-source améliore donc la structure du discours qui est automatisé.

3.5 Réseaux de neurones

Ce module donne la possibilité à l'utilisateur de créer des neurones et de les connecter via une interface graphique (figure 2).

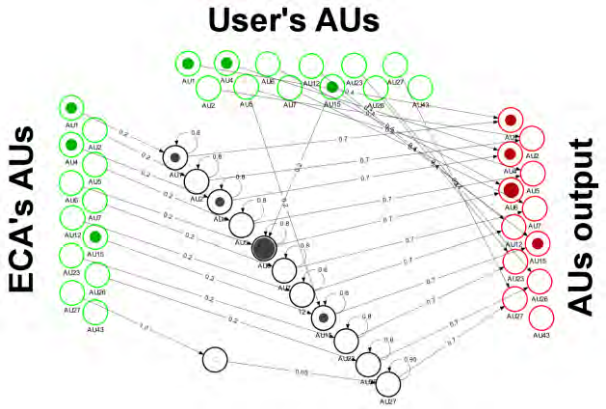


Figure 2. Utilisation d'un réseau de neurone

Ces neurones sont de type McCulloch & Pitts [13] et leur fonction d'activation est linéaire. La figure 2 présente un exemple d'utilisation : ce réseau combine les AU de l'utilisateur et de l'ACA pour obtenir une résonance motrice en temps réel des expressions.

4 Travaux en cours

4.1 Amélioration des modules actuels

Les réseaux de neurones sont à améliorer (fonction d'apprentissage, fonction d'activation non-linéaire), tout comme les modules liés aux communications avec des programmes externes.

4.2 Interactions entre ACA dans un même environnement

La volonté de faire interagir plusieurs ACA dans VIB a initié la création d'un gestionnaire d'environnements, agrégeant plusieurs environnements locaux en un environnement global dans lequel des agents de VIB peuvent venir se brancher à volonté. Chaque agent reste ainsi installé sur sa machine hôte, tandis que le module gestionnaire d'environnements de VIB connecte l'environnement de cet agent à l'ensemble des environnements qu'il gère. Chaque modification de l'environnement sur une machine hôte sera synchronisée avec tous les autres environnements. De la même façon chaque émission de FAP et de BAP sera répercutée sur tous les hôtes, permettant ainsi à chaque agent d'interagir dans un environnement commun.

Des ajouts de nouveaux modules sont possibles et prévus, dépendant des travaux de chacun et suivant les projets en cours.

Remerciements

Nous remercions toute la Greta team du laboratoire LTCI, en particulier Ken Prepin.

Bibliographie

[1] R. Niewiadomski, E. Bevacqua, M. Mancini & C. Pelachaud, Greta: an interactive expressive ECA system, *Proceedings of The 8th International Conference on Autonomous Agents and Multi-agent Systems*, pp. 1399-1400, 2009.
[2] E. Bevacqua, K. Prepin, E. de Sevin, R. Niewiadomski & C. Pelachaud, Reactive behaviors in saiba architecture, *Workshop Towards a Standard Markup Language for Embodied Dialogue Acts - AAMAS'09*, 2009
[3] M. Schröder & J. Trouvain, The German text-to-speech synthesis system MARY:a tool for research, development and teaching, *International Journal of Speech Technology 6.4*, pp. 365–377, 2003.
[4] Acapela Group, http://www.acapela-group.com
[5] Cereproc, https://www.cereproc.com
[6] P. Ekman, & W. V. Friesen, The repertoire of nonverbal behavior: Categories, origins, usage, and coding, *Semiotica 1*, pp. 49–98, 1969.
[7] A. Kendon, Gesture: Visible action as utterance, *Cambridge University Press*, 2004.
[8] D. McNeill, Hand and Mind: What Gestures Reveal About Thought, *The University of Chicago press*, 1992.
[9] ISO/IEC 14496, MPEG-4 International Standard, Moving Picture Experts Group, http://mpeg.chiariglione.org/standards/mpeg-4
[10] Apache Software Foundation, https://projects.apache.org/indexes/category.html#network-server
[11] J. Wagner, F. Lingenfelser & E. André, The social signal interpretation framework (SSI) for real time signal processing and recognition, *Interspeech*, pp. 3245–3248, 2011.
[12] C. Rich & C. Sidner, Using Collaborative Discourse Theory to Partially Automated Dialogue Tree Authoring, *14th Int. Conf. on Intelligent Virtual Agents*, 2012.
[13] W. S. McCulloch, W. Pitts, A Logical Calculus of the Ideas Immanent in Nervous Activity, *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115-133, 1943.

Machine Learning for Interactive Systems : Challenges and Future Trends

Olivier Pietquin[1] Manuel Lopes[2]

[1] Université Lille 1 - LIFL (UMR 8022 CNRS/Lille 1) - France
[2] Inria Bordeaux Sud-Ouest - France

olivier.pietquin@univ-lille1.fr - manuel.lopes@inria.fr

Abstract

Machine learning has been introduced more than 40 years ago in interactive systems through speech recognition or computer vision. Since that, machine learning gained in interest in the scientific community involved in human-machine interaction and raised in the abstraction scale. It moved from fundamental signal processing to language understanding and generation, emotion and mood recognition and even dialogue management or robotics control. So far, existing machine learning techniques have often been considered as a solution to some problems raised by interactive systems. Yet, interaction is also the source of new challenges for machine learning and offers new interesting practical but also theoretical problems to solve. In this paper, we address these challenges and describe why research in machine learning and interactive systems should converge in the future.

Keywords

Machine learning, interactive systems

1 Introduction

Communication between humans involves complex signals such as speech, gestures, facial expressions, body movements, written texts, *etc.* These signals convey high-level information such as semantics, emotions, context but can take highly variable forms. For instance, there might be many acoustic realisations for one word sequence, many word sequences for one meaning, *etc.* To enable machines to interact with humans in a natural manner, this variability has to be handled. On another hand, machine learning is the branch of artificial intelligence that addresses the problem of learning intelligent behaviours from data. To deal with communicative signal variability, machine learning has naturally been introduced very early in Human-Machine Interaction (HCI). The first and probably major achievement of machine learning in HCI is the introduction of Hidden Markov Models (HMM) in Automatic Speech Recognition (ASR) in the mid 70's [32, 31] which remains the standard method for ASR today. At the same time, data driven methods for text-to-speech synthesis (TTS) were developped [63]. It is only much later than machine learning has been exploited as a mean to interpret higher-level information such as semantics [66] or facial expression [53]. As for ASR and TTS, high-level analysis methods also gave rise to new synthesis methods like data-driven language generation [81].

In this paper, we are interested in machine learning methods intervening at a higher level : interaction management. Indeed, building an interactive system is not only about putting together all these input and output processing modules. There is a need for a intermediate module for sequencing the interaction. Taking past inputs and outputs into account, the interaction manager is in charge of deciding what should be the next system output. The interaction manager is probably one of the latest components of an interactive system that benefited from machine learning techniques. In the late 90's, the spoken dialogue management has been cast into a sequential decision making problem that could be solved by machine learning methods sucs as reinforcement learning [91] have been introduced with the aim of making the interaction more natural in a measurable way [44]. This seminal work led to many other applications of reinforcement learning to Spoken Dialogue Systems (SDS) [87, 71, 43] but also to other types of interacting systems such as tutoring applications [30, 69, 51], museum guides [95], car driving assistance [76], recommender systems [26] and even robotics bar tenders [22].

In the following, we address the different challenges arising when taking the sequential nature of interaction into account. We first describe how interaction can be seen as a sequential decision making problem in Section 2. We than explain why and how this decision making problem has been extended to handle partial observability in Section 3. After 15 years of research in this area, these methods have proven to be efficient in finding good interaction strategies but not to be efficient in terms of data. Data sparsity thus remains a problem addressed in Section 4. Thanks to improvement in data efficiency, there has been a lot of work to enable systems to learn online, from interactions. In Section 5, paradigms to improve efficiency by actively learn new skills will be presented. Recently, going even further active learning a new trend of research emerged : imitation learning. This will be explained in Section 6. From this,

we will see that interaction provides now totally new problems to machine learning and we will summarize these in Section 7 before coming to our conclusion.

2 Interaction as a sequential decision making process

Interaction management is actually about deciding on what to do in a given context, knowing that this context will be influenced by the decision. It is thus a sequential decision making problem where local decisions influence future ones and the quality of the interaction. To optimize this process, planning algorithms [23] were first proposed. Yet, planning makes a lot of assumptions such as being able to enumerate all the possible contexts or knowing transition probabilities between states given actions. Also, the objective has to be known in advance so that the optimal path in the graph can be computed. Once the plan is computed, it can hardly be modified even though the interaction goes wrong.

The machine learning answer to the sequential decision making optimisation problem is Reinforcement Learning [91]. Although this has been studied for a long time [2], it's only in the 90's that is has been applied to real world problems because of the so called *curse of dimensionality*. In this paradigm, an agent (e.g. interactive systems) faces a dynamic system (e.g. humans) that steps from states to states as an effect of the actions of the agent. The agent therefore learns to perform the sequence of actions that makes the system go through desired states. To assess the quality of a state, the agent perceives rewards after each action it performs in the environment. It thus tries to follow a path in the state space that offers the best cumulative reward. If one assumes that human-machine interaction is a turn-taking process (which is a strong assumption which is more and more contested in incremental systems [88]), than interaction management becomes such a sequential decision making problem.

Using reinforcement learning requires casting the task into the Markov Decision Processes (MDP) paradigm [2]. An MDP is formally a t-uple $\{S, A, R, T, \gamma\}$ where S is the state space, A is the action space, $R : S \rightarrow \mathbb{R}$ is the reward function, $T : S \times A \rightarrow \mathcal{P}(S)$ is a set of Markovian transition probabilities and γ is a discount factor to be defined later. The optimisation of the decision making problem consists in finding a policy $\pi : S \rightarrow \mathcal{P}(A)$ that maps states to actions in such a way that the cumulative rewards obtained by following this policy is maximized. To do so, the quality of a policy is measured in every state as the expected cumulative reward that can be obtained by following the policy starting from this state. This measure is called the value function $V^\pi : S \rightarrow \mathbb{R}$:

$$V^\pi(s) = E\left[\sum_{i=0}^{\infty} \gamma^i R(s_i) | s_0 = s, a_i = \pi(s_i)\right] \quad (1)$$

One can define an order on value functions such as $V^1 > V^2$ if $\forall s\ V^1(s) > V^2(s)$. The optimal policy π^* is the one that maximizes the value function for every state : $\pi^* = \arg\max_\pi V^\pi$. Many algorithms have been proposed in the literature to attempt at solving this problem [91], especially when the transition probabilities are not known, and this is still an active research area.

To optimize human-machine interaction management within this framework, one has to cast this task into an MDP. This has been first proposed in the late 90's [45]. The state space is the set of all possible interaction contexts and actions are the communicative acts the system can perform. The transition probabilities are usually unknown and several definitions for the reward function can be found in the literature. It is generally argued that the user satisfaction should be used as a reward [90] which can be approximated as a linear combination of objective measures that can be gathered during the interaction [97]. Yet, this reward is most often a very simple handcrafted function [45, 71, 98]. To define such a reward which is very task-dependent. If the system is devoted to goal-oriented dialogues, social chat, emotion control etc. it of course has to be different.

3 Partial Observability and non-Markovian processes

The MDP framework makes several strong assumptions. For instance, the dialogue contexts cannot be perfectly observed due to the recognition error introduced by the speech and the semantic analysers. The task is therefore non-Markov in the observation space. To meet the Markov assumption made by the MDP framework, the underlying states have to be inferred from observations using what is called a belief tracker. For example, the *Hidden Information State* [99] paradigm builds a list of the most probable current situations given the past observations, which is supposed to be a Markovian representation allowing for taking decisions in the MDP framework.

To take into account the perceptual aliasing problem introduced by error-prone speech and language understanding modules, Partially Observable MDP (POMDP) have been proposed to model the dialogue management task [83] and the tutoring task [79]. Yet, solving the POMDP problem requires the transition and observation models to be known which also requires a lot of assumptions and engineering work. There has been a lot of work to make this approach tractable and suitable for learning online making this approach very promising [98, 15].

There has been some attempts to either learn a Markov state representation online [16] or to learn a policy without making the Markov assumption [17].

4 Data sparsity

Data sparsity is often a problem because of the difficulty of collecting data. To alleviate this problem, interaction simulation based on user modeling [86, 67, 46] together with error modeling (ASR *etc.*) [75, 70, 94] is most often

used to artificially expand training datasets. However, the learnt strategies are sensible to the quality of the user model which is very difficult to assess [85, 74].

An alternative to this bootstrapping method is to use generalization frameworks adapted to RL such as approximate dynamic programming. Although this idea was first proposed very early [3] it took a long time before it has been studied in the field of reinforcement learning [27, 42, 91]. Because it was very new in machine learning at the time RL was first introduced in interactive systems, very few attempts to apply generalization methods in the framework of interaction management can be found in the literature. In [29], the authors use the SARSA(λ) algorithm [91] with linear function approximation which is known to be sample inefficient. In [47], LSPI [42] is used with feature selection and linear function approximation. Recently, Fitted Value Iteration (FVI) [27] have also been applied to dialogue management [8, 73]. All these studies report *batch* learning of dialog policies from fixed sets of data and thus learn in an *off-policy* manner, meaning that they learn an optimal policy from observations generated with another policy (which is mandatory for learning from fixed sets of data). It also means that, once a strategy is learnt from these datasets, it doesn't evolve anymore while, of course, one cannot expect to have a representative enough dataset for complex tasks.

5 Online and active learning

To alleviate the problem of incompleteness and inconsistency of data collected offline, online learning of dialogue management strategies has recently been made possible. Examples are Gaussian Processes [24], Natural Actor Critic [35] or Kalman Temporal Differences [72]. The two formers [24, 35] report the use of *online* and *on-policy* algorithms which requires permanently changing the policy to be learnt (an issue known as the dilemma between exploration and exploitation). These changes to the policy made during learning are visible to the user which may cause problems in real applications at the early stage of learning where the changes in the policy can lead to very bad behaviors of the dialogue manager. Thus, user simulation is still required. The later [72] makes possible *online* and *off-policy* learning which means that the system can learn online by observing a non-optimal policy in action (e.g. an hand-crafted safe but suboptimal strategy) which makes easier to deal with the trade-off between exploration (try new actions) and exploitation (use what was learn). To make online learning safer (to avoid the online learner to take very bad action), active learning has been proposed [14]. This method estimates the uncertainty about the outcomes of actions and decides to explore the most uncertain but promising actions. This approach has shown to perform very efficiently online in simulation [15] and in real world [25].

The exploration vs. exploitation dilemma is traditionally addressed by the bandit theory in the machine learning community [93]. This theory has gained in interest these few last years and has been very recently applied to interactive systems in tutoring [51] as well as in recommandation systems [48]. Yet it still relies on a trial-an-error process as RL.

Interactive robotics goes further by taking into account the ability of the human to provide information about how to perform the task. It has been suggested that *interactive learning*, human-guided machine learning or learning with human in-the-loop, might be a new perspective on robot learning that combines the ideas of learning by demonstration, learning by exploration, active learning and tutor feedback [19, 50, 18]. Under this approach the human user is considered as a teacher that interacts with the robot and provides extra feedback. Approaches have considered extra reinforcement signals [92], action requests [28, 52], disambiguation among actions [10], preferences among states [54], iterations between practice and user feedback sessions [34, 40] and choosing actions that maximize the user feedback [38, 39].

Another reason to use interactive systems to make the machine to learn from humans is that when the users train the system they might become more comfortable with using it and accept it. See the work from [62] for a study on this subject. The queries of the robot will have the dual goal of allowing the robot to deal with its own limitations and give the user information about the robot's uncertainty on the task being learned [21, 9].

6 Learning from Demonstrations

As sketched in the previous section, Learning from Demonstration (LfD) is also a recent avenue of research for interactive systems. Indeed, it is not a strong assumption to say that humans are experts in interaction that should be used as model for machines. Here again, several approaches can be envisioned.

First, one can cite Inverse Reinforcement Learning (IRL) [84, 59]. Many criticisms have been done to the reinforcement learning approach to interaction management [64, 65]. Especially, one criticism that has not been much addressed, is that these algorithms require providing the learning agent with a reward after each interaction. Although there have been attempts to define objective reward functions such as the PARADISE framework [97], this reward is indeed generally handcrafted by the system designer who introduces some expertise in the system [45, 71, 98] but also a strong bias. Very little attention has been paid to the particular problem of defining the best reward function for interactive systems.

Defining the appropriate reward function that will lead to a desired behavior is actually a real problem in the field of reinforcement learning. It is sometimes very hard to define in terms of mathematics although it is easy to demonstrate examples of optimal behaviors. Giving driving lessons is such a task where demonstrating a good behavior is easier than associating a reward to each couple of contexts and actions. Interaction management is also such a task since it

is very natural for human beings to interact with each other although it is much harder to isolate contexts and associate a reward to each possible action in these contexts.

Learning a reward function from demonstrations of the optimal behavior is known as the inverse reinforcement learning (IRL) problem. It is an ill-posed problem since the zero-reward is a solution whatever the expert policy (in other words, if you receive a zero-reward whatever you do, every policy is optimal). It also suffers from the same scaling-up problem as reinforcement learning when dealing with large state spaces.

IRL can be seen as a way to transfer the behavior of an expert to an artificial agent. It is of major importance in human-machine interaction where naturalness of the interaction is a desired feature. Indeed, since quantifying naturalness and user satisfaction is tricky, imitating the behavior of human operators can be a solution as suggested in [65]. This solution has been used to model human behaviours [6, 68] or for learning the reward of a dialogue system [11] but it is still very tricky to use because most algorithms suppose that the direct RL problem can be solved as many time as needed or that any number of random samples of interactions can be generated [1, 58, 36]. Yet, this is not true since solving the direct RL problem or gathering random data requires interacting with humans with whom the system cannot be random. New paradigms that do not make these assumptions have been recently proposed [37, 77] and applied to Embodied Conversional Agents applications [61]

Learning how to solve a task after seeing it being done has also been suggested has an efficient way to program robots. Typically, the burden of selecting informative demonstrations has been completely on the side of the teacher. Active learning approaches endow the learner with the power to select which demonstrations the teacher should perform. Several criteria have been proposed : game theoretic approaches [89], entropy [52, 55], query by committee [33], membership queries [56], maximum classifier uncertainty [10], expected myopic gain [13, 12] and risk minimization [20].

One common goal is to find the correct behavior, defined as the one that matches the teacher, by repeatedly asking for the correct behavior in a given situation. Such idea as been applied in situations as different as navigation [52, 55], simulated car driving [10] or object manipulation [52].

Directly under the inverse reinforcement learning formalism, one of the first approaches were proposed by [52]. The teacher can directly ask about the reward value at a given location [80] and it has been shown that reward queries can be combined with action queries [56]. Active inverse reinforcement learning can also be seen as a way to infer the preferences of the teacher. This problem of *preference elicitation* has also been addressed in several domains [5, 96, 4].

7 New challenges for machine learning

As shown before, interactive systems have many properties that require innovative machine learning techniques such as the sequential nature of interaction, the partial observability of inputs or the non-deterministic behaviour of users. Although these fields are still under active research (like imitation learning), there are many other big challenges brought by interactive systems to machine learning that will undoubtedly generate fundamental research in this field.

A first one is to clearly understand the theoretical properties of such systems. Machine learning has became a very theoretical field with time which can create a big gap between the interests of different communities. But on another hand, using machine learning in human-computer interaction requires theoretical proofs since empirical ones are hard to obtain. For instance, guarantees about security are often required before using robots in an inhabited area.

There are reasons to believe that an interactive perspective on learning from demonstration might lead to better results (even for the same amount of data). The theoretical aspects of these interactive systems have not been explored, besides the directly applied results from active learning. One justification for the need and expected gain of using such systems is discussed by [82]. Even if an agent learns from a good demonstration then, when executing that learned policy, its error will grow with T^2 (where T is the horizon of the task). The reason being that any deviation from the correct policy moves the learner to a region where the policy has a worse fit. If a new demonstration is requested in that new region then the system learns not only how to execute a good policy but also how to correct from small mistakes. Such observation, as the authors refer, was already given by [78] without a proof.

Having a human on the loop we have to consider the risks involved by a decision or the cost in terms of tiredness of making many queries in an interactive learning setting. Estimating risks in a sequential decision making process is a real machine learning challenge [57]. Studies and algorithms have also addressed the problem of deciding when to ask. Most approaches will just ask to user whenever the information is needed [60] or when there is high uncertainty [10]. A more advanced situation considers making queries only when it is too risky to try experiments [20].

Another challenge is to take into account the fact that human users may also change their behaviour with time. Its not only that this makes the environment of the machine-learning agent non-stationary but adversarial. Indeed, the users adapt their behaviour to the one of the machine which itself learns from the observations they make from the human behaviour. This *co-adaptation* phenomenon is very poorly addressed in the HCI literature [7] (although it is also known in brain-computer interaction [41]) but it is also not common in the machine learning community because

it brings very tricky problems to solve [49].
These are only few examples of unsolved challenges, but there are many others such as scalability, weakly supervised learning, transfer learning and so on.

8 Conclusion

In this paper, we described a list of challenges induced by interactive systems that were addressed by means of machine learning. Especially, we were interested in the problem of managing interactions which is intrinsically sequential. Although interactive systems were at the origin of major signal processing and machine learning achievements initially (as for HMMs), it became a consumer of machine learning techniques in the last decades in the field of sequential decision making. It is now again a source of big challenges for the machine learning community and, especially, it offers a panel of killing applications that has the potential to increase the visibility of machine learning. For these reasons, we believe that linked between communities will be tighter than ever in the near future.

Références

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of ICML 2004*, page 1, 2004.

[2] R. Bellman. *Dynamic Programming*. Dover Publications, sixth edition, 1957.

[3] R. Bellman and S. Dreyfus. Functional approximation and dynamic programming. *Mathematical Tables and Other Aids to Computation*, 13 :247–251, 1959.

[4] E. Brochu, N. de Freitas, and A. Ghosh. Active preference learning with discrete choice data. In *Advances in Neural Information Processing Systems*, 2007.

[5] U. Chajewska, D. Koller, and R. Parr. Making rational decisions using adaptive utility elicitation. In *National Conf. on Artificial Intelligence*, pages 363–369. Menlo Park, CA ; Cambridge, MA ; London ; AAAI Press ; MIT Press ; 1999, 2000.

[6] S. Chandramohan, M. Geist, F. Lefevre, O. Pietquin, et al. User simulation in dialogue systems using inverse reinforcement learning. *Proceedings of Interspeech 2011*, pages 1025–1028, 2011.

[7] S. Chandramohan, M. Geist, F. Lefevre, O. Pietquin, M.-I. Supelec, and F. Metz. Co-adaptation in spoken dialogue systems. In *Proceedings of the IWSDS 2012*, Ermenonville, France, 2012.

[8] S. Chandramohan, M. Geist, and O. Pietquin. Optimizing Spoken Dialogue Management with Fitted Value Iteration. In *Proceedings of Interspeech'10*, Makuhari (Japan), 2010.

[9] C. Chao, M. Cakmak, and A. Thomaz. Transparent active learning for robots. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE Inter. Conf. on*, pages 317–324, 2010.

[10] S. Chernova and M. Veloso. Interactive policy learning through confidence-based autonomy. *J. Artificial Intelligence Research*, 34 :1–25, 2009.

[11] H. R. Chinaei and B. Chaib-draa. Learning dialogue pomdp models from data. In *Proceedings of the 24th Canadian Conference on Advances in Artificial Intelligence*, Canadian AI'11, pages 86–91, Berlin, Heidelberg, 2011. Springer-Verlag.

[12] R. Cohn, E. Durfee, and S. Singh. Comparing action-query strategies in semi-autonomous agents. In *Inter. Conf. on Autonomous Agents and Multiagent Systems*, 2011.

[13] R. Cohn, M. Maxim, E. Durfee, and S. Singh. Selecting Operator Queries using Expected Myopic Gain. In *2010 IEEE/WIC/ACM Inter. Conf. on Web Intelligence and Intelligent Agent Technology*, pages 40–47, 2010.

[14] L. Daubigney, M. Gasic, S. Chandramohan, M. Geist, O. Pietquin, and S. Young. Uncertainty management for on-line optimisation of a POMDP-based large-scale spoken dialogue system. In *Proceedings of Interspeech 2011*, page 1301âĂŞ1304, Florence (Italy), August 2011.

[15] L. Daubigney, M. Geist, S. Chandramohan, and O. Pietquin. A Comprehensive Reinforcement Learning Framework for Dialogue Management Optimisation. *IEEE Journal of Selected Topics in Signal Processing*, 6(8) :891–902, December 2012. pdf.

[16] L. Daubigney, M. Geist, and O. Pietquin. Model-free POMDP optimisation of tutoring systems with echo-state networks. In *Proceedings of SIGDial 2013*, pages 102–106, Metz (France), August 2013.

[17] L. Daubigney, M. Geist, and O. Pietquin. Particle Swarm Optimisation of Spoken Dialogue System Strategies. In *Proceedings of Interspeech 2013*, Lyon (France), August 2013.

[18] R. Dillmann. Teaching and learning of robot tasks via observation of human performance. *Robotics and Autonomous Systems*, 47(2) :109–116, 2004.

[19] R. Dillmann, O. Rogalla, M. Ehrenmann, R. Zollner, and M. Bordegoni. Learning robot behaviour and skills based on human demonstration and advice : the machine learning paradigm. In *Inter. Symposium on Robotics Research (ISRR)*, volume 9, pages 229–238, 2000.

[20] F. Doshi, J. Pineau, and N. Roy. Reinforcement learning with limited reinforcement : using bayes risk for active learning in pomdps. In *25th Inter. Conf. on Machine learning (ICML'08)*, pages 256–263, 2008.

[21] T. Fong, C. Thorpe, and C. Baur. Robot, asker of questions. *Robotics and Autonomous systems*, 42(3) :235–243, 2003.

[22] M. E. Foster, S. Keizer, Z. Wang, and O. Lemon. Machine learning of social states and skills for multi-party human-robot interaction. In *Proceedings of the workshop on Machine Learning for Interactive Systems (MLIS 2012)*, page 9, Montpellier, France, 2012.

[23] R. Freedman. Atlas : A plan manager for mixed-initiative, multimodal dialogue. In *Proceedings of the AAAI-99 Workshop on Mixed-Initiative Intelligence*, pages 1–8. Citeseer, 1999.

[24] M. Gasic, F. Jurcicek, S. Keizer, F. Mairesse, B. Thomson, K. Yu, and S. Young. Gaussian processes for fast policy optimisation of pomdp-based dialogue managers. In *Proceedings of SIGDIAL'10*, Tokyo, Japan, 2010.

[25] M. Gašić, F. Jurčiček, B. Thomson, K. Yu, and S. Young. On-line policy optimisation of spoken dialogue systems via live interaction with human subjects. In *Proceedings of ASRU*, pages 312–317, 2011.

[26] N. Golovin and E. Rahm. Reinforcement learning architecture for web recommendations. In *Proceedings of the International Conference on Information Technology : Coding and Computing (ITCC 2004)*, volume 1, pages 398–402, Las Vegas, Nevada, USA, 2004.

[27] G. Gordon. Stable Function Approximation in Dynamic Programming. In *ICML'95*.

[28] D. Grollman and O. Jenkins. Dogged learning for robots. In *Robotics and Automation, 2007 IEEE Inter. Conf. on*, pages 2483–2488, 2007.

[29] J. Henderson, O. Lemon, and K. Georgila. Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Computational Linguistics*, 2008.

[30] A. Iglesias, P. Martínez, R. Aler, and F. Fernández. Learning teaching strategies in an adaptive and intelligent educational system through reinforcement learning. *Applied Intelligence*, 31(1) :89–106, 2009.

[31] F. Jelinek. *Statistical Methods for Speech Recognition*. Language, Speech and Communications Series. Mit Press, 1997.

[32] F. Jelinek, L. Bahl, and R. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21(3) :250–256, May 1975.

[33] K. Judah, A. Fern, and T. Dietterich. Active imitation learning via reduction to iid active learning. In *UAI*, 2012.

[34] K. Judah, S. Roy, A. Fern, and T. Dietterich. Reinforcement learning via practice and critique advice. In *AAAI Conf. on Artificial Intelligence (AAAI-10)*, 2010.

[35] F. Jurcicek, B. Thomson, S. Keizer, M. Gasic, F. Mairesse, K. Yu, and S. Young. Natural Belief-Critic : a reinforcement algorithm for parameter estimation in statistical spoken dialogue systems. In *Proceedings of Interspeech'10*, Makuhari (Japan), 2010.

[36] E. Klein, M. Geist, B. Piot, and O. Pietquin. Inverse reinforcement learning through structured classification. pages 1–9, South Lake Tahoe, Nevada, USA, 2012.

[37] E. Klein, B. Piot, M. Geist, and O. Pietquin. A cascaded supervised learning approach to inverse reinforcement learning. In H. Blockeel, K. Kersting, S. Nijssen, and F. Zelezny, editors, *Proceedings of ECML/PKDD 2013*, volume 8188 of *Lecture Notes in Computer Science*, pages 1–16, Prague (Czech Republic), September 2013. Springer.

[38] W. Knox and P. Stone. Interactively shaping agents via human reinforcement : The tamer framework. In *fifth Inter. Conf. on Knowledge capture*, pages 9–16, 2009.

[39] W. Knox and P. Stone. Combining manual feedback with subsequent mdp reward signals for reinforcement learning. In *9th Inter. Conf. on Autonomous Agents and Multiagent Systems (AAMAS'10)*, pages 5–12, 2010.

[40] P. Korupolu, V.N., M. Sivamurugan, and B. Ravindran. Instructing a reinforcement learner. In *Twenty-Fifth Inter. FLAIRS Conf.*, 2012.

[41] S. Koyama, S. M. Chase, A. S. Whitford, M. Velliste, A. B. Schwartz, and R. E. Kass. Comparison of brain–computer interface decoding algorithms in open-loop and closed-loop control. *Journal of computational neuroscience*, 29(1-2) :73–87, 2010.

[42] M. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 2003.

[43] O. Lemon and O. Pietquin. Machine learning for spoken dialogue systems. In *Proceedings of Interspeech'07*, pages 2685–2688, Anvers, Belgium, 2007.

[44] E. Levin, R. Pieraccini, and W. Eckert. Learning dialogue strategies within the markov decision process framework. In *Proceedings of ASRU 1997*, pages 72–79. IEEE, 1997.

[45] E. Levin, R. Pieraccini, and W. Eckert. Using Markov decision process for learning dialogue strategies. In *Proceedings of ICASSP 98*, volume 1, pages 201–204, Seattle, Washington, USA, 1998.

[46] E. Levin, R. Pieraccini, and W. Eckert. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, 8(1) :11–23, 2000.

[47] L. Li, S. Balakrishnan, and J. Williams. Reinforcement Learning for Dialog Management using Least-Squares Policy Iteration and Fast Feature Selection. In *InterSpeech'09*, Brighton (UK), 2009.

[48] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[49] M. L. Littman. Friend-or-foe q-learning in general-sum games. In *ICML*, volume 1, pages 322–328, 2001.

[50] A. Lockerd and C. Breazeal. Tutelage and socially guided robot learning. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ Inter. Conf. on*, volume 4, pages 3475–3480, 2004.

[51] M. Lopes, B. Clement, D. Roy, and P.-Y. Oudeyer. Multi-armed bandits for intelligent tutoring systems. *submitted to Journal of Artificial Intelligence in Education*, 2013.

[52] M. Lopes, F. S. Melo, and L. Montesano. Active learning for reward estimation in inverse reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD'09)*, 2009.

[53] K. MASE. Recognition of facial expression from optical flow. *IEICE transactions*, 74(10) :3473–3483, 1991.

[54] M. Mason and M. Lopes. Robot self-initiative and personalization by learning through repeated interactions. In *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI'11)*, 2011.

[55] F. S. Melo and M. Lopes. Learning from demonstration using mdp induced metrics. In *Machine learning and knowledge discovery in databases (ECML/PKDD'10)*, 2010.

[56] F. S. Melo and M. Lopes. Multi-class generalized binary search for active inverse reinforcement learning. *submitted to Machine Learning*, 2013.

[57] O. Mihatsch and R. Neuneier. Risk-sensitive reinforcement learning. *Machine learning*, 49(2-3) :267–290, 2002.

[58] G. Neu and C. Szepesvári. Training parsers by inverse reinforcement learning. *Machine learning*, 77(2-3) :303–337, 2009.

[59] A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of ICML 2000*, pages 663–670, Stanford, CA, USA, 2000.

[60] M. Nicolescu and M. Mataric. Learning and interacting in human-robot domains. *Systems, Man and Cybernetics, Part A : Systems and Humans, IEEE Transactions on*, 31(5) :419–430, 2001.

[61] R. Niewiadomski, J. Hofmann, J. Urbain, T. Platt, J. Wagner, B. Piot, H. Cakmak, S. Pammi, T. Baur, S. Dupont, M. Geist, F. Lingenfelser, G. McKeown, O. Pietquin, and W. Ruch. Laugh-aware virtual agent and its impact on user amusement . In *Proceedings of AAMAS2013*, pages 619–626, Saint Paul, USA, May 2013.

[62] T. Ogata, N. Masago, S. Sugano, and J. Tani. Interactive learning in human-robot collaboration. In *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ Inter. Conf. on*, volume 1, pages 162–167, 2003.

[63] J. P. Olive and N. Spickenagel. Speech resynthesis from phoneme-related parameters. *The Journal of the Acoustical Society of America*, 59(4) :993–996, 1976.

[64] T. Paek. Reinforcement learning for spoken dialogue systems : Comparing strengths and weaknesses for practical deployment. In *Proceedings of the Interspeech Dialog-on-Dialog Workshop (2006)*, 2006.

[65] T. Paek and R. Pieraccini. Automating spoken dialogue management design using machine learning : An industry perspective. *Speech Communication*, 50(8) :716–729, 2008.

[66] R. Pieraccini, E. Levin, and E. Vidal. Learning how to understand language. In *Proceedings of Eurospeech'93*, pages 1407–1412, 1993.

[67] O. Pietquin. Consistent goal-directed user model for realisitc man-machine task-oriented spoken dialogue simulation. In *Proceedings of ICME 2006*, pages 425–428, Amsterdam, Netherlands, 2006.

[68] O. Pietquin. Inverse Reinforcement Learning for Interactive Systems. In *Proceedings of the IJCAI workshop on Machine Learning for Interactive Systems (MLIS 2013)*, pages 71–75, Beijing (China), August 2013. Invited Speaker.

[69] O. Pietquin, L. Daubigney, and M. Geist. Optimization of a tutoring system from a fixed set of data. In *Proceedings of the ISCA workshop on Speech and Language Technology in Education*, pages 1–4, Venice, Italy, 2011.

[70] O. Pietquin and T. Dutoit. Dynamic bayesian networks for nlu simulation with applications to dialog optimal strategy learning. In *Proceedings of ICASSP 2006*, volume 1, pages 49–52, Toulouse, France, 2006.

[71] O. Pietquin and T. Dutoit. A probabilistic framework for dialog simulation and optimal strategy learning. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2) :589–599, 2006.

[72] O. Pietquin, M. Geist, and S. Chandramohan. Sample Efficient On-line Learning of Optimal Dialogue Policies with Kalman Temporal Differences. In *Proceedings of IJCAI 2011*, pages 1878–1883, Barcelona, Spain, July 2011. Oral Presentation.

[73] O. Pietquin, M. Geist, S. Chandramohan, and H. Frezza-Buet. Sample-Efficient Batch Reinforcement Learning for Dialogue Management Optimization. *ACM Transactions on Speech and Language Processing*, 2011.

[74] O. Pietquin and H. Hastie. A survey on metrics for the evaluation of user simulations. *Knowledge Engineering Review*, 28(01) :59–73, February 2013. first published as FirstView.

[75] O. Pietquin and S. Renals. ASR System Modeling For Automatic Evaluation And Optimization of Dialogue Systems. In *Proceedings of ICASSP 2002*, volume I, pages 45–48, Orlando, (USA, FL), May 2002.

[76] O. Pietquin, F. Tango, and R. Aras. Batch reinforcement learning for optimizing longitudinal driving assistance strategies. In *Proceedings of the IEEE Symposium on Computational intelligence in vehicles and transportation systems (CIVTS 2011)*, pages 73–79, Paris, France, 2011.

[77] B. Piot, M. Geist, and O. Pietquin. Boosted and reward-regularized classification for apprenticeship learning. In *Proceedings of AAMAS2014*, Paris (France), May 2014.

[78] D. Pomerleau. Neural network perception for mobile robot guidance. Technical report, DTIC Document, 1992.

[79] A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto. Faster teaching by pomdp planning. In *Artificial intelligence in education*, pages 280–287. Springer, 2011.

[80] K. Regan and C. Boutilier. Eliciting additive reward functions for markov decision processes. In *Inter. Joint Conf. on Artificial Intelligence (IJCAI'11)*, Barcelona, Spain, 2011.

[81] E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, 2000.

[82] S. Ross and J. A. D. Bagnell. Efficient reductions for imitation learning. In *13th Inter. Conf. on Artificial Intelligence and Statistics (AISTATS)*, May 2010.

[83] N. Roy, J. Pineau, and S. Thrun. Spoken dialogue management using probabilistic reasoning. In *Proceedings of ACL 2000*, pages 93–100. Association for Computational Linguistics, 2000.

[84] S. Russell. Learning agents for uncertain environments. In *Proceedings of COLT 1998*, pages 101–103, Madison, Wisconsin, USA, 1998.

[85] J. Schatzmann, M. N. Stuttle, K. Weilhammer, and S. Young. Effects of the user model on simulation-based learning of dialogue strategies. In *Proceedings of ASRU 2005*, December 2005.

[86] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The Knowledge Engineering Review*, 21(2) :97–126, June 2006.

[87] K. Scheffler and S. Young. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proceedings of HTL 2002*, pages 12–19, San Diego, Californie, USA, 2002. Morgan Kaufmann Publishers Inc.

[88] D. Schlangen and G. Skantze. A general, abstract model of incremental dialogue processing. In *Proceedings of EACL 2009*, pages 710–718, 2009.

[89] A. P. Shon, D. Verma, and R. P. N. Rao. Active imitation learning. In *AAAI Conf. on Artificial Intelligence (AAAI'07)*, 2007.

[90] S. Singh, M. Kearns, D. Litman, and M. Walker. Reinforcement learning for spoken dialogue systems. In *Proceedings of NIPS99*, 1999.

[91] R. Sutton and A. Barto. *Reinforcement Learning : An Introduction*. Cambridge Univ Press, 1998.

[92] A. Thomaz and C. Breazeal. Teachable robots : Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7) :716–737, 2008.

[93] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.

[94] B. Thomson, M. Gasic, M. Henderson, P. Tsiakoulis, and S. Young. N-best error simulation for training spoken dialogue systems. In *Proceedings of SLT 2012)*, pages 37–42. IEEE, 2012.

[95] S. Thrun, M. Beetz, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Haehnel, C. Rosenberg, N. Roy, et al. Probabilistic algorithms and the interactive museum tour-guide robot minerva. *The International Journal of Robotics Research*, 19(11) :972–999, 2000.

[96] P. Viappiani and C. Boutilier. Optimal bayesian recommendation sets and myopically optimal choice query sets. In *Advances in Neural Information Processing Systems*, 2010.

[97] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. PARADISE : A framework for evaluating spoken dialogue agents. In *Proceedings of EACL 1997*, pages 271–280. Association for Computational Linguistics, 1997.

[98] J. D. Williams and S. Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2) :393–422, 2007.

[99] S. Young, M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. The hidden information state model : A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2) :150–174, 2010.

Nonverbal behavior of a virtual agent expressing attitudes in a group

B. Ravenet[1] A. Cafaro[2] M. Ochs[2] C. Pelachaud[2]

[1] Institut Mines-Télécom ; Télécom Paristech ; CNRS LTCI
[2] CNRS LTCI ; Télécom Paristech

{ravenet,cafaro,ochs,pelachaud}@telecom-paristech.fr

Abstract

Embodied Conversational Agents have been widely used to simulate dyadic interactions with users. We want to explore the context of expression of interpersonal attitudes in simulated group conversations. We are presenting a model for conversational groups of humans and agents, where the agents are able to exhibit a variety of nonverbal behaviors (e.g gestures, facial expressions, proxemics) depending on the interpersonal attitudes that they want to express within the group while talking. The model combines corpus-based and theoretical-based approaches. A multi-layer framework to represent conversational group of humans and agents is also proposed. Finally, we present a preliminary implementation of this model. The capacity of the model to convey different virtual agents' interpersonal attitudes is illustrated on a scenario characterized in the proposed framework.

1 Introduction

While embodied conversational agents (ECAs) have been mainly studied in dyadic interaction settings, there is also a growing interest for small group situations. A dyadic interaction is a 2-interactant configuration (two persons speaking to each other), whereas a small group situation implies generally three to twenty interactants [3]. There have been several proposals for models simulating agents in groups but few of them have considered interpersonal attitudes and its impact on the nonverbal behavior as we propose in this paper. In our work, we base our theoretical assumptions on the literature of Human and Social Sciences about human-human interaction in groups. Based on this literature [17][11], we propose a model to represent small conversational groups of humans and agents (three to five participants), in which the agents are able to adapt and exhibit different nonverbal behaviors when talking, depending on the interpersonal attitudes that they want to express. Interpersonal attitude is an "affective style that can be naturally or strategically employed in an interaction with a person or a group of persons"[24]. We are using the representation from Argyle to manipulate agent's interpersonal attitudes [1]. An interpersonal attitude is represented on two axes, a status axis (ranging from submissive to dominant) and a liking axis (going from friendly to hostile).

In order to model the influence of such interpersonal attitudes on an ECA's nonverbal behavior, our approach is based on a combination of behavior models coupling a data-based model of conversational gestures and a rule-based model of group formation that simultaneously influence the ECAs' nonverbal behavior.

In the next section, we review related work about groups of ECAs and expression of attitudes. In the third section, we present our model of conversational groups for ECAs and users. In the fourth section, we present our combined model for social attitudes in small group conversations. In the fifth section, we describe a preliminary implementation of this model using the VIB platform [19], the Unity3D[1] game engine and the Impulsion AI Engine[2]. Finally, we provide further insights about our approach and some limitations.

2 Related Agents Work

One of the earliest works about groups of ECAs is [26]. They built a dialog model for multiple agents within a system called the Mission Rehearsal Exercice. This system was able to generate the nonverbal behavior of the agents following the script of given scenario. The relations between the characters were defined by their roles (E.g.: subordinates, teammates or superior). In a more recent work, the system was used to implement a model of group conversational nonverbal behavior [12] and group formation [13]. While these works attempted to model conversational groups as we are doing, they did not model the influence of the interpersonal attitudes on the nonverbal behavior exhibited.

In [9], the system Demeanour supported the design of virtual characters within a group with different social attitudes. The system generated different gaze, gesture and posture behavior but it did not manage the group formation dynamics.

In [21], they designed a group model similar to ours where an user took part in a task resolution interaction with a group of agents. This model is implemented within a game

[1] https://unity3d.com/
[2] www.impulsionproject.net

application and it is used to trigger the actions and reactions of the agents along with the evolution of their inner state but it does not deal with nonverbal behavior generation. As their model of interaction is clearly oriented for problem-solving interactions, in our model, we take a different approach by considering conversational interactions and we focus on the nonverbal behavior of conversational participants.

In [20], they have implemented a model that arranges virtual agents following Kendon's F-Formation. The agents are able to rearrange their formations and they glance at newcomers when they arrive in the group. However, ECAs' interpersonal attitudes have not been considered.

In [15] they designed the nonverbal behaviour of ECAs depending on their role in the conversation and their relations in a specific scenario set in an old western bar. The ECA interaction was scripted and their model was not dealing with group formation.

In a previous work, we collected a corpus of nonverbal behavior for an agent, to express different attitudes in a dyadic-conversation as speaker [22]. We have built a Bayesian model of nonverbal behavior for a talking agent from the collected corpus. This model was integrated with a dialog manager in a job interview simulation application [6]. The dialog manager was in charge of choosing the next sentence and the attitude during the interaction to adapt to the level of anxiety of the user. We conducted a perceptive study in order to assess if the attitudes displayed by the agent were correctly perceived by users. Results showed that the agent could convey interpersonal attitudes. However, in these works we have only considered dyadic-interactions. This work, extends our previous model by combining it with a new rule-based model for the autonomous generation of proxemics and body orientation behavior supporting group formations. The full list of parameters we are considering are gaze, interpersonal distance, body orientation, activation of gestures and head movements, spatial extent and power of gestures, facial expressions and head orientations. The framework for group interaction proposed in [21] is interesting but it has only adressed problem-solving interaction. Therefore we are proposing another framework aimed at conversational interaction.

3 Framework For Conversational Group Representation

In order to represent ECAs' conversational groups and in particular their interactions, we propose a framework grounded in human social psychology literature [17][11].

In [17], McGrath presents the works that have been done to model an interacting group. He describes a general framework of interactivity. He starts by explaining that a group, which takes part in an interaction process, is firstly defined by its members. Each of these members has a set of properties (E.g.: descriptive like the personnality or the gender). The members are also related to each other by their relations and roles, defining the group structure. This is the first panel of informations about the group. The second panel is the contextualization of the group by an environment and a task. Indeed a group of members involves them doing something somewhere. Each of these panels influences the behavior of the members while interacting. Finally, the interaction is described using three layers. The first one is the communication layer. Each behavior of a member can be seen as a communication from this member to all the others. In a second layer, this communication carries an action (the interaction, with regards to the current task) and interpersonal informations. Finally, the last layer is the impact of the behavior on the current task of the group and on the other members (E.g.: evolution of their goals or social relations). We are using the elements of this framework to design our group model as follows:

- **Interaction Process**: Our group is in a conversation.
- **Members and group structure**: Each member has a particular interpersonal attitude to express towards each other member.
- **Contextualization**: We are specifying the location and the type of conversation (E.g.: smalltalk at the coffee break, a political debate, a lecture).
- **Communication and action**: In our conversational group, the members communicate particular utterances along with information about their interpersonal attitude through the nonverbal behavior exhibited (E.g.: different facial expressions, body orientations or gesturing behaviors).

In [2], Bales presents a set of twelve categories to describe the various social interactions that happen in a small group interaction. Positive social-emotional interactions (show solidarity, show tension release and agree), negative social-emotional interactions (show antagonism, show tension and disagree) and task-oriented questions (ask for orientation, ask for opinion and ask for suggestion) and task-oriented answers (give orientation, give opinion and give suggestion). He claims that every human interaction should fall into one and only one of these categories. These categories are not dedicated to a particular context and are useful to describe a general-level approach. Therefore as we are focusing on conversational groups, we are specifying our model using definitions from Linguistics. Based on [11], we consider three levels in a conversation. The higher level is the *speech situation*. It defines the context of the conversation (e.g. a lunch meal, a political debate or a lecture). We use this as the *environment parameter* of McGrath model. Levinson [16] refers to them as activity types. Within this speech situation will occur several *speech events*. A *speech event* is composed of one or several speech acts. Examples of speech events include asking for a direction, telling a joke or explaining an opinion. This is the *task parameter*. Finally, the lower level is the

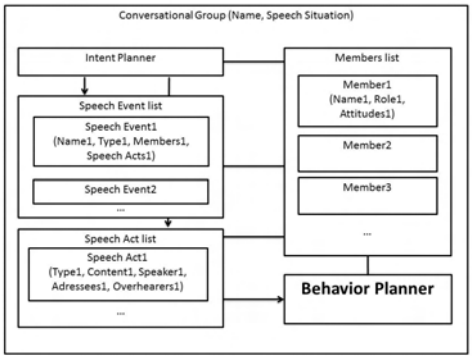


Figure 1: A schematic view of our framework for small conversational groups showing the core entity *Conversational Group*, the *Members* and the list of *Speech Events*.

speech act. The speech act is basically the miminal unit of a speech event. It is usually one utterance and we are using it as the *action parameter*. Speech act are very well descibed by Searle in [25] where he presents a classification of these acts into five categories: assertives, directives, commissives, expressives and declaratives.

Finally, based on the works presented above, the first entity of our framework as depicted Figure 1 is the *Conversational Group*. This entity contains the list of members and the attitude of each member towards all the others. It also contains information such as *Group Name* and *Speech Situation*. The entity also contains the list of potential, current and past *Speech Events*. A *Speech Event*, in turn, contains the list of members participating in this event, a list of *Speech Act*, name and type (e.g. asking a question, telling a joke or explaining something). The Speech Act contains the type (from Searle's classification [25]), the content of the speech, the speaker, the list of adressees and the list of bystanders referenced from the list of group members. A group member contains an identifier and a role.

Within this framework, our model for expressing ECA's interpersonal attitudes is divided into two parts and it is designed to be *SAIBA* compliant [27]. Therefore, an *Intent Planner* is responsible of choosing the communicative intentions of the ECAs, it also builds the list of *Speech Events* and *Speech Acts* that arise in a configuration depending on the role of the ECAs and the *Speech Situation*. A *Behavior Planner* is in charge of transforming communicative functions (or intents) into multimodal behaviors. In this work, we have focused on the *Behavior Planner* and on the speaker role of an ECA. We will consider listener and bystander behavior in future work.

4 Influencing The Nonverbal Behavior : *Behavior Planner*

4.1 Nonverbal Behavior And Interpersonal Attitude

Interpersonal attitudes can be expressed with nonverbal behavior in both dyadic [7, 4, 5] and small group interactions [18, 8, 23]. In dyadic interactions, a more dominant person tends to do more gestures [7], less expressive facial expressions are associated with a submissive attitude [5] and mutual gaze is a sign of dominance or friendliness [4]. These are examples of particular nonverbal behaviors accompanying the speech. But we also want to explore nonverbal behavior related to the group formation and expression of attitude in such a new interaction configuration. In [18], Mehrabian describes eye gaze, posture and distance as important behaviors that impact the evaluation of attitude in such interactions. He presented several research studies that focused on the relation between interpersonal attitude and a variety of nonverbal behavior, including proxemics based on the work of Hall [10]. This was used in [8] to infer social relations from the interpersonal distances between members of a group. Based on Hall's Proxemics work and Kendon's F-Formations [14] to recognize the group formations, they report that people with a higher social distance keep a higher physical distance between them. In [23], they identify an emergent dominant leader in a small group using speaking and gazing cues while building an annotated corpus of small group interactions.

4.2 Two-stage influence

Central to our model is the Behavior Planner component, which upon receiving intents about expressing specific attitudes needs to produce the proper nonverbal behavior to be exhibited by the ECA. We are doing it in two stages happening simultaneously. On one hand we are influencing the nonverbal behavior related to conversational and performative intents (e.g. facial expression, gestures, head orientation). On the other hand, we are influencing the behavior related to group formations and cohesion (e.g. gaze behavior, interpersonal distance and body orientation). We limited the generated conversational nonverbal behavior only for the ECA that is speaking. As we are integrating two models that both influence the nonverbal behavior of an agent, we define the following mechanism to combine them: on each modality, the two stages are given a weight (which sum equals to 1) to indicate the degree of influence each model has on the modality. Figure 2 shows for each modality (i.e. nonverbal behavior affected by the models) the weights corresponding to the degree of influence of conversational behavior (left weight) and group formation (right weight).

First Stage: Nonverbal behaviors accompanying the speech. The nonverbal conversational behavior that we are considering in our model is the following: presence of gestures and head movements, type of facial expressions,

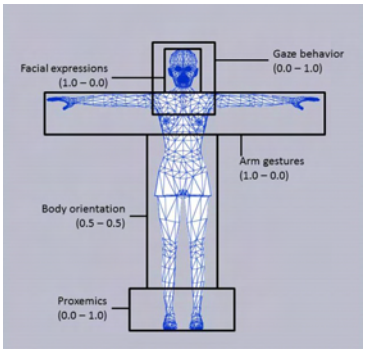


Figure 2: The influence of our combined models on the ECA's generated nonverbal behavior. For each affected nonverbal modality the weights in parenthesis indicate the influence that conversational behavior (left num.) and group formation (right num.) models have on it.

head orientation, presence of gaze avoidance, spatial extent and power of the gestures. Depending on the speech act and the desired expressed attitude, the nonverbal behavior generated should vary. In order to do this we integrated the model developed in [22] with the current model. We are manipulating the probabilities to select particular values for our parameters following this network. A possible outcome for a dominant attitude would be for instance wide and powerful gestures and an upward head, no gaze avoidance and a neutral facial expression. For a friendly attitude, the agent might perform the speech act using a smiling face, tilting his head on the side with wide and smooth gestures. However, this model was developed with dyadic interactions in mind. A possible issue that may arise here and we think that should be adressed is to determine if these behaviors are still compatible with the small group formation and simulation dynamics.

Second Stage: Group formation. The second part of our Behavior Planner is the influence of the attitude on the ECA behavior that manages the group formation and cohesion, in particular the interpersonal distance, the gaze behavior and the body orientation. Based on Hall's proxemics [10] and Kendon F-Formation [14] theories, our model adds on top of these a set of rules to configure this spatial organization depending on the social attitude. When performing a speech act, the model chooses for the speaking agent which other member (human or agent) is its preferred target for a glance, the importance of maintaining an body orientation related to the group or to the adressee and how close it wants to stand to each other member within its social space. For instance, the agent should have a higher probability to glance at the group member towards which it expresses submissiveness or friendliness, stand closer with group members towards which it expresses friendliness or a neutral status level and it should orient its body more directly towards group members with which it expresses submissiveness.

Combining the models. Since we are using two models to control the behavior of speaking ECAs, we are using a weight on each modality to decide which model control which modality. For now we have fixed weights on every modalities except the body orientation. Indeed, each stage take care of its own modalities. Our conversational model is sending the next speech act to perform to our Behavior Planner. This Behavior Planner takes also as input the interpersonal attitudes of the agent towards all the other agents. The first stage computes the upper body nonverbal behavior (facial expression, presence of gestures and head movement, head orientation, spatial extent and power of gestures) for this speech act and the interpersonal attitude towards the adressee. The other stage, computes the body orientation, the interpersonal distance and the group member which is looked at within an F-Formation. On top of this, the combined model computes the preferred target, the weights for the body orientation modality (more weight from the group formation model resulting in an orientation more consistent with the group and less towards the adressee) and the desired interpersonal distances between all characters in their social spaces.

5 Implementation

The preliminary implementation that we present relies on two separate technologies, the VIB platform and the Impulsion AI engine. The VIB platform is a SAIBA compliant platform for the generation and realization of multimodal behavior for ECAs. In [22], we extended the *Behavior Planner* of this platform with our bayesian network to generate the agents' nonverbal behavior to express different social attitudes in dyadic interactions. The nonverbal behaviors supporting this scenario were gesturing, gazing, facial expressions, head orientation and gaze avoidance. The Impulsion AI engine is a software platform developed to improve ECAs nonverbal behavior in social simulations with particular emphasis on F-formation systems (i.e. group conversations) and gatherings (e.g. multiple groups sharing the same environment). The engine is grounded on Scheflen's human territories and Kendon's F-Formation [14] theories and it provides ECAs with autonomous generation and realization of gaze, proxemics and body orientation behavior supporting a simulated group conversation. Impulsion works with the Behavior Trees technology underneath and provides an extendable API that allows a designer to customize ECAs behaviors and specify how the agents interact among their-selves and with the user's controlled character (i.e. the avatar) Both VIB and Impulsion have been deployed within the Unity3D game engine. In this preliminary implementation of our model we geared up a set of ECAs with an integrated version of VIB and Impulsion. From a theoretical perspective VIB allows the ECAs to express social attitudes and Impulsion handles their territorial and gaze behavior in simulated F-Formations for group conversations as described earlier in Section 4. On a software engineering perspective, we have coordinated this

integration by allowing VIB to control the upper body part of our characters (gestures and facial expressions, the head orientation is not handled by VIB in this implementation), while Impulsion is controlling the character's interpersonal distance, body orientation and gaze behavior. In particular, VIB's Behavior Planner indicates the nonverbal behavior corresponding to the attitudes that our ECAs need to express and Impulsion handles the believability of the group simulation. Furthermore, gaze attention, body orientations and interpersonal distances are controlled to allow an user's avatar to join a group of ECAs. This integration is still work in progress and presents two challenging issues that we need to address. First the whole orchestration of nonverbal behavior needs to be consistent with the intended social attitudes that we aim to express. The timing and correct synchronization of the behaviors produced by the two engines and exhibited by our ECAs need to be carefully handled. Secondly, at a lower level, we are working on blending the resulting animations corresponding to the behaviors exhibited (e.g. idle body movements with arms gestures) to obtain a more realistic and believable simulation.

6 Conclusion

In this paper, we have presented a model for conversational groups of humans and agents and a preliminary implementation of the Behavior Planner of this model. We have used an approach combining two models of social interaction, one dedicated to conversational nonverbal behavior and the other for small group formation and territorial cohesion. This is a novel approach, however it introduces some challenging issues that we need to address: on a theoretical level, we need to assess if two separate models of social behavior are compatible when combined together to generate believable and consistent behavior. We are aware that the model for attitudes in dyadic interactions cannot simply be migrated to small group interactions. This new social context has different requirements due to the different spatial arrangements of the ECAs involved and the need to clearly define the addressee for each separate nonverbal modality (e.g. body oriented towards a participant while gazing at another). We plan to dynamically change the weights for each modality according to the desired social context we are simulating. On a software level perspective, the blending and synchronization of the nonverbal behavior generated by the two engines adopted (i.e. VIB and Impulsion) need to be considered. The model and the implementation presented in this paper is specifically focused on the speaker ECA in a small group. However, we have plan to consider the listeners and bystanders nonverbal behavior as well. Finally, we plan to run a user evaluation study aimed at validating the model and assessing that the expressed social attitudes together with the simulated group formation mechanics are yielding believable results.

References

[1] Michael Argyle. *Bodily Communication*. University paperbacks. Methuen, 1988.

[2] Robert F Bales. *Interaction process analysis; a method for the study of small groups.* Addison-Wesley, 1950.

[3] Steven A Beebe and John T Masterson. *Communication in small groups: principles and practices*. Boston: Pearson Education, Inc, 2009.

[4] Judee K. Burgoon, David B. Buller, Jerold L. Hale, and Mark A. de Turck. Relational Messages Associated with Nonverbal Behaviors. *Human Communication Research*, 10(3):351–378, 1984.

[5] Judee K. Burgoon and Beth A. Le Poire. Nonverbal cues and interpersonal judgments: Participant and observer perceptions of intimacy, dominance, composure, and formality. *Communication Monographs*, 66(2):105–124, 1999.

[6] Zoraida Callejas, Brian Ravenet, Magalie Ochs, and Catherine Pelachaud. A computational model of social attitudes for a virtual recruiter. In *Autonomous Agent and Multiagent Systems*, 2014.

[7] Dana R. Carney, Judith A. Hall, and LavoniaSmith LeBeau. Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior*, 29:105–123, 2005.

[8] Marco Cristani, Giulia Paggetti, Alessandro Vinciarelli, Loris Bazzani, Gloria Menegaz, and Vittorio Murino. Towards computational proxemics: Inferring social relations from interpersonal distances. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 290–297. IEEE, 2011.

[9] Marco Gillies, I. Barry Crabtree, and Daniel Ballin. Customisation and context for expressive behaviour in the broadband world. *BT Technology Journal*, 22(2):7–17, 2004.

[10] Edward Twitchell Hall and Edward T Hall. *The hidden dimension*, volume 1990. Anchor Books New York, 1969.

[11] Dell Hymes. Toward ethnographies of communication. *Language and literacy in social practice*, page 11, 1994.

[12] Dušan Jan and David R Traum. Dialog simulation for background characters. In *Intelligent Virtual Agents*, pages 65–74. Springer, 2005.

[13] Dušan Jan and David R Traum. Dynamic movement and positioning of embodied agents in multiparty conversations. In *Proceedings of the Workshop on Embodied Language Processing*, pages 59–66. Association for Computational Linguistics, 2007.

[14] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[15] Jina Lee and Stacy Marsella. Modeling side participants and bystanders: The importance of being a laugh track. In HannesHÃűgni VilhjÃąlmsson, Stefan Kopp, Stacy Marsella, and KristinnR. ThÃşrisson, editors, *Intelligent Virtual Agents*, volume 6895 of *Lecture Notes in Computer Science*, pages 240–247. Springer Berlin Heidelberg, 2011.

[16] Stephen C Levinson. Activity types and language. *Linguistics*, 17(5-6):365–400, 1979.

[17] Joseph Edward McGrath. *Groups: Interaction and performance*, volume 14. Prentice-Hall Englewood Cliffs, NJ, 1984.

[18] Albert Mehrabian. Significance of posture and position in the communication of attitude and status relationships. *Psychological Bulletin*, 71(5):359, 1969.

[19] Radoslaw Niewiadomski, Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. Greta: an interactive expressive eca system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '09, pages 1399–1400, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems.

[20] Claudio Pedica and Hannes Högni Vilhjálmsson. Spontaneous avatar behavior for human territoriality. *Applied Artificial Intelligence*, 24(6):575–593, 2010.

[21] Rui Prada and Ana Paiva. Believable groups of synthetic characters. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 37–43. ACM, 2005.

[22] Brian Ravenet, Magalie Ochs, and Catherine Pelachaud. From a user-created corpus of virtual agent's non-verbal behaviour to a computational model of interpersonal attitudes. In *Proceedings of Intelligent Virtual Agent (IVA) 2013*, 2013.

[23] Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast, and Daniel Gatica-Perez. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*, 7(1-2):39–53, 2013.

[24] Klaus Scherer. What are emotions? and how can they be measured? *Social Science Information*, 2005.

[25] John R. Searle. A classification of illocutionary acts. *Language in Society*, 5:1–23, 3 1976.

[26] David Traum and Jeff Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, pages 766–773. ACM, 2002.

[27] Hannes Vilhjálmsson, Nathan Cantelmo, Justine Cassell, Nicolas E Chafai, Michael Kipp, Stefan Kopp, Maurizio Mancini, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, et al. The behavior markup language: Recent developments and challenges. In *Intelligent virtual agents*, pages 99–111. Springer, 2007.

Caractérisation d'unités gestuelles en vue d'une interaction humain-avatar

I. Renna[1] S. Delacroix[2] F. Catteau[1] C. Vincent[1] D. Boutet[1]

[1]Structures Formelles du Langage, UMR 7023 (CNRS / Université Paris 8)
[2]Laboratoire d'Analyse du Mouvement, Institut National de Podologie, Paris

ilaria.renna@gmail.com

Domaine principale de recherche: RFP
Papier soumis dans le cadre de la journée commune: NON

Résumé

Nous présentons ici une méthode de caractérisation d'unités gestuelles coverbales, en vue d'une exploitation dans une interaction humain-avatar. Nous avons enregistré 12 types de gestes avec un système de capture de mouvement. Nous avons utilisé les signaux de position obtenus afin d'en dégager des unités gestuelles à l'issue d'une segmentation de la partie significative. Pour soutenir notre analyse linguistique des gestes, nous présentons les hypothèses biomécaniques, notre méthode de segmentation, les hypothèses de caractérisation et les résultats obtenus.

Mots Clef

Unités gestuelles, segmentation de la partie significative d'un geste, caractérisation de geste.

Abstract

In this paper we present a method to characterize co-verbal gestures unities to be exploited in a human-avatar interaction. We recorded 12 different kinds of gesture with a motion capture system and exploited the obtained position signals to find gesture unities after a stroke segmentation. To prove a linguistic gestures analysis, we present the biomechanical assumptions, our segmentation method and its results as well as the characterization assumptions and their results.

Keywords

Gesture unities, stroke segmentation, gesture characterization.

1 Introduction

La caractérisation du sens des gestes est faite classiquement en fonction de descriptions égocentrées [17] appréhendant les éléments gestuels selon une description globale dans un repère du corps.

Nous voulons montrer que l'on peut caractériser le sens de différentes Unités Gestuelles (UG) sémantiquement proches à partir de formes distribuées sur le membre supérieur dans des repères multiples non égocentrés, centrés sur chacun des segments (main, avant-bras, bras), ce qui facilite une caractérisation automatique des gestes enregistrés en capture de mouvement (section 2).

Cette caractérisation servira de base pour alimenter un algorithme génétique qui anime un agent virtuel : les caractéristiques de plusieurs bases gestuelles seront hybridées pour donner lieu à des nouveaux gestes grâce auxquels un agent virtuel sera amené à interagir avec un acteur réel lors de performances théâtrales.

La segmentation de la partie significative des gestes (*stroke* [15], section 4) est un préalable nécessaire à la caractérisation : on ne peut pas caractériser le sens d'un geste sans savoir à quel moment il intervient. Pour cela, une segmentation automatique est présentée et testée par rapport aux vérités terrain constituées par la segmentation réalisée par deux annotateurs (section 4). Cette opération effectuée et validée, la caractérisation automatique repose sur un centrage par rapport à la variation du mouvement d'un degré de liberté (ddl) — la pronosupination (section 5). Naturellement, la caractérisation sémantique des gestes repose sur un modèle linguistique dont les grandes lignes sont exposées dans la section 3. Ce modèle, basé sur des constantes forme/sens, se décline en plusieurs niveaux dont chacun apporte une information sur la structuration du sens [2].

2 Présentation de la base de données

La base de données étudiée est composée de 91 gestes symboliques coverbaux et isolés réalisés selon un étiquetage sémantique contrôlé [1 et 2]. C'est l'une des 4 bases de données du projet CIGALE dont l'objectif est de créer une interaction avatar-humain.

Les gestes coverbaux présents couvrent l'ensemble des ddl du membre supérieur et sont sémantiquement autonomes (voir 3.1). Certains gestes sont réalisés sur l'ensemble du membre supérieur alors que d'autres peuvent n'être exécutés que sur les doigts, par exemple.

2.1 Capteurs et modèle biomécanique

Les gestes d'un acteur sont recueillis à l'aide d'un système de capture de mouvement 3D composé de 24 caméras numériques infrarouges (VICON, 120 fps), qui assure la fiabilité de la compréhension du geste et l'efficacité de la caractérisation des mouvements de l'avatar. Pour modéliser les segments corporels en trois dimensions, une liste des marqueurs cutanés est établie (*marker-set* de 90 points, Fig. 1).

Celle-ci référence les positions anatomiques à utiliser pour modéliser chaque segment comme un solide indéformable. Généralement, trois repères anatomiques non alignés suffisent à définir un segment. Dans notre modèle (Fig. 2), les segments tronc, bras, avant-bras et main ont été définis à partir des coordonnées spatiales

des mires, selon une méthode standardisée (détail dans la légende). Cette dernière permet la création des trois axes orthogonaux pour chaque système de coordonnées segmentaires ([23], [8]). Pour cela, les centres articulaires du poignet, du coude, de l'épaule ainsi que ceux de la région cervicale et lombaire sont calculés [8].

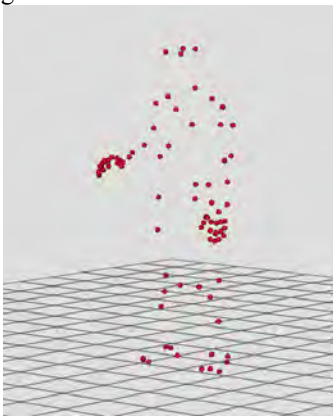


Figure 1. Visualisation du *marker-set*.

Afin de décrire les mouvements articulaires tridimensionnels de l'épaule, du coude et du poignet à chaque instant du geste, les systèmes de coordonnées de chaque articulation sont définis à partir des systèmes de cordonnées segmentaires adjacents à l'articulation. Pour cela, une séquence de rotations successives autour d'axes mobiles permettant d'obtenir la cinématique articulaire grâce aux angles d'Euler est utilisée [23]. La séquence mobile de rotation permet de définir le système de coordonnées articulaires en utilisant les axes de deux segments adjacents : un axe du segment proximal, un axe du segment distal et un axe flottant perpendiculaire aux deux précédents.

Grâce à ce modèle biomécanique, les différents mouvements articulaires du poignet, du coude et de l'épaule sont calculés. Ainsi, les mouvements de flexion palmaire/dorsale et d'adduction/abduction du poignet sont calculés. Ceux-ci correspondent aux mouvements de flexion/extension et d'adduction/abduction de la main tel que décrit dans les schémas d'actions (section 3). Les mouvements d'extension/flexion et de supination/pronation du coude correspondent respectivement aux mouvements d'extension/flexion de l'avant-bras et de supination/pronation de la main pour les schémas d'actions (voir 3.2). Enfin, les mouvements de rétropulsion/antépulsion, d'abduction/adduction et de rotation externe/interne de l'épaule sont mesurés. Ceux-ci correspondent respectivement à une extension/flexion, abduction/adduction et rotation extérieure/intérieure du bras pour les schémas d'actions.

3 Recadrage linguistique en vue d'une interaction humain-avatar

Les gestes coverbaux enregistrés correspondent à des emblèmes ou *quote gestures* ([19] et [13]), c'est-à-dire des gestes sémantiquement autonomes, à la signification indépendante du discours verbal associé.

Les 91 gestes se répartissent en une douzaine d'UG. Les significations sont les suivantes : rejeter, refuser, mépriser, déconsidérer, passer, accepter, considérer quelque chose, considérer quelqu'un, offrir, s'en fiche, s'engager, révérer. Ces étiquettes sémantiques ont été testées et validées auprès d'une population francophone dans un travail antérieur [2].

Chacune de ces UG répond à un schéma d'action singulier qui met en œuvre une partie ou l'ensemble des segments du membre supérieur.

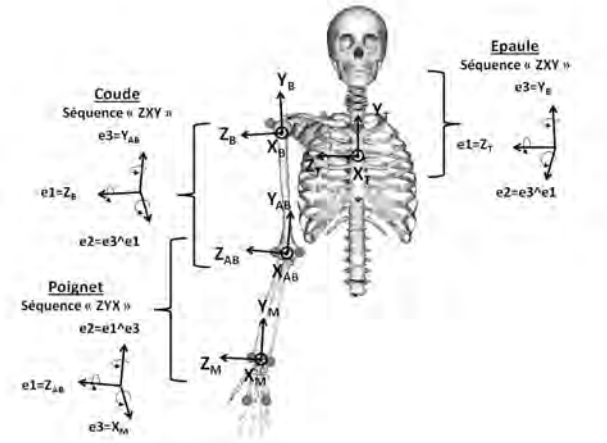


Figure 2. Modèle biomécanique. Pour la main, l'origine du système de cordonnées est situé au centre articulaire du poignet. L'axe Y est le vecteur unitaire reliant le centre des 2e et 5e têtes métacarpiennes à l'origine. L'axe X est le vecteur unitaire normal au plan contenant l'origine et les 2e et 5e têtes métacarpiennes. L'axe Z est le produit vectoriel des axes X et Y. Pour l'avant-bras, l'origine du système de cordonnées est situé au centre articulaire du coude. L'axe Y est le vecteur unitaire reliant le centre articulaire du poignet à l'origine. L'axe X est le vecteur unitaire normal au plan contenant l'origine et les processus styloïdes de l'ulna et du radius. L'axe Z est le produit vectoriel des axes X et Y. Pour le bras, l'origine du système de cordonnées est situé au centre articulaire de l'épaule. L'axe Y est le vecteur unitaire reliant le centre articulaire du coude à l'origine. L'axe X est le vecteur unitaire normal au plan contenant l'origine, l'épicondyle et l'épitrochlée. L'axe Z est le produit vectoriel des axes X et Y. Pour le tronc, l'origine du système de cordonnées est situé au centre articulaire cervical. L'axe Y est le vecteur unitaire reliant le centre articulaire lombaire à l'origine. L'axe Z est le vecteur unitaire normal au plan contenant l'origine, centre articulaire lombaire et l'espace supra sternal. L'axe X est le produit vectoriel des axes Y et Z.

La description sous forme de schémas d'action repose sur la mise en mouvement de différents ddl des segments du membre supérieur dans un ordre précis. Le mouvement est transféré en fonction des moments d'inertie qui sont attachés à chaque ddl et en fonction du mouvement conjoint (involontaire) de l'axe longitudinal (Rotation extérieure/intérieure ou Pronation/supination) associé à toute articulation à deux ddl ([5], [16] et [4]). Ainsi, pour l'UG « refuser » par exemple (Schéma 1), le schéma d'action déploie le geste de la main vers l'avant-bras.

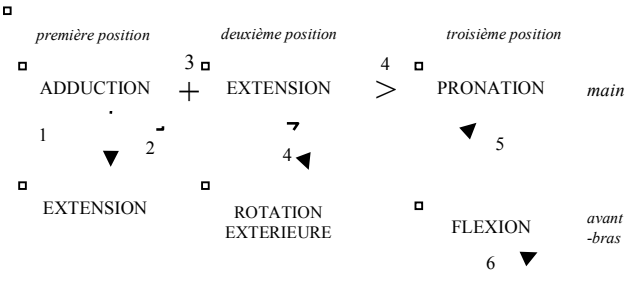


Schéma 1. Schéma d'action de l'UG « refuser ».

3.1 Flux de propagation du mouvement

Dans le schéma d'action, la position du pôle de l'adduction (mouvement vers l'auriculaire dans le plan de la paume) donne le sens de propagation du mouvement. Si l'adduction figure en première ou en deuxième position, le flux de propagation du mouvement est distal-proximal ; il va globalement de la main vers l'avant-bras. Au contraire, si l'adduction est en troisième position, alors elle est un mouvement conséquent des deux premiers et, à ce titre, ne présente pas une amplitude importante. Le geste est initié sur l'avant-bras et se propage alors vers la main selon un flux distal-proximal. On définit ainsi deux types d'UG. Les 8 premières UG de la liste ci-dessus sont structurées sur la main, tandis que les 4 dernières (offrir, s'en fiche, s'engager et révérer) le sont sur le bras. Lors d'une étude précédente, des UG identiques présentées en vidéo ont montré un taux de reconnaissance significatif selon une méthode des juges [2].

3.2 Schéma de l'enchaînement des mouvements sur la main

L'enchaînement des mouvements sur la main suit une structuration telle que le mouvement ou la position des deux premiers ddl entraînent le mouvement involontaire du troisième ddl. Ce troisième mouvement est dû soit à une contrainte biomécanique liée à un mouvement autour de l'axe longitudinal (pronation/supination), soit à un enchaînement entraîné par le moment d'inertie. Dans les deux cas, les pôles du mouvement en troisième position sont parfaitement déterminables et répondent à un ordre d'enchaînement des deux mouvements précédents tel que leur ordre impacte le pôle du troisième mouvement. Ainsi, la suite ADD.EXTEN entraîne un mouvement involontaire de PRONATION, tandis que l'ordre opposé, EXTEN.ADD implique un mouvement de SUPINATION ([1] et [2]).

3.3 Regroupement des UG par famille de sens

Le repérage de l'ordre des pôles mis en mouvement relève tout autant de l'amplitude du mouvement, de la succession temporelle de l'apparition du mouvement, de la position initiale et de l'accélération sans qu'il soit aisé d'en hiérarchiser les critères qui varient même de manière intra-individuelle. En revanche, il est possible de regrouper les UG par champ sémantique sur une base formelle (Schéma 2).

Dans un premier temps, il s'agit de déterminer le flux de propagation du mouvement : soit le geste part de la main et le mouvement remonte sur l'avant-bras, soit il part du bras et diffuse vers la main (Main et Bras dans le schéma). Pour la branche de la main (Schéma 2, à gauche), la position de la pronosupination initiale au geste est soit marquée, soit non marquée. Au niveau suivant, on examine le mouvement de la pronosupination par rapport à la position initiale. On obtient ainsi 8 schémas d'action manuels. Pour la branche du bras (Schéma 2, à droite), on examine la position ou le mouvement de l'ADD/ABD du bras. Au niveau suivant, le mouvement de la pronosupination permet de distinguer les 4 UG organisées sur le bras.

Chacune de ces UG a un label sémantique. Un premier niveau de regroupement hyperonymique compose 4 ensembles sémantiques : i/ Positionnement par rapport aux choses, ii/ Considération ou jugement, iii/ Implication et iv/ Intérêt. Ce niveau sémantique correspond au 2e niveau de disjonction formelle dans le schéma. On peut également procéder à un regroupement sémantique en deux parts correspondant à la première disjonction formelle (Main ou Bras) : Positionnement par rapport au monde *versus* Positionnement par rapport à la relation.

Ainsi, à différents niveaux de différenciation formelle correspond un étiquetage sémantique spécifique.

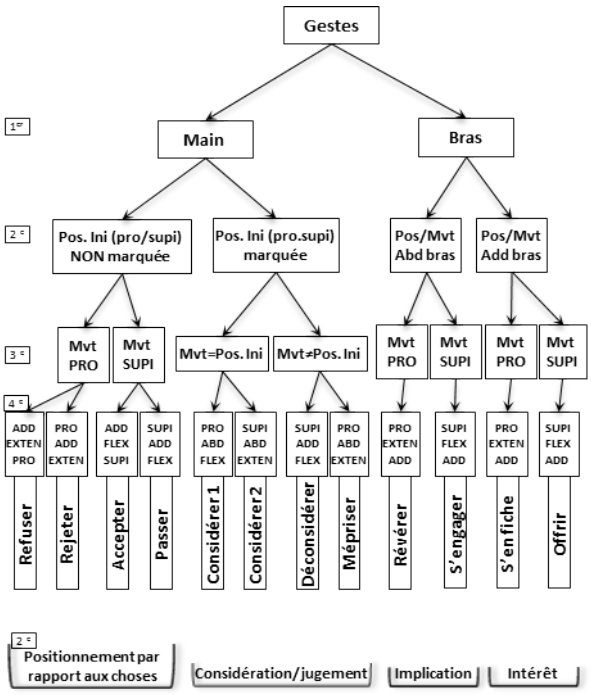


Schéma 2. Présentation formelle des gestes en vue de leur caractérisation sémantique.

4 Segmentation des signaux gestuels

Dans notre base de données, chaque signal gestuel est constitué d'un enchaînement pose en T, geste, pose en T. Une segmentation automatique est nécessaire afin d'extraire le geste.

La séquence généralement admise est celle correspondant à une suite de 4 phases ([17] et [6]) :

1- Position de repos
2- La préparation (pré-*stroke*)
3- Le cœur (*stroke*)
4- La rétraction (post-*stroke*)

Cette séquence décrit la structure du geste.

La difficulté réside dans l'impossibilité de trouver un critère objectif automatique pour extraire le *stroke,* la partie sémantiquement significative. Cette opération est complexe même pour un humain et reste incertaine [21].

Dans notre cas, nous effectuons la segmentation sur la base de propriétés morpho-cinématiques (*morpho-kinetics* selon Kendon [14]). En effet, la préparation du mouvement consiste en un mouvement balistique qui amène le(s) bras vers le cœur du mouvement [3]. Cette balistique consiste en une accélération puis une décélération à l'approche de la pose finale, puis une

accélération et une décélération symétriques aux premières pour revenir à la position de repos.
Les poses en T sont également caractérisées par des mouvements d'accélération et de décélération.
En conséquence on a décidé d'extraire le *stroke* de chaque geste en considérant la valeur absolue de la dérivée en Y des positions de l'index (considéré dans tous les cas comme le membre du corps qui bouge le plus) : les minima de ce signal représentent le passage entre décélération et accélération. Pour la segmentation automatique on considère donc que le *stroke* est la partie comprise entre la phase minimale qui précède le deuxième maximum (caractéristique du début du *stroke*) et la phase minimale qui suit l'avant-dernier maximum (fin du *stroke*) (Fig. 3). Pour éviter de prendre en compte des phase maximales et minimales dues au bruit (petits mouvements d'ajustement ou de préparation) un seuil est fixé à 0.5 car nous considérons qu'un mouvement sémantiquement valide présente un pic supérieur à cette valeur.

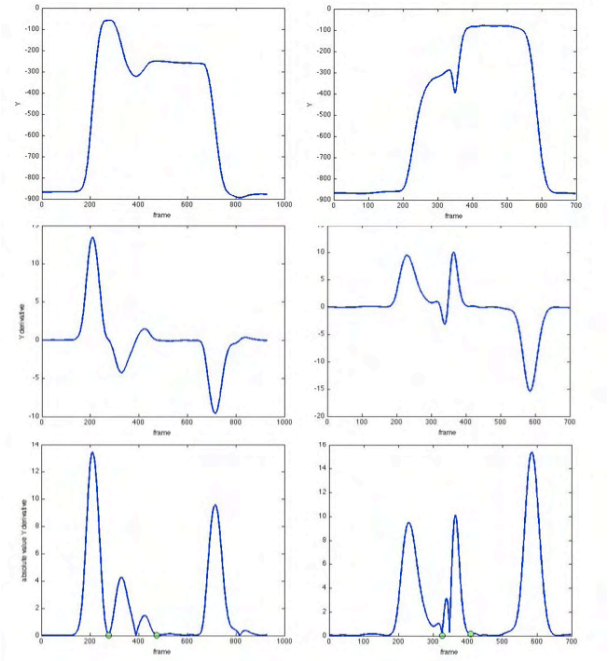


Figure 3. A gauche : signaux du geste « révérer » ; à droite : signaux d' « accepter ». De haut en bas : positions de l'index en Y, vitesse (dérivée) et valeur absolue de la vitesse avec individuation automatique du *stroke* comprise entre deux points verts.

4.1 Evaluation de la segmentation

Pour évaluer les méthodes de segmentation automatique, il est nécessaire de comparer les performances de la segmentation réalisée automatiquement avec celle de juges humains (codeurs). En général, les méthodes pour effectuer cette comparaison utilisent la segmentation d'un seul codeur comme référence [11]. Une telle référence ne devrait pas être considérée comme unique vérité terrain, étant donné que l'accord inter-annotateurs est souvent assez faible [12]. De plus, pour s'assurer qu'une segmentation automatique n'est pas biaisée par rapport aux choix d'un codeur, elle devrait être comparée directement au travail de plusieurs codeurs [10]. Pour évaluer, d'une part, l'accord inter-annotateurs, et d'autre part, la segmentation automatique, nous avons décidé d'adopter deux méthodes : l'*Accurate Temporal Segmentation Rate* (ATSR) (taux de segmentation temporelle précis) [20] et le F-score [22].
L'ATSR est une mesure basée sur le temps, qui permet d'évaluer la performance en termes de précision de détection des début et fin de *stroke* pour chaque geste, alors que le F-score donne plus d'informations sur la précision et permet l'individuation de la typologie d'erreur.
Trois comparaisons sont effectuées avec ces deux méthodes :

1. la segmentation automatique est comparée au premier annotateur, considéré comme vérité terrain (cas 1) ;
2. la segmentation automatique est comparée au second annotateur considéré comme vérité terrain (cas 2) ;
3. les deux annotateurs sont comparés (cas 3).

Pour chaque geste considéré, l'ATSR a été calculé de la manière suivante : l'*Absolute Temporal Segmentation Error* (ATSE) (erreur de segmentation temporelle absolue) est évaluée en additionnant l'erreur temporelle absolue entre la vérité terrain et le résultat de l'algorithme pour les événements de début et de fin, le tout divisé par la durée totale de l'occurrence du *stroke* mesurée à partir de la vérité terrain, comme formalisé dans l'équation 1.
Une fois les ATSE obtenues, les mesures d'ATSR sont calculées en soustrayant l'ATSE moyenne à 1, de manière à obtenir le taux de précision comme montré dans l'équation 2. Une segmentation parfaitement précise produit un ATSR de 1.

$$ATSE = \frac{\left|\text{Start}_{GT} - \text{Start}_{Alg}\right| + \left|\text{Stop}_{GT} - \text{Stop}_{Alg}\right|}{\text{Stop}_{GT} - \text{Start}_{GT}} \qquad (1)$$

$$ATSR = 1 - \frac{1}{n}\sum_{i=1}^{n} ATSE(i) \qquad (2)$$

L'équation 1 compte les différences qui se produisent image par image, donc une erreur est prise en compte même quand les annotations diffèrent de seulement quelques images.
Pour limiter cet effet, il est possible de fixer un seuil de tolérance α de manière à ce que

si ATSE(i) < α, alors ATSE(i) = 0. (3)

Comme, en général, un *stroke* dure environ 100 images, nous fixons α =0.2. Ceci correspond à une différence globale de α *100=20 images (ce qui signifie environ 0.17s à la fréquence d'acquisition de 120i/s) ce qui est un choix adéquat comparé à la durée de la vérité terrain, considérant que, en moyenne, il est facile d'avoir 10 images de décalage pour chaque début et fin.
Nous obtenons : pour le cas 1, ATSR=0.6038 ; pour le cas 2, ATSR=0.5857 ; pour le cas 3, ATSR=0.8707.
Ce genre de méthode manque néanmoins d'exhaustivité car il ne fournit pas le type d'erreur.
Nous pouvons, en fait, avoir 5 types d'erreur (Fig. 4).
De fait, il est important de savoir si la segmentation automatique est erronée, mais préserve le *stroke* (Fig. 4, erreur 2) ou si elle le coupe (tous les autres cas).

De manière à évaluer la qualité de notre segmentation et l'accord inter-annotateurs, considérons la précision (*p*) et le rappel (*r*) ([9] et [18]) : la précision est la fraction de détections qui sont des vrais positifs plutôt que des faux positifs (équation 4), alors que le rappel est la fraction de vrais positifs qui sont détectés plutôt que manqués (équation 5). En termes probabilistes, la précision est la probabilité que la détection soit valide, et le rappel est la probabilité que les données de vérité terrain soient détectées.

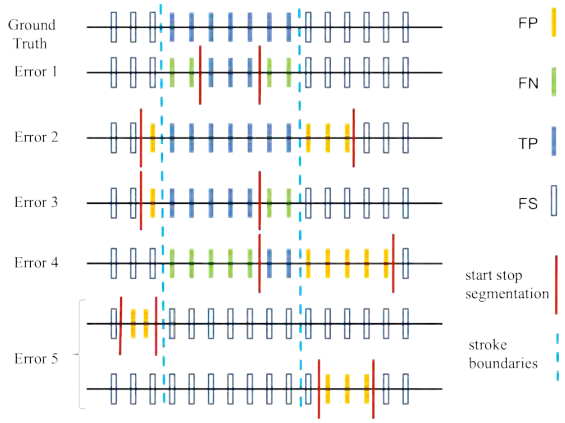


Figure 4. Différentes erreurs de segmentation possibles. FP = Faux Positif, FN = Faux Négatif, TP = Vrai Positif et FS = Hors Segmentation.

$$p = \frac{TP}{TP + FP} \quad (4) \qquad r = \frac{TP}{TP + FN} \quad (5)$$

La précision et le rappel peuvent être combinés dans le F-score de cette manière : $F_\beta = (1+\beta^2) * \frac{p * r}{\beta^2 p + r}$ (6)

Quand le paramètre β=1, le F-score est dit équilibré et s'écrit F_1 : $F_1 = 2 * \frac{p * r}{p + r}$ (7)

Le score F_1 peut être vu comme une moyenne pondérée de la précision et du rappel. F_1 est compris entre 0 (moins bonne valeur) et 1 (meilleure valeur).

Les résultats obtenus sont résumés dans le tableau 1.

	p	*r*	F_1
Cas 1	0.7430	0.9216	0.8227
Cas 2	0.7358	0.9151	0.8157
Cas 3	0.9077	0.9053	0.9065

Tableau 1. Résultats obtenus pour les trois cas d'étude.

En général, des valeurs élevées de F_1 sont obtenues ; *r* est plus élevé que *p* dans la comparaison avec la segmentation automatique, ce qui signifie que l'algorithme renvoie la plupart des résultats pertinents, alors que *p* est plus élevé pour l'accord inter-annotateurs : ce qui est obtenu le plus, ce sont les accords pertinents.

Les résultats concernant les types d'erreur sont présentés dans le tableau 2.

	Cas 1	Cas 2	Cas 3
Erreur 1	4	5	17
Erreur 2	53	49	15
Erreur 3	24	16	54
Erreur 4	10	21	5
Erreur 5	0	0	0

Tableau 2. Erreurs mesurées dans les cas d'étude.

Il est important de souligner que dans les cas 1 et 2, l'erreur 2 se produit le plus : ceci signifie que la méthode de segmentation préserve le *stroke*. De plus, l'erreur la plus basse est la coupe du *stroke* : nous pouvons affirmer que la méthode de segmentation utilisée est robuste pour analyser le type des gestes présentés car elle permet d'avoir des F-score élevés avec une coupure du *stroke* très basse.

Pour l'accord inter-annotateurs, nous soulignons que, quand ils se trompent, c'est principalement parce que l'un d'eux coupe le *stroke* (erreur 1) ou parce que l'un anticipe l'autre (erreur 3).

Cette méthode de segmentation pour simple et robuste qu'elle soit demande néanmoins à être étendue et évaluée pour des situations différentes.

5 Caractérisation des composantes des schémas d'action

Afin de caractériser les schémas d'action pour chacun des gestes coverbaux enregistrés, les signaux segmentés sont transformés en données cinématiques selon la modélisation biomécanique (section 2.1). Les mouvements des différents ddl de chaque articulation du membre supérieur droit (épaule, coude et poignet) sont pris en compte et normalisées temporellement sur 101 points [7].

Pour la caractérisation, on part de la constatation que dans une communication humain-humain, pour n'importe quel type de geste (réalisé sur l'ensemble du membre supérieur ou juste sur un de ses segments), les mouvements des ddl de pronosupination sont les plus visibles. Il est donc décidé, dans un premier temps, de focaliser l'analyse sur la partie des signaux alignée temporellement avec la zone de pronosupination contenant la plus grande variation. Les paramètres biomécaniques tels que les positions initiales et finales de chaque ddl ainsi que leur amplitude maximale sont pris en considération (Fig.5).

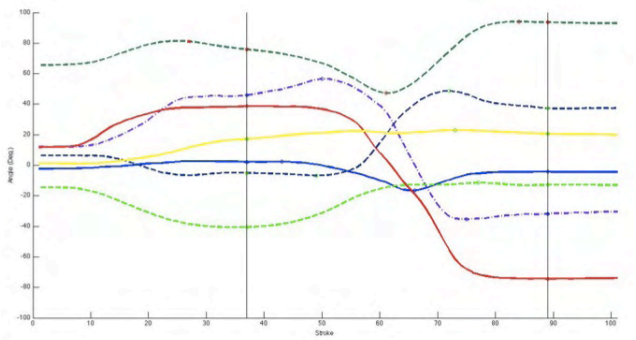


Figure 5. Signaux des différents segments du membre supérieur. Pour la main : en rouge la supination/pronation, en bleu ciel l'abduction/adduction et en violet pointillé la flexion/extension. Pour l'avant-bras en vert foncé pointillé l'extension/flexion. Pour le bras : en bleu foncé pointillé la rotation interne/externe, en vert pointillé l'abduction/adduction et en jaune l'extension/flexion. Les lignes verticales indiquent la plus grande variation de pronosupination.

La caractérisation a été effectuée sur les 33 gestes (des 91 captés) mobilisant des mouvements de tous les segments du membre supérieur. Pour cela, les étapes considérées de l'arbre de décision présenté au schéma 2 vont i/du premier nœud (détermination du flux manuel

ou brachial) ii/ au quatrième (séparation des gestes par la pronosupination).
Pour le premier nœud, il s'agit de déterminer si la propagation du mouvement part du bras (flux proximal-distal) ou de la main (distal-proximal). Pour cela, on a calculé : 1/ l'instant où pour la position initiale apparaissent le min. et le max. de chaque ddl à l'intérieur du *stroke* découpé automatiquement ; 2/ la différence temporelle entre le min. et le max. des ddl d'un segment à l'autre (bras [offrir, s'en fiche, s'engager et révérer], avant-bras et main [pour tous]). Dans ce dernier calcul, le choix de la valeur min. ou max. pour tel ou tel ddl correspond à la position initiale et donc, *a priori*, à l'opposé du pôle en mouvement repéré au cours du *stroke*. Si le mouvement de la main est une EXTEN (valeur positive), alors la position initiale correspond à un minimum (flexion, valeur négative). Ainsi par exemple, pour la ligne supérieure du schéma 1 qui illustre les pôles en mouvement du geste « refuser » : ADD.EXTEN>PRO, on choisit comme positions initiales la valeur max. de l'ADD/ABD, la valeur min. de la FLEX/EXTEN et la valeur min. de la SUPI/PRO.
On a fixé un seuil minimal de 10 images, correspondant à un volant de 2 images vidéo à 25i/s, pour la différence temporelle permettant de connaître le flux.
Sur 33 gestes testés qui recouvrent les 12 UG présentées dans la section 3 (chaque UG a été réalisée entre 2 et 3 fois), la détermination du flux par cette méthode valide 87,88% de ce qui était attendu. Parmi les 4 cas non validés, 3 sont en dessous du seuil des 10 images et ne répondent donc à aucun flux déterminable et un seul cas (une réalisation de « révérer ») montre un flux inverse de ce qui était attendu.
A l'autre bout de l'arbre de décision (schéma 2), la quatrième étape de la caractérisation — celle des pôles en mouvement pour déterminer le schéma d'action — a été faite selon une méthode avec deux types de données (voir a/ et b/ ci-dessous).
Dans un premier temps, les calculs concernent la moyenne des deux ou trois réalisations par UG (33 gestes en tout) et portent donc sur 12 UG moyennées dont on connaît *a priori* les pôles en mouvement ainsi que les étiquettes sémantiques. Dans un second temps, on détermine l'amplitude maximale pour chaque ddl, a/ soit à l'intérieur du bornage de la pronosupination tel qu'il est présenté dans la figure 5 ; b/ soit plus largement, à partir du *stroke*, en calculant la différence entre la position finale et la position initiale de chaque ddl. On obtient ainsi les pôles en mouvement pour l'ensemble des ddl qui caractérisent l'UG, c'est-à-dire 60 ddl pour la somme des 12 UG.
Les résultats avec le premier type de données (a/ dans le bornage de pronosupination) donnent un taux de reconnaissance de 76,67%. L'autre option (b/ dans le *stroke* avec la différence de position finale et initiale) voit un taux de caractérisation bien meilleur : 90%. Sur les 60 ddl, seuls 6 pôles attendus voient leur pôle opposé apparaître. Dans les deux options, sur une moyenne de 6 ddl mesurées par UG, le pôle le plus sujet à erreur est l'ABD/ADD de la main (a/ 36% des erreurs, b/ 67%) ; il s'agit du pôle présentant la plus petite amplitude (25° et 35°).
Les étapes intermédiaires — 2 et 3 — de caractérisation (voir schéma 2) consistent à déterminer :
- pour l'étape 2, le marquage des positions initiales de la pronosupination et le mouvement d'ABD *vs* ADD du bras ;
- pour l'étape 3, la position initiale et le mouvement de la pronosupination identique *vs* opposé et le pôle du mouvement entre PRO et SUPI.
La caractérisation du mouvement ABD *vs* ADD du bras et PRO *vs* SUPI est effectuée sans aucun problème. En revanche le marquage des positions initiales de la pronosupination (étape 2) ne donne pas les résultats escomptés. Seule la différence d'amplitude de la rotation intérieure/extérieure entre le début et la fin du *stroke* est significative dans ce cas. Pour un intervalle de confiance à 95%, il n'y a pas de zones de recouvrement entre "rejeter/refuser", d'une part, et "mépriser" d'autre part. Pour le trio "passer/accepter/déconsidérer", ce non recouvrement est également vérifié. Ainsi, il convient de modifier le critère de marquage ddl (PRO/SUPI) de l'étape 2 en un différentiel d'amplitude de rotation extérieure/intérieure plus ou moins marqué.
Pour l'étape 3, l'identité ou l'opposition entre la position initiale et le mouvement de pronosupination est un bon critère puisque pour un intervalle de confiance à 95%, il n'y a pas de zones de recouvrement entre "rejeter/refuser/mépriser", d'un côté, et "considérer quelque chose", de l'autre. Il en va de même entre "passer/accepter/déconsidérer", d'une part, et "considérer quelqu'un", d'autre part.
En résumé, les seules étapes qui ne donnent pas entière satisfaction sont donc la 1re (un seul cas d'inversion pour « révérer ») et la 4e (90% des pôles attendus). Les étapes intermédiaires sont fiables à 100%.

6 Conclusion

Dans ce travail, nous avons présenté une méthode de caractérisation de 12 unités gestuelles sémantiquement proches à partir de formes distribuées sur le membre supérieur. Pour cela une capture du mouvement a été effectuée ainsi qu'une méthode de segmentation automatique. Les tests faits à partir du protocole de segmentation montrent sa robustesse dans l'individuation du *stroke* nécessaire pour la caractérisation gestuelle.
Les méthodes simples de caractérisation répondent pour l'instant à l'exigence d'associer, à chaque étape, un étiquetage sémantique à la caractérisation formelle. C'est le cas, puisque les UG qui partagent les mêmes pôles se différencient uniquement par l'ordre d'apparition dans le schéma d'action. Or, nous n'avons pas encore fait cette caractérisation pour 4 d'entre eux. Pour cette étude, nous ne pouvons discriminer, d'une part, « refuser » de « rejeter » et, d'autre part, « accepter » de « passer ». Par contre ces deux groupes sont étiquetables : d'un côté, un *positionnement négatif par rapport aux choses* et de l'autre, le même type de positionnement *positif* cette fois. Ainsi, tous les gestes ont un étiquetage sémantique associé avec une granularité variable. Il reste à éprouver ces méthodes de caractérisation pour des réalisations ne mettant en mouvement qu'un seul segment comme la main.

Bibliographie

[1] Boutet, D., Une morphologie de la gestualité : structuration articulaire. *Cahiers de linguistique analogique,* n°5, Abell, pp. 80–115, décembre 2008.

[2] Boutet, D., Structuration physiologique de la gestuelle : modèle et tests. *Lidil* 42, 77–96, 2010.

[3] Chellali R., Renna I., Bernier E., Détection et Reconnaissance des Gestes Emblématiques, *Interaction Homme-Machine pour l'Apprentissage Humain -IHMA –RFIA*, 2012.

[4] Cheng, P. L., Simulation of Codman's paradox reveals a general law of motion, *Journal of Biomechanics*, 39(7), 1201–1207, 2006.

[5] Codman, E. A. *The shoulder: rupture of the supraspinatus tendon and other lesions in or about the subacromial bursa*. RE Kreiger, 1934.

[6] Corradini, A. et Cohen, P. R, Speech-gesture Interface for Handfree Painting on a Virtual Paper suing Partial Neural Networks as Gesture Recognizer, *Proceedings IJCNN'02, HI*, 2293–2298, 2002.

[7] Desroches, G., Dumas, R., Pradon, D., Vaslin, P., Lepoutre, F.-X., Chèze, L., Upper limb joint dynamics during manual wheelchair propulsion. *Clinical Biomechanics* 25, pp. 299–306, 2010.

[8] Dumas, R., Chèze, L., Verriest, J.-P., Adjustments to McConville et al. and Young et al. body segment inertial parameters. *Journal of Biomechanics* 40(7), 543–553, 2007.

[9] Fawcett. T., An introduction to ROC analysis. *Pattern Recogn. Lett.* 27(8), pp. 861–874, 2006.

[10] Fournier, C., Inkpen, D., Segmentation Similarity and Agreement, *NAACL HLT '12 Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2012.

[11] Gonzalez Preciado M., Computer Vision Methods for Unconstrained Gesture Recognition in the Context of Sign Language Annotation, PhD thesis, Toulouse, 2012.

[12] Marti A. Hearst. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics* 23(1), 33–64. MIT Press, Cambridge, MA, USA.

[13] Kendon, A., How Gestures Can Become like Words. In *Cross-Cultural Perspective in Nonverbal Communication*, 131–141. C.J. Hogrefe, Toronto & Lewiston, N.Y.: Fernando Poyatos, 1988.

[14] Kendon, A. An agenda for gesture studies. *Semiotic Review of Books* 7(3), 8–12, 1996.

[15] Kita, Sotaro, Ingeborg van Gijn, et Harry van der Hulst. « Gesture and Sign Language in Human-Computer Interaction ».. *Lecture Notes in Computer Science*. 1371:23–35, Springer, Berlin / Heidelberg, 1998.

[16] MacConaill, M. A. « The Movements of Bones and Joints ». *Journal of Bone & Joint Surgery, British Volume* 30-B(2), 322–326, 5 janvier 1948.

[17] McNeill, D., *Hand and Mind : What Gestures Reveal about Thought*. University of Chicago Press, Chicago & London, 1992.

[18] Olson, D. L. et Delen, D.. *Advanced Data Mining Techniques* (1st ed.). Springer, 2008.

[19] Payrató, L., A pragmatic view on autonomous gestures: A first repertoire of Catalan emblems, *Journal of Pragmatics* 20(3), 193–216, 1993.

[20] Ruffieux, S., Lalanne, D., Mugellini, E., ChAirGest: a challenge for multimodal mid-air gesture recognition for close HCI. *ICMI,* 483–488, 2013.

[21] M. Sigalas, H. Baltzakis, P. E. Trahanias. Gesture recognition based on arm tracking for human-robot interaction. *IROS*, 5424–5429, 2010.

[22] Van Rijsbergen, C. J., Information Retrieval (2nd ed.). Butterworth, 1979.

[23] Wu, G., van der Helm, F. C., Veeger, H. E., Makhsous, M., Van Roy, P., Anglin, C., Nagels, J., Karduna, A. R., McQuade, K., Wang, X., Werner, F. W., Buchholz, B., ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion--Part II: shoulder, elbow, wrist and hand. *Journal of Biomechanics* 38(5), 981–992, 2005.

Un modèle logique de théorie de l'esprit pour un agent virtuel dans le contexte de simulation d'entretien d'embauche

M. Belkaïd[1] N. Sabouret[2]

[1] ETIS, UMR 8051, Cergy
[2] LIMSI-CNRS, UPR 3251, Orsay

Résumé

Dans le contexte de la simulation d'entretiens d'embauche, le recruteur virtuel doit être capable de se représenter et de raisonner sur les états mentaux de l'utilisateur en s'appuyant sur des indices non-verbaux qui sont des indices de ses émotions et de son attitude sociale. Dans cet article, nous proposons un modèle formel de théorie de l'esprit (ToM) pour des agents virtuels dans le contexte d'interaction humain-agent, en nous concentrant sur les dimensions affectives des interactions. Ce modèle combine deux paradigmes de théorie de l'esprit et s'appuie sur une logique modale de type BDI dans laquelle sont décrites les règles d'inférence sur les états mentaux, les émotions et les relations sociales entre les acteurs. Nous présentons les résultats d'une étude préliminaire sur l'impact d'un tel modèle dans le contexte de la simulation d'entretien d'embauches.

Mots Clef

Théorie de l'esprit, modèles cognitifs, approches logiques, agents virtuels intelligents.

Abstract

In job interview simulation, the virtual interviewer must be capable of representing and reasoning about the user's mental state based on social cues that inform the system about his/her affects and social attitude. In this paper, we propose a formal model of Theory of Mind (ToM) for virtual agent in the context of human-agent interaction that focuses on the affective dimension. It relies on a hybrid ToM that combines the two major paradigms of the domain. Our framework is based on modal logic and inference rules about the mental states, emotions and social relations of both actors. Finally, we present preliminary results regarding the impact of such a model on natural interaction in the context of job interviews simulation.

Keywords

Theory of Mind, Cognitive Models, Logic-Based Approaches, Intelligent Virtual Agents.

1 Introduction

Le travail présenté dans cet article se situe dans le contexte de l'utilisation d'agents virtuels pour la simulation d'entretien d'embauche, qui a reçu un intérêt croissant de la communauté ces dernières années [2, 5, 16]. En effet, sur le plan sociétal, l'aide à l'insertion professionnel est un objectif majeur de nos sociétés (le taux de chômage chez les moins de 25 ans en Europe a dépassé 25%) et il a été montré que la simulation d'entretien d'embauche, en particulier avec des agents virtuels, peut à améliorer la confiance en soi et les compétences sociales des jeunes [8, 16, 26]. Sur le plan théorique, la simulation d'entretien d'embauche est une situation contrôlée dans laquelle il semble plus facile d'étudier la reconnaissance, le raisonnement et la synthèse de comportements affectifs, qui sont des problèmes trop difficiles pour être abordés dans un cadre général.

Notre objectif est de faire des agents dont la réaction verbale et non-verbale est cohérente avec les entrées non-verbales (sourires, expressions émotionnelles, mouvements du corps). Alors que la majorité des modèles comportementaux pour les agents virtuels proposent des modèles plutôt réactifs[16, 20, 21], nous proposons de raisonner sur les états mentaux de l'interlocuteur pour adapter le comportement des agents. La théorie de l'esprit (ou ToM, pour *Theory of Mind*) est la capacité qu'ont les humains et les primates à interpréter, prédire et même influence le comportement des autres[4]. Dans le contexte d'agents intelligents pour la pratique de compétences sociales, cette capacité nous semble être la clef vers des comportements plus réalistes.

Dans la section suivante, nous présentons brièvement les recherches qui sont à la base de nos travaux. Les sections 3 présentent l'architecture générale et notre modèle logique de ToM. La section 4 présente notre implémentation dans le contexte de la simulation d'entretiens d'embauche. Nous présentons les grandes lignes d'une expérimentation préliminaire dans la section 5 et nous discutons des résultats et des perspectives dans la section 5.2.

2 Travaux connexes

Les modèles d'appraisal comme CPM [25] ou OCC [19] peuvent être utilisés pour permettre aux agents virtuels de raisonner sur la dimension affective de l'interaction, portée par le comportement non-verbal des deux interlocuteurs. Ainsi, [1, 10] proposent des implémentations BDI de OCC. Le modèle de double appraisal proposé dans FAtiMA [3],

bien qu'il ne repose pas sur un modèle logique, est un premier exemple de théorie de l'esprit basée sur OCC. Notre objectif est de définir un modèle logique de théorie de l'esprit orienté vers les émotions, en s'appuyant sur les modèles logiques BDI comme [15] et [10] qui proposent une formalisation du raisonnement sur les états mentaux de l'interlocuteur.

En sciences humaines, un débat subsiste sur la nature des mécanismes de théorie de l'esprit : les défenseurs de la *theory-theory*(TT) postulent que la ToM s'appuie sur des règles de sens commun[7], alors que les partisans de la *simulation-theory* (ST) [12] defendent l'idée d'une projection dans l'état mental de l'interlocuteur. Plusieurs travaux ont montré qu'aucune de ces deux visions n'étaient suffisante [27], ce qui a donné naissance à des approches hybrides [7][12].

Les modèles informatique de la ToM choisissent en général l'une ou l'autre des approches. Ainsi, [3] repose sur une approche ST alors que [6, 23] se situent dans une approche TT. Lorsqu'elles sont combinées (comme dans [14]), elles sont implémentées de manière disjointes. Notre proposition est de définir un modèle logique qui combine de manière naturelle ces deux approches au sein d'un même moteur de raisonnement.

3 Architecture et modèle logique

Notre architecture, illustrée sur la figure 1, est composée des éléments suivants. Les états mentaux sont les croyances, attitudes, buts et intentions des agents. Les croyances portent sur des faits, des règles du monde (au sens de la TT) et sur les états mentaux des autres agents. Les attitudes décrivent l'évaluation de l'état du monde et, par extension, ses buts. Dans notre modèle les intensions ne portent que sur la prochaine action. Le moteur d'inférence comprend un modèle délibératif de type *folk-psychology* (au sens de la ST) qui permet de mettre à jour les croyances de l'agent, un modèle de raisonnement de sens commun (au sens de la TT, dans laquelle on retrouve des règles spécifiques au domaine, la simulation d'entretien d'embauche dans notre cas) et enfin le modèle affectif basé sur OCC.

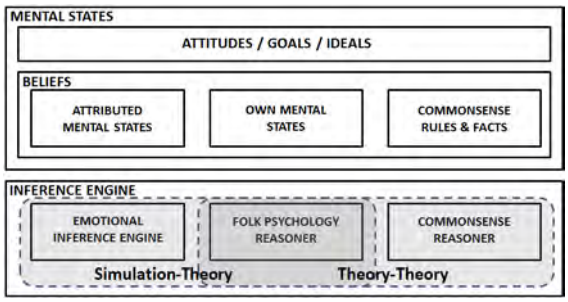


FIGURE 1 – Notre architecture hybride de ToM.

3.1 Syntaxe du modèle logique

Soit ATM un ensemble de propositions décrivant des faits (par exemple : "le salaire proposé est élevé"), ACT un ensemble d'action (par exemple : "se présenter"), ILL un ensemble d'actes de langages, AGT un ensemble d'agents (l'agent virtuel, son interlocuteur et éventuellement d'autres individus), EMO un ensemble de catégories d'émotions. Nous appelons *événements* les actions dans lesquelles l'un des interlocuteurs prend part, comme dans [18]. Un événement $e \in EVT$ est un tuple dans $AGT \times AGT \times (ACT \cup ILL(ATM))$ représentant l'agent qui effectue l'action, celui qui la subit et enfin l'action elle-même, qui peut être aussi un acte de langage (par exemple : "dire que *le salaire est élevé*"). Nous y ajoutons un degré de plausibilté comme cela se fait habituellement en BDI.
Notre langage est défini par la grammaire suivante :

$$\begin{aligned} &Evt : \epsilon ::= \langle a, (a|\varnothing), \alpha\rangle \mid \langle a, a, Spk(\varsigma, \varphi)\rangle \\ &Prp : \pi ::= p \mid \epsilon \mid Like^k_{a,b} \mid Dom^k_{a,b} \\ &Fml : \varphi ::= \pi \mid Bel^l_a(\varphi) \mid Att^k_a(\varphi) \mid Int_a(\varphi) \mid \\ &\qquad Emo^i_{a,(b|\varnothing)}(\varepsilon, \varphi) \mid N(\varphi) \mid U(\varphi, \varphi) \mid \neg\varphi \mid \varphi \wedge \varphi \end{aligned} \tag{1}$$

avec $a, b \in AGT$, $\alpha \in ACT$, $p \in ATM$, $\epsilon \in EVT$, $\varepsilon \in EMO$, $\varsigma \in ILL$, $l, i \in [0, 1]$, $k \in [-1, 1]$. $Like$, Dom, Bel, Att et Int sont des opérateurs de la logique modale et N et U sont les opérateurs temporels *Next* et Until de la logique temporelle LTL et CTL* [22]. Les autre opérateur temporels F et G ainsi que les opérateurs booléens $\top$, $\bot$, $\vee$ et $\Rightarrow$ sont définis de manière classique. De plus, dans la description des événements, nous autorisons l'utilisation de "$-$" pour désigner n'importe quel atome.
Notre modèle de relation sociale est basé sur [17] : $Like^k_{a,b}$ détermine le degré d'appréciation et $Dom^k_{a,b}$ le degré de dominance.
$Bel^l_a(\varphi)$ décrit une croyance, comme dans [10], et se lit "l'agent a croit que φ avec une certitude l". De même, $Att^k_a(\varphi)$ décrit une attitude. Dans notre modèle, cet opérateur sera utilisé pour décrire des désirs, des idéaux et des buts, qui seront représentés avec leurs propres opérateurs modaux comme dans [1, 13].
Dans la suite, nous utiliserons l'opérateur $\stackrel{\text{def}}{=}$ pour la définition de nouveaux opérateurs, alors que l'opérateur $\stackrel{\text{def}}{\Longrightarrow}$ sera utilisé pour décrire nos règles d'inférences. Ainsi :

$$\begin{aligned} Des^k_a(\varphi) &\stackrel{\text{def}}{=} Att^k_a(F(\varphi)) \\ Ideal^{k>0}_a(\varphi) &\stackrel{\text{def}}{=} Att^{k>0}_a(G(\varphi)) = Des^{-k<0}_a(\neg\varphi) \end{aligned} \tag{2}$$

Un désir est quelque chose envers lequel l'agent a une attitude positive, alors qu'un idéal est quelque chose que l'agent souhaiterait toujours vrai. L'objet d'une attitude, d'un désir ou d'un idéal peut être un atome ("préserver la forêt") ou une formule plus complexe comme une croyance ou un opérateur temporel.

Comme en BDI [24], nous notons $Int_a(\varphi)$ les intentions (plans) d'un agent.
$Emo^i_{a,(b|\varnothing)}(\varepsilon, \varphi)$ représente les émotions. Conformément à [11], une émotion ε est toujours à propos d'un fait $varphi$ et peut être dirigée vers un agent b, avec $\varepsilon \in EMO$ la catégorie émotionnelle et i l'intensité.
Nous introduisons l'opérateur $Resp_a$ pour décrire la responsabilité directe (contrairement à [1, 13], nous ne considérons pas le cas où l'agent est responsable d'une situation qu'il aurait pu éviter) :

$$Resp_a(\epsilon) \stackrel{\text{def}}{=} (\epsilon = \langle a, -, - \rangle) \tag{3}$$

3.2 Semantique

Notre sémantique est basé sur la théorie des mondes possibles. Soit $\mathcal{F} = \langle \mathcal{W}, \mathcal{B}, \mathcal{D}, \mathcal{I}, \mathcal{E} \rangle$ avec :

- W l'ensemble non-vide des mondes possibles,
- $\mathcal{B} : AGT \rightarrow (W \rightarrow 2^W)$ la fonction qui associe à chaque agent $a \in AGT$ et à chaque monde $w \in W$ l'ensemble des mondes accessibles par les croyances $\mathcal{B}_a(w)$,
- $\mathcal{D} : AGT \rightarrow (W \times [0,1] \rightarrow 2^W)$ la fonction qui associe à chaque agent $a \in AGT$ et à chaque monde $w \in W$ avec un degré de désirabilité l l'ensemble des mondes accssibles par les désirs $\mathcal{D}_a(w, l)$,
- $\mathcal{I} : AGT \rightarrow (W \rightarrow 2^W)$ la function qui associe à chaque agent $a \in AGT$ et à chaque monde $w \in W$ l'ensembles des mondes accessibles par l'intention $\mathcal{I}_a(w)$,
- $\mathcal{E} : EVT \rightarrow W$ la fonction qui associe à chaque évenement $\epsilon \in EVT$ le monde résultant.

Soit un modèle $\mathcal{M} = \langle \mathcal{F}, \mathcal{V} \rangle$ (avec $\mathcal{V} : W \rightarrow ATM$ une fonction d'évaluation). Nous notons $\mathcal{M}, w \models \varphi$ le fait que φ est vrai dans w. Les valeurs de vérités sont définies de manière classique par induction :

- $\mathcal{M}, w \models p$ ssi $p \in \mathcal{V}(w)$;
- $\mathcal{M}, w \models \neg\varphi$ ssi $\mathcal{M}, w \models \varphi$ n'est pas vrai ;
- $\mathcal{M}, w \models \varphi \wedge \psi$ ssi $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$;
- $\mathcal{M}, w \models Bel^l_a(\varphi)$ ssi $\frac{card(\mathcal{GB}_a(w))}{card(\mathcal{B}_a(w))} = l$ avec $\mathcal{GB}_a(w) = \{v \in \mathcal{B}_a(w) ; \mathcal{M}, v \models \varphi\}$;
- $\mathcal{M}, w \models Des^l_a(\varphi)$ ssi $\mathcal{M}, v \models \varphi \; \forall v \in \mathcal{D}_a(w, l)$;
- $\mathcal{M}, w \models Int_a(\varphi)$ ssi $\mathcal{M}, v \models \varphi \; \forall v \in \mathcal{I}_a(w)$;
- $\mathcal{M}, w \models \epsilon$ ssi $\mathcal{M}, v \models \top \; \forall v \in \mathcal{E}(\epsilon)$;

Dans la prochaine section, nous présentons quelques règles d'inférence de notre modèle. Pour des raisons de place, toutes les règles ne peuvent pas être décrites dans cet article mais nous présentons les plus significatives. Lorsque cela sera nécessaire, nous noterons f la combinaison des degrés de croyances, d'intensité émotionnelle, etc qui sera décrite dans la section 4.

3.3 Quelques règles

Comme dans [10, 1], les relations de croyance $\mathcal{B}$ sont transitive et euclidiennes :

$$Bel^l_a(\varphi) \stackrel{\text{def}}{\Longrightarrow} Bel^1_a(Bel^l_a(\varphi)) \tag{4}$$

Ainsi, les agents sont conscients de leurs propres états mentaux et de leurs relations sociales. Cependant, choisissons que $\mathcal{B}$ ne forme pas une relation binaire :

$$Bel^l_a(\varphi) \stackrel{\text{def}}{\Longrightarrow} Bel^{1-l}_a(\neg\varphi) \tag{5}$$

Nous définissons $0.5 < mod\_th < str\_th$ deux limites pour représenter le fait que l'agent croit peu ($l < mod\_th$), moyennement ($mod\_th < l < str\_th$) ou fortement ($l > str\_th$) quelque chose. Enfin, nos agents sont capables de déductions :

$$Bel^l_a(\psi) \wedge Bel^{l'}_a(\psi \Rightarrow \varphi) \stackrel{\text{def}}{\Longrightarrow} Bel^{f(l,l')}_a(\varphi) \tag{6}$$

Les attitudes peuvent être positives ou négatives mais on les suppose consistants :

$$\mathcal{M}, w \models (Att^k_a(\varphi) \wedge Att^{k'}_a(\neg\varphi)) \text{ iff } k = -k' \tag{7}$$

Toutefois, un agent peut désirer quelque chose qui conduit à quelque chose de non-désiré. C'est au moment du choix d'action que ce conflit sera géré :

$$Des^k_a(\varphi) \wedge Bel^{l>str\_th}_a(\psi \Rightarrow F(\varphi)) \wedge \neg IncDes^k_a(\psi) \stackrel{\text{def}}{\Longrightarrow} N(Des^k_a(\psi)) \tag{8}$$

avec $IncDes^k_a(\varphi)$ le désir inconsistant :

$$IncDes^k_a(\varphi) \stackrel{\text{def}}{=} (Bel^{l>str\_th}_a(\varphi \Rightarrow \neg\psi) \wedge Des^{k'>0}_a(\psi)) \vee (Bel^{l>str\_th}_a(\varphi \Rightarrow \psi) \wedge Des^{k'<0}_a(\psi)) \tag{9}$$

Ainsi, le désir φ est inconsistant si l'agent croit fortement qu'il conduit à un fait indésirable ψ. Conformément au modèle BDI [24], nous définissons les buts comme des désirs consistants et qu'il croit atteignables. Pour cela, nous introduisons la limite des_th :

$$Goal^{k>0}_a(\varphi) \stackrel{\text{def}}{=} Des^{k>des\_th}_a(\varphi) \wedge Bel^l_a(F(\varphi)) \wedge \neg IncDes^k_a(\varphi) \tag{10}$$

Enfin, un but est transformé en intention lors que l'agent peut l'atteindre :

$$Goal^{k>0}_a(\epsilon) \wedge Resp_a(\epsilon) \stackrel{\text{def}}{\Longrightarrow} N(Int_a(\epsilon)) \tag{11}$$

ou, comme dans [6], parce qu'il croit qu'il existe un moyen de l'atteindre :

$$Goal^{k>0}_a(\varphi) \wedge Bel^{l>str\_th}_a(\psi \Rightarrow F(\varphi)) \wedge \neg IncDes^k_a(\psi) \wedge Bel^{l'}_a(F(\psi)) \stackrel{\text{def}}{\Longrightarrow} N(Int_a(\psi)) \tag{12}$$

Lors qu'un agent a une intention et peut la réaliser, il le fait :

$$Int_a(\varphi) \wedge Bel^{l>str\_th}_a(\psi \Rightarrow F(\varphi)) \stackrel{\text{def}}{\Longrightarrow} Int_a(\psi) \qquad Int_a(\epsilon) \wedge Resp_a(\epsilon) \stackrel{\text{def}}{\Longrightarrow} N(\epsilon) \tag{13}$$

Dans notre modèle, comme dans [18, 9, 3], les attitudes sont influencées non seulement par les croyances, mais aussi par la relation sociale :

$$\begin{aligned} Bel_a^{l>str\_th}(\varphi) \wedge Att_a^k(F(\varphi)) \wedge Bel_a^{l'}(Att_b^{k'}(F(\varphi))) \\ \wedge Like_{a,b}^h \wedge Dom_{a,b}^{h'} \stackrel{\text{def}}{\Longrightarrow} Att_a^{f(k,k',h,h')}(\varphi) \\ Bel_a^{l>str\_th}(Des_b^k(\varphi)) \wedge Like_{a,b}^{k'>0} \stackrel{\text{def}}{\Longrightarrow} N(Des_a^{f(k,k')}(\varphi))) \end{aligned} \tag{14}$$

Enfin, nos règles d'appraison sont conformes à ce qui se fait classiquement dans la litérature [1, 13, 10]. Par exemple :

$$\begin{aligned} Bel_a^l(\gamma) \wedge Att_a^{k>0}(\gamma) &\stackrel{\text{def}}{\Longrightarrow} N(Joy_a^{i=f(l,k)}(\gamma)) \\ Bel_a^l(F(\gamma)) \wedge Des_a^{k<0}(\gamma) &\stackrel{\text{def}}{\Longrightarrow} N(Fear_a^{i=f(l,k)}(\gamma)) \\ Bel_a^l(\gamma) \wedge Ideal_a^k(\gamma) \wedge Bel_a^{l'}(Rsp_b(\gamma)) & \\ &\stackrel{\text{def}}{\Longrightarrow} N(Admiration_{a,b}^{i=f(l,l',k)}(\gamma)) \end{aligned} \tag{15}$$

Enfin, les règles de sens commun dépendent du domaine. Dans le contexte de l'entretien d'embauche, nous aurons par exemple :

$$\begin{aligned} &Des_r^{0.77}(\neg Emo_{\_,c}^i(distress)) \wedge i > 0.5 \\ &Bel_r^{0.8}(Att_c^{-0.5} salary\_is\_bad) \end{aligned}$$

Pour représenter un agent qui ne souhaite pas rendre le candidat triste et qui pense que les candidats souhaitent généralement obtenir un bon salaire.

4 Implémentation

Le modèle théorique que nous avons présenté dans la section précédent est censé être indépendant du domaine, mis à part les règles de sens commun. Dans notre implémentation pour la simulation d'entretien d'embauche, dans le cadre du projet TARDIS [2]. L'intérêt de l'entretien d'embauche pour la validation de notre modèle est qu'il s'agit de situations de dialogue diadique semi-structurés où le recruteur a souvent la possibilité de raisonner sur les états mentaux et les émotions du candidat.

Notre cadre logique et le moteur d'inférence ont été impémentés en SWI-Prolog. Ce moteur a été couplé avec un programme C++ qui gère le tour de parole et la communication entre les modules. Conformément au modèle BDI, à chaque tour, l'agent interprète les émotions exprimées par son interlocuteur pour générer une liste d'actions possibles, en sélectionner une pour mettre à jour ses intentions et les exécuter.

La difficulté lors de l'implémentation de notre modèle est de fixer les limites pour les croyances et les buts, et de définir les fonctions de combinaison que nous avons noté f dans les formules de la section précédente. Dans notre implémentation, nous avons choisi de manière arbitraire de fixer : $mod\_th = 0.5$, $str\_th = 0.75$ et $des\_th = 0.7$.

Les fonctions de combinaison peuvent être regroupées en deux catégories :

– Pour la dynamique des attitudes et des croyances (par exemple l'équation 14), nous utilisons une simple moyenne suivie d'une normalisation :

$$f(k, k') = ((k + k')/4) + 0.5$$

– Pour les émotions (équation 15), nous combinons une influence linéaire de l'attitude (par exemple, la joie est linéairement corrélée à l'attitude envers l'objet de l'émotion) avec une influence logarithmique de degré de croyance. Ainsi, nous obtenons des émotions plus fortes avec des croyances relativement faibles :

$$f(l, k) = \frac{k}{2} \times \frac{Log(2l - 1) - min}{min} + 0.5$$

où min est la limite de $Log(x)$ lorsque $x \to 0$, soit la plus petite valeur possible dans l'ordinateur. Le facteur $2l - 1$ permet d'ajuster la valeur dans $[0, 1]$ avant le calcul de l'intensité, puis nous le réajustons dans $[0.5, 1]$ pour obtenir des intensités plus significatives.

Les règles de sens commun correspondant à la situation d'entretien d'embauche définissent l'ensemble des actes de dialogue (parler du salaire, de l'expérience) et des attentes en termes d'impact (dire au candidat qu'il est en retard devrait le mettre mal à l'aise). L'agent choisit alors les actions à effectuer, c'est-à-dire les actes de langage, en fonction de ses buts courants (en terme d'état affectif de l'interlocuteur et de sujets à aborder : par exemple, je souhaite mettre l'interlocuteur à l'aise mais je dois aborder la question du salaire). De plus, nous avons définis des opérateurs spécifiques pour décrire la confiance en soi, la motivation et la compétence professionnelle du candidat. Les degrés de croyances pour ces faits sont calculés en fonction de ses réactions émotionnelles aux questions de l'agent à l'aide de règles simple (de type TT). Par exemple, des hésitations lors de la réponse à une question sont un indice de non-confiance en soi.

La perception des états affectifs à travers le comportement non-verbal de l'interlocuteur est gérée par un autre module dans le projet TARDIS(voir [2]). Dans cet article, comme nous le verrons, nous supposons que nous avons des entrées numériques dont on ne sait pas comment elles ont été obtenues à partir de capteurs.

5 Évaluation préliminaire

Notre premier objectif était d'étudier l'impact d'un tel modèle de théorie de l'esprit sur la qualité ou la difficulté d'un entretien d'embauche avec un agent virtuel. Pour cela, nous avons évalué notre modèle en situation avec 30 sujets – 11 femmes et 19 hommes – issus du personnel de notre laboratoire, qui ont interagit à travers une interface graphique simplifiée avec notre modèle logique. Aucun agent virtuel

n'était présent : l'état mental du recruteur et ses croyances en terme de confiance en soi, de motivation et de compétence du candidat étaient représentés par des barres de progression (dont les valeurs allaient de -1 à 1). De même, aucun capteur n'était utilisé : les utilisateurs devaient saisir leur comportement à l'aide de 8 indicateurs (ascenseur à placer entre 0 et 1) : soulagé, embarrassé, hésitant, stressé, mal à l'aise, concentré, agressif, je m'ennuie. Le choix de ces indicateurs provient des recherches du projet TARDIS [2].

Les sujets devaient jouer le rôle du candidat à un poste de secrétaire et pouvaient adopter la personnalité qu'ils souhaitaient. à l'issue de l'entretien, ils devaient remplir un questionnaire de 11 items (sur une échelle de Lickert à 5 valeurs) portant sur la crédibilité des réactions et la qualité des évaluations faites par l'agent. Chaque sujet interagissait avec l'une des trois versions possible de l'agent : celui dont la base de connaissance avait pour objectif de mettre le candidat à l'aise (PROFIL_A), celui qui posait des questions classiques sans but précis sur l'état mental de l'utilisateur (PROFIL_B) et celui qui posait des questione embarrassantes (PROFIL_C). Les 3 agents utilisent le même moteur de raisonnement.

Notre première hypothèse est que la variation du profil aura un impact sur la réaction émotionnelle des candidats (du moins, telle qu'elle est exprimée par les 8 indicateurs manipulés par le sujet). Notre mesure porte sur la somme des intensités émotionnelles exprimées.

5.1 Resultats

Nos résultats montrent, suivant un test de Kruskal-Wallis, un effet principal du profil (PROFIL_X) sur la somme des intensités émotionnelles ($Chi^2(2, 629) = 11.435; p < 0.01$) et, en particulier, sur l'embarras exprimé ($Chi^2(2, 629) = 6.231; p < 0.05$) et sur la concentration exprimée ($Chi^2(2, 629) = 9.218; p < 0.01$). Cela signifie que le profil du recruteur a un impact sur les affects décrits (et peut-être exprimés en situation réelle) par les sujets. Un test de Mann-Whitney montre aussi que test les participants qui ont interagit avec le profil A (compréhensif) choisissent plus d'embarras et plus de concentraction, alors que ceux qui ont interagit avec le profil C (difficile) choisissent plus de stress, de "mal à l'aise" et de concentration. Nos résultats sont illustrés sur la figure 2.

5.2 Discussion

La théorie de l'esprit est un phénomène complexe qui fait intervenir d'autres processus cognitifs (dont la mémorisation, les capacités de raisonnement, etc) et perceptifs (*e.g.* l'interprétation des signaux sociaux). C'est pourquoi il est difficile non seulement de la modéliser, mais aussi de l'évaluer. Un protocole simple comme celui que nous avons présenté ici n'est pas suffisant pour évaluer l'impact de la ToM sur la qualité de l'entraînement à l'entretien d'embauche. Pour commencer, l'utilisation d'une interface graphique et non d'un agent virtuel ne permet pas aux utilisateurs de s'immerger dans la situation. Notre hypothèse est

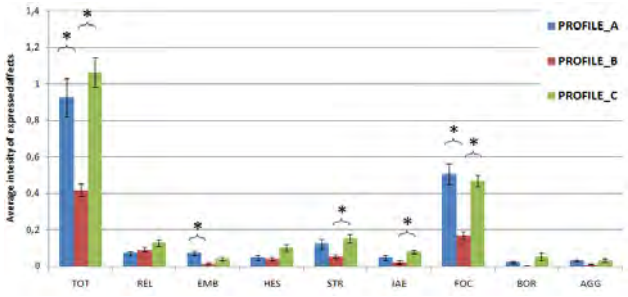


FIGURE 2 – Intensité moyenne et écart-type choisi par les participants en fonction du profil et des affects considérés. Les effets significatifs (*) sont sur les affects embarras, stress, mal à l'aise, concentré et sur la somme totale des intensités.

que le dispositif complet, tel qu'il est proposé dans le projet TARDIS, devrait permettre d'évaluer la réactivité et les processus de raisonnement de l'agent virtuel. Cependant, dans la litérature, il n'existe pas de protocole expérimental pour l'évaluation de la qualité d'une théorie de l'esprit dans une situation d'interaction. Les travaux les plus proches [14, 23] évaluent la capacité de la ToM artificielle à expliquer les choix dans le scénario en comparant les décisions avec le modèle de la tâche. Mais ces modèles ne portent que sur la partie verbale de l'interaction, alors que notre ToM s'intéresse à la composante co-verbale (par la sélection d'actes expressifs).

Cependant, notre étude préliminaire nous a permis de mettre en évidence quelques résultats pour l'évaluation d'un tel modèle dans un contexte de simulation d'entretien. Tout d'abord, le fait que le profil B (neutre) n'utilise pas de théorie de l'esprit pour sélectionner les questions qu'il pose conduit à des réactions affectives moins importantes. Il semblerait donc que la ToM sur les émotions dans le contexte de l'interaction dialogique ait un réel impact sur la réaction de l'interlocuteur. De plus, cette étude montre la nécessité de disposer de différents profils de recruteurs (au niveau du modèle de raisonnement, pas au niveau de l'apparence ou du comportement non-verbal) pour éliciter des émotions différentes.

L'utilisation d'un modèle logique présente un autre intérêt majeur pour l'entraînement des candidats à l'emploi : chaque déduction peut être retrouvée, expliquée et fournie à l'utilisateur, pour lui permettre de mieux comprendre l'impact de ses réactions sur les états mentaux de l'agent virtuel, et l'interprétation qu'il a fait de l'état mental du candidat. C'est pourquoi nous poursuivons actuellement nos travaux pour l'intégration de modèles logiques de ToM au sein d'agents virtuels.

Références

[1] C. Adam, A. Herzig, and D. Longin. A logical formalization of the OCC theory of emotions. *Synthese*, 168(2) :201–248, Feb. 2009.

[2] K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, H. Jones, M. Ochs, C. Pelachaud, K. Porayska-Pomsta, R. Paola, and N. Sabouret. The TARDIS framework : intelligent virtual agents for social coaching in job interviews. *Proceedings of the Tenth International Conference on Advances in Computer Entertainment Technology (ACE-13). Enschede, the Netherlands. LNCS 8253*, page in press, 2013.

[3] R. Aylett and S. Louchart. If I were you : double appraisal in affective agents. In *Proceedings of the 7th international joint conference on Autonomous Agents and MultiAgent Systems*, pages 1233–1236, 2008.

[4] S. Baron-Cohen. *Mindblindness : An essay on autism and theory of mind*. MIT press, 1997.

[5] L. M. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer. Cicero - towards a multimodal virtual audience platform for public speaking training. In *Proc. 2013 International Conference on Intelligent Virtual Agents*, pages 116–128, 2013.

[6] T. Bosse, Z. A. Memon, and J. Treur. A recursive BDI agent model for Theory of Mind and its applications. *Applied Artificial Intelligence*, 25(1) :1–44, 2011.

[7] G. Botterill and P. Carruthers. *The philosophy of psychology*. Cambridge University Press, 1999.

[8] J. Bynner and S. Parsons. Social Exclusion and the Transition from School to Work : The Case of Young People Not in Education, Employment, or Training (NEET). *Journal of Vocational Behavior*, 60(2) :289–309, Apr. 2002.

[9] C. Castelfranchi. Modelling social action for AI agents. *IJCAI'97 Proceedings of the Fifteenth international joint conference on Artifical intelligence - Volume 2*, 103(1) :1567–1576, 1997.

[10] M. Dastani and E. Lorini. A logic of emotions : from appraisal to coping. In *Proceedings of the 11thiInternational conference on Autonomous Agents and Multiagent Systems*, pages 1133–1140, 2012.

[11] N. H. Frijda. *The emotions*. Cambridge University Press, 1986.

[12] A. I. Goldman. *Simulating minds : The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press, 2006.

[13] N. Guiraud, D. Longin, E. Lorini, S. Pesty, and J. Rivière. The face of emotions : a logical formalization of expressive speech acts. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, pages 1031–1038, 2011.

[14] M. Harbers. Explaining agent behavior in virtual training. *SIKS dissertation series*, 2011(35), 2011.

[15] A. Herzig and D. Longin. A logic of intention with cooperation principles and with assertive speech acts as communication primitives. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems : part 2*, pages 920–927. ACM, 2002.

[16] M. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. Picard. MACH : My Automated Conversation coacH. In *Proc. 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM Press, 2013.

[17] T. F. Leary et al. *Interpersonal diagnosis of personality*. Ronald Press New York, 1957.

[18] M. Ochs, N. Sabouret, and V. Corruble. Simulation of the Dynamics of Nonplayer Characters' Emotions and Social Relations in Games. *Computational Intelligence and AI in Games, IEEE Transactions on*, 1(4) :281–297, 2009.

[19] A. Ortony, G. L. Clore, and A. Collins. The Cognitive Structure of Emotions, 1990.

[20] A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobreperez, S. Woods, C. Zoll, and L. Hall. Caring for agents and agents that care : Building empathic relations with synthetic agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 194–201, Washington, DC, USA, 2004. IEEE Computer Society.

[21] L. Pareto, D. Schwartz, and L. Svensson. Learning by guiding a teachable agent to play an educational game. *in Education Building Learning*, pages 1–3, 2009.

[22] A. Pnueli. The temporal logic of programs. In *Foundations of Computer Science, 1977., 18th Annual Symposium on*, pages 46–57. IEEE, 1977.

[23] D. V. Pynadath, N. Wang, and S. C. Marsella. Are you thinking what I'm thinking ? An Evaluation of a Simplified Theory of Mind. In *Intelligent Virtual Agents*, pages 44–57. Springer, 2013.

[24] A. S. Rao and M. P. Georgeff. Modeling Rational Agents within a BDI-Architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, 1991.

[25] K. R. Scherer. Emotion and emotional competence : conceptual and theoretical issues for modelling agents. *Blueprint for Affective Computing*, pages 3–20, 2010.

[26] M. Sieverding. 'Be Cool !' : Emotional costs of hiding feelings in a job interview. *International Journal of Selection and Assessment*, 17(4), 2009.

[27] K. Vogeley, P. Bussfeld, A. Newen, S. Herrmann, F. Happe, P. Falkai, W. Maier, N. J. Shah, G. R. Fink, and K. Zilles. Mind reading : neural mechanisms of theory of mind and self-perspective. *Neuroimage*, 14(1) :170–181, 2001.

Subjective Evaluation of a BDI-based Theory of Mind model

Atef Ben Youssef Nicolas Sabouret Sylvain Caillou

LIMSI-CNRS, UPR3251, Université Paris-Sud, 91405 Orsay

Résumé

La théorie de l'esprit (ToM) joue un rôle important dans les interactions affectives et plusieurs modèles logiques ont été proposé dans le domaine des Agents Conversationnels Animés pour améliorer leur crédibilité. Cependant, l'intégration d'un tel modèle avec un agent virtuel et l'évaluation du résultat restent des questions ouvertes. Dans cet article, nous présentons l'évaluation d'un modèle BDI de théorie de l'esprit basé sur une étude subjective. Nous présentons tout d'abord le principe général du modèle de ToM considéré et les dimensions que nous avons évaluées. Nous présentons ensuite notre protocole et nous montrer que l'utilisation d'un modèle de ToM améliore la crédibilité de l'agent virtuel, à la fois au niveau cognitif et au niveau expressif.

Mots Clef

Évaluation, Théorie de l'Esprit, Interaction.

Abstract

Theory of Mind (ToM) plays an important role in affective interactions and several logic-based models of ToM have been proposed in the literature to enhance the credibility of intelligent virtual agents. However, the evaluation of the impact of such a model remains a difficult question. In this paper, we present an evaluation of a Belief-Desire-Intension (BDI)-based ToM model using a subjective study based on human-agent interaction. We first briefly present the main principles of the considered ToM model and the dimensions to evaluate. We then present our protocol and we show that the use of the ToM model improved the believability of the vitual agent, both at the cognitive and at the expressive level.

Keywords

Evaluation, Theory of Mind, Human-Agent Interaction.

1 Introduction

During the last decade, intelligent virtual agent have become able to express more and more complex behaviour, including affective reactions to complex situations [15, 17, 9] and to user interactions [14, 20]. For instance, in the Semaine platform [20], the virtual agent is able to adopt a mimickery behaviour to simulate active listening behaviour while expressing different personalities. Such agents' behaviours generally rely on reactive models in which the actions and facial expressions are directly a result of the inputs (perceived affects or contextual information).

Several approaches in human-agent interaction try to complete the reactive behaviour of virtual agents using cognitive models such as the appraisal theory [15], based on the interpretation of the interaction situation. Some of these models also consider the mental state of the interlocutor [19, 4]. In these models, the agent is capable to represent and reason not only on the situation, but also on the interpretation of the situation by the interlocutor and the resulting affects. This cognitive process is called Theory of Mind (ToM) [5] and several authors have shown that it plays an important role for the appraisal of complex empathetic emotions (sorry-for, gloating, etc) [19, 18].

In addition to the affective reasoner, a ToM model for Human/Agent interaction must be able to represent and reason about the communicative intention and its relation to the social attitude, ranging from cooperation and competition to social manipulation, as was pointed out by [8]. One common proposal to support such adaptability is to rely on the Belief-Desire-Intention (BDI) model: several computational models of emotions have already been proposed (e.g. [1, 11]) that show that the BDI model is a good basis to represent and to reason about the interlocutor's mental state. The underlying idea of such models is that the combination of ToM and BDI logics in affective computing makes leads to a more realistic mental state of the agent. Hopefully, this should lead to more realistic reactions at the expressive level, compared to models that directly map the recorded signal of the interlocutor behaviours to the agent's one and ignore all other levels of the interaction such the internal reasoner.

However, the evaluation of the real impact of a ToM+BDI model on the believability of the virtual agents still remains an open question. As will be discussed in the next section, models are evaluated in terms of expressiveness, but not in the context of real human-agent interaction. In this paper, we first discuss the difficulties that arise for the evaluation of logic-based ToM models (section 2). We then present briefly the two models that we used in our evaluation: a ToM+BDI model and a simple rule-based reference model (section 3). We then present our evaluation study (section 4) and the experimental results (section 5). We show what can be learned from such experience both for the improvement of a BDI model of ToM and for the evaluation of logical models for virtual agents in general.

2 Positionning and related work

2.1 Evaluation of a ToM model

Theory of Mind is a complex process that relies on various other cognitive and perceptual processes. It is not only hard to model but also to assess. In the human-science literature, there exist validated methods to evaluate whether subjects – generally children – have ToM abilities and use them [7]. Transposing these methods to logic-based models and, ultimately, to interactive virtual agents, raises several difficulties. The work by Harbers on virtual agents' actions explanation in the context of human training [13] points out some of these difficulties from the computational point of view. In this work, the course of events and the agent's actions and explanations must be specified in advance for different scenarios. In this context, the ToM model is evaluated based on whether it matches these specifications. This is not sufficient to validate its logical structure or its impact on the agent's flexibility. Similarly, Pynadath et al. [19] build expectations about user's actions – based on formal models in the specific context of wartime negotiations – in order to model a simplified ToM. They compare these expectations with the user's actual behavior using logical rules from the appraisal theory. This idea of double-appraisal is already present in [4] and one originality of Pynadath et al. is to extend it and to use their reverse-appraisal to validate the implementation of ther ToM. However, the evaluation of such a model only validates the task model that was used for the appraisal, and not the validity of the ToM w.r.t. the interlocutor's understanding of the interactional situation.

In general, extending computational evaluation of ToM models to human/agent interaction situation requires to get rid of a complete description of the task, since human interaction and dialogue is hardly represented by a clear, logical and deterministic model [21]. We then face the following dilemma. On the one hand, ToM evaluation protocols proposed in the psychology literature generally do not integrate the interactional aspect: the evaluation is about the understanding of what the character feels, not about what it might think about an interaction situation. On the other hand, computer-science models for ToM evaluation tend to focus on the quality of the task model recognition or user preferences detection, not on the ToM itself.

Our aim is to evaluate how the complex representation of general rules about the dialogue task in BDI, coupled with a representation of each interlocutor's goals and expectations, can help the agent produce complex affects and show more realistic expression during the interaction. This is the reason why, unlike other mentionned evaluation protocols, we propose to use a subjective approach in which the user does not need to make any interpretation of the virtual agent's beliefs and goals: the participant will not need a theory of the agent's mental state and reasoning. To this purpose, we propose to consider the context of a job interview dialogue. Job interviews are a good example of semi-structured dyadic interactions where recruiters have several opportunities to reason about candidates' mental and affective states. Moreover, the use of virtual agents for job interview simulation appears as a promising way to increase applicant's self-confidence [14, 2].

2.2 Affect recognition

In addition to these conceptual difficulties, several technical limits come that make the setup of an evaluation protocol even more difficult. In order to bring any contribution, a logical model requires inputs in terms of detected affects. However, speech recognition to provide semantic information to the logical model is generally not feasible. Current techniques for audio-visual emotion recognition [22] are complex to use and show good results in controlled situations. Similarly, the expression of affects through facial expression, voice and body gesture remains a difficult research problem and studies show that human beings have difficulties in identifying the expressed affects [3]. However, in the context of human-computer interaction, it seems not possible to get rid of these difficulties. Our approach consists in having the user "enter" its affective state by positionning a set of bars that represent the intensity of their affect's (one for each affect). This is not entirely satisfactory but it allows us to get rid of the affect recognition difficulties.

2.3 Affect expression

As far as mental state expression is concerned, we have tried different approaches. In a previous study [6], we used a very simple Graphical User Interface (GUI) to show the outputs of our ToM: the mental state of the virtual agent was represented by 4 bars (see figure 1). This only gives to the user a very limited representation of the agent's mental state. As one could have expected, our study showed that the participants were not able to figure out the dynamics of the agent's reasoning and to distinguish it from a simple reactive model: the ToM, in this interaction setting, did not appear more credible than a simple hard-coded linear combination of inputs. For these reasons, we propose in this paper to evaluate the perception of the mental state based on the expressions of a virtual agent. We express the affective output of our ToM model using a male virtual avatar designed using the Multimodal Affective and Reactive Character (MARC) system [10].

3 Comparison of two models

To evaluate the performance of our ToM, we plugged two different cognitive models to an agent to compare their impact on the job-interview interaction: our BDI-based model, implemented in Prolog, and a simple reactive model (which plays the role of placebo in our validation experiment) in which the ToM outputs are the result of linear function of the affective inputs (weighed arithmetic mean). These two models are presented hereafter.

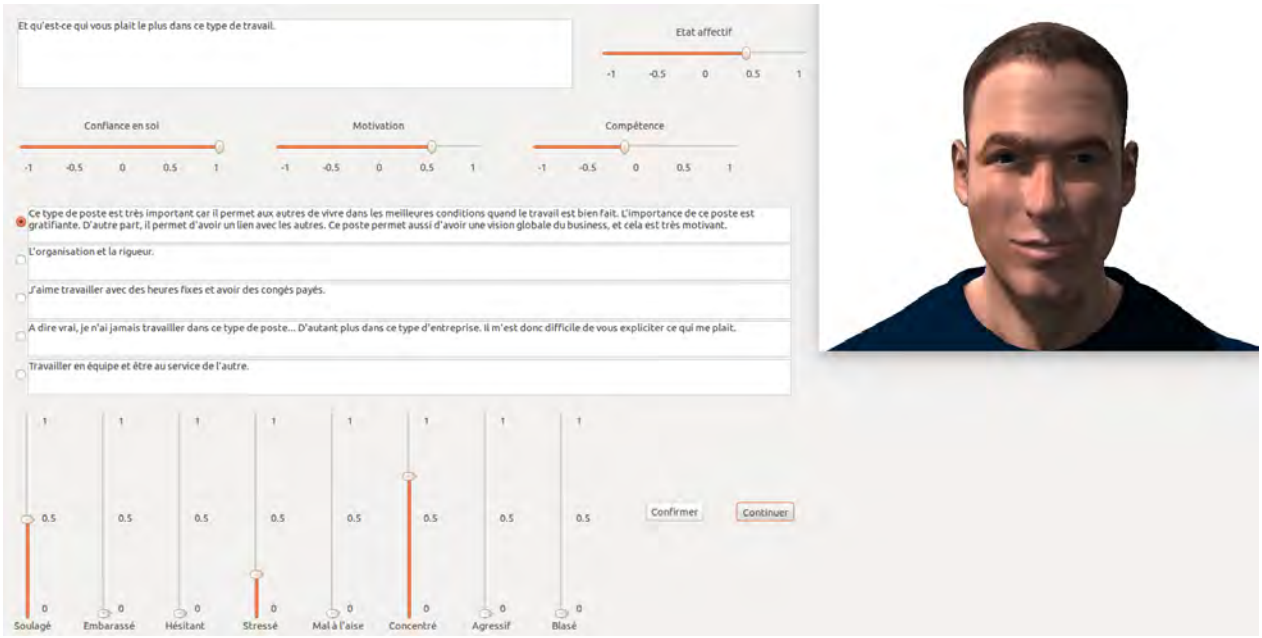


Figure 1: Graphical User Interface (GUI) used on the subjective evaluation

This figure shows the bar-based interface (on the left) and the MARC agent (on the right). In the bar-based interface, the user can enter its affective state at each step of the interaction (lower part), while selecting an answer (central part). He/she can also see a simple representation of the agent's mental state represented by 4 bars (upper part). In this study, the user were told to ignore the 4 bars and to concentrate on the MARC agent's expression, which was displayed in fullscreen mode

3.1 Theory of Mind (ToM) model

The detailed logical rules of the ToM model was previously published by Belkaid et al. [6]. In this paper, we focuses on its evaluation. Note that it is important to have in mind the data it manipulates about the interactional situation. The core of the model is about dialogue selection in human-agent interaction. All task-related information (*i.e.* what are the possible dialogues, what are the general rules about the interaction task) are defined as domain rules. The system uses a deliberative reasoner for intention generation (action selection) and mental state's update (belief revision). The agent's mental state includes not only beliefs, desires (goals) and intentions (dialogical actions) but also *attitudes* toward any logical predicate: a fact, a topic of discussion or even one of the interlocutor's goal. For instance, the agent believes that a compliment to the user about her CV should make him/her more happy. A understanding recruiter agent will have a positive attitude toward believing that the user has a good qualification for the job.
The agent is also provided with an OCC-based [16] model of appraisal for computing affective states for the virtual agent. We use the 8 following emotions: joy, distress, happy-for, sorry-for, admiration, reproach, pride and shame. These emotions correspond to what domain experts reported that could be expressed by recruiter in a job interview context (but the list could be extended, for example with hope/fear). The emotion that is displayed by the agent is selected amongst the highest-intensity emotions (using a probabilistic selection based on the emotion intensities).
In addition to its attitude toward its interlocutor, the agent computes beliefs about the interlocutor's self-confidence, motivation and qualification, based on its reaction to the questions and some domain rules. For instance, hesitating in the job description topic can indicate they are not qualified enough while being focused when introducing themselves denotes a good self-confidence level.

3.2 Placebo model

The so-called *placebo* model serves as a baseline system to evaluate the quality of our ToM model. For the computation of the attitude as for the computation of the beliefs about self-confidence, motivation and qualification, this model uses a simple linear mapping of input affective states:

$$Value(v) = \frac{\sum_{a \in Affects} \alpha_{v,a} Input(a)}{\sum_{a \in Affects} \alpha_{v,a}} \quad (1)$$

with $Affect$ the set of affect categories in inputs, $input(a)$ the input value for affects a and v one of attitude, self-confidence, motivation or qualification. Each $\alpha_{v,a}$ is the weight of input a in the computed value v and they were fixed to -1, 0 or 1. For instance, the input "focused" counts positively for the attitude: $\alpha_{attitude,focused} = 1$.
These variables dynamics is simulated using a simple linear regression:

$$v_{t+1} = 0.3 \times Value(v) + 0.7 \times v_t \quad (2)$$

with v_0 set to the neutral value 0 for all 4 variables (attitude, self-confidence, motivation or qualification).
The placebo does not compute OCC emotions. The output emotion is randomly selected from the 8 emotions: joy, distress, happy-for, sorry-for, admiration, reproach, pride and shame.
The following sections presents in detail the experimental setting and the result of our study.

4 Experimental setting

Figure 1 show the GUI used in this study that couple the inputs and outputs in the same screen. Both models have the same inputs and outputs. They receive a set of intensity values for affect categories (lower part of Figure 1: 2 positive affects and 6 negative ones). The list of the candidate affects was defined in the frame of a larger research project based on 1) a corpus of job interviews in which we could identify the expressed affects of both the candidate and the interviewer) [2] and 2) what can be detected with the state-of-the-art social signal processing systems [22].
The MARC agent model can be easily coupled with cognitive models and it has a rich set of action units that can be triggered to express basic emotions [9]. The output of the model is composed of one emotion in the OCC taxonomy [16] (which represents the reactive emotion to display), a value of attitude (in $[-1, 1]$) representing how the agent feels toward the candidate, and 3 task-specific variables evaluated by the model: the participant's self-confidence, motivation and qualification for the job (in $[-1, 1]$) as seen by the agent.
The goal of our study is to evaluate in what a virtual agent animated using ToM model is either more credible or more engaging than a virtual agent with a simple reactive model (in our case, the placebo system). This is not a trivial answer since many variables come into play: the perception of the user affects, the quality of the information displayed to the user, the task model, the expressiveness of the virtual agents, *etc*. In order to reduce the number of these variables, we consider the following setting:

- The dialogues are scripted and the user select one possible answer among a predefined list at each step. The verbal content is ignored by both the ToM and the placebo models.
- The user affects are not detected. Instead, the user selects within a predefined set of possible affects combination that represents its possible mental states. He/she can change the values at will.
- The questions by the virtual agent are displayed through text, in a separate window, so that the user can focus on the agent's reaction (in terms of emotion and attitude expressions) rather than on the next questions.

The virtual agent plays the role of the recruiter and asks questions to the user about the salary, the experience... Each topic is associated with some expectations about the impact of the question. In the ToM model, the agent selects questions (or topics) based on its current *goals* in terms of affective state for the interlocutor. However, the placebo model does not have this capability. For this reason, in our experiments, we forced both models to follow a scripted interaction, in which questions are asked in a fixed order.
Once the agent has displayed its question, the user is given a set of 5 possible predefined answers, each one associated with a value (in $[0, 1]$) for all the affects in inputs: aggressive, bored, embarrassed, focused, ill at ease, hesitating, relieved and stressed. The proposed answer and the values of affects have been designed based on the result of our preliminary study [6], by selecting a balanced set of answers and affect values given by the participants of this study. Some of these values have been modified to better express the intended attitude and to balance between positive and negative attitudes. However, the user can change the predefined values of the affects by moving the slider.
When he or she submits his/her answer, these values are sent to the AI model (ToM or placebo) which computes the ouputs: emotion, new value of attitude and domain-specific values (self-confidence, motivation and qualification). The virtual agent first expresses the emotion during 2 seconds using a set of action units defined in [12]. It then switches its expression mode toward the attitude: a positive attitude will be expressed using the joy action units, whereas a negative attitude will be expressed using the anger action units, these actions units providing the best rendering in the pretests we have done on the MARC agent.
The participant were told to focus on the agent's expression. Once the agent remains still, they can read the next question and continue the job interview.

5 Results

At the end of simulation, the participants were asked to fill a questionnaire with 6 questions about the quality of the interaction. Each question was rated on a 11 point Likert scale (from not satisfied at all (0) to very satisfied (10)) related to:

- the *difficulty* of the interview,
- the *credibility of the global behaviour* of the recruiter,
- the impression about *taking into account* their *responses* and their *emotional reactions*,
- the *enjoyability of the interaction* with the agent,
- the *utility* of such a simulation with a virtual agent for the preparation of a real job interview,

This test was performed by 10 native-speakers. Five of them interacted with the placebo version of the agent and the other five interacted with the ToM version. The average evaluations over users for the main criteria are shown in Figure 2.

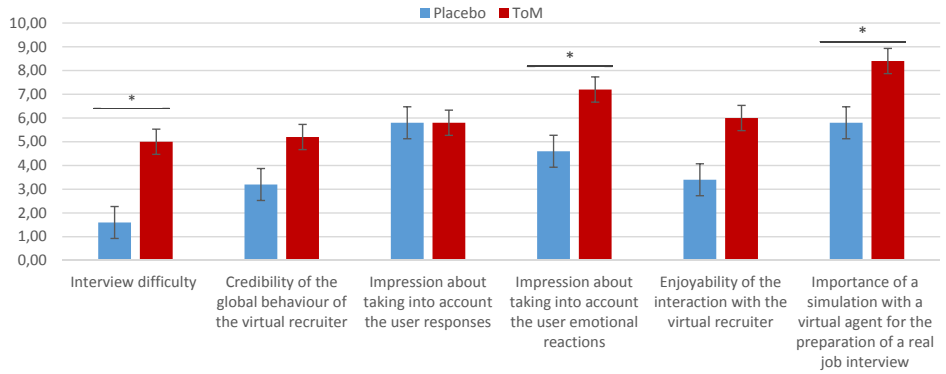


Figure 2: Subjective evaluation results over 10 users. The * denote significant differences ($p < 0.05$) between the two models

Several observations can be made. First, users interacting with ToM agent found, to a large extend, that the interview was harder than those interacting with the placebo. The statistical result is highly significant ($p < 0.05$). This could be explained by the fact that the ToM model gives higher impression that the virtual recruiter is following up the user's response using the displayed affects. There is also a tendency that the users found the global behaviour of the ToM agent more credible than the placebo agent behaviour, but the difference is not as clear as for the difficulty. One reason might be that, in both models, the agent does not take at all into account the verbal content of the participant selected answers (which the subjects were not aware of). This might impact the credibility of the agent, w.r.t. participant's expectations. Yet, these two observations together tend to show that the ToM agent was perceived as more challenging while no clear difference could be made in terms of affective behaviour's credibility.

The second observation is that the ToM model gives a significantly higher impression that the user's reaction were taken into account (which is what we expected).

The third observation is that the users interacting with the ToM agent found it significantly more useful for the preparation of a job interview than those "training" with the placebo. This result alone is surprising since no task indicator (not shown on the figure) was evaluated as more accurate in the ToM agent than in the placebo agent. In other words, users did not make any difference between the quality of the task-specific indicators (self-confidence, motivation and qualification), but they consider the ToM agent as more useful for training (average score of 8.4 for the ToM agent and 5.8 for the placebo). One explanation could reside in the fact that, to some not statistically significant extent, the users evaluated the ToM agent as more pleasant to interact with. We still need to clarify the reasons behind these results, based on the detailed comments of the participants. However, it tends to show that the perception of the agent's interest cannot be only explained in terms of affects expression and interpretation at the conscious level, and that the use of a ToM in an intelligent virtual agent goes beyond making it just "smarter".

6 Conclusion & perspectives

This paper presents an evaluation protocol for logic-based ToM models in virtual agents for use in interaction tasks. The originality of this protocol is that it is based only on the evaluation of the agent's affects expression. We first presented the conceptual and operational difficulties that arise when trying to evaluate what such models bring to the credibility of virtual agents. We proposed a protocol based on the subjective evaluation of two different versions of an intelligent virtual agent in a dialogue task (job interview simulation). In the first version, the complete logical model is used whereas in the second one, it has been replaced by a simple reactive model. In both conditions, the agent computes and expresses an attitude and task-specific indicators. The subjects evaluated the difficulty and interest of ToM-based virtual agents at a much higher level compared to the reactive model, while they did not perceive significant differences in termes of credibility of affects expression. This tends to show that the perception of the agent's interest cannot be only explained in terms of affects expression and interpretation at the conscious level.

These results must be taken with care since many different variables come in hand in our evaluation protocol. Our first limitation is that we have only a reduced set of subjects: while the statistical results are significant, we intend to extend the number of subjects in the upcoming weeks. Moreover, the participants' feedback gave us interesting insights on their understanding of the virtual agent and this led us to consider the following perspectives. First, the virtual agent's slow change of attitude, while realistic in terms of recruiter's behaviour, seems to be an obstacle for the acceptability of the agent: people tend to expect from virtual agents more forgiving than from humans. Second, the credibility of the virtual agent's affects expression (emotions and attitude) is still a problem for the users, independently from the cognitive model behind. For this reason, we should establish this credibility and use it as a baseline for our future evaluations.

The evaluation of a ToM model in an intelligent virtual agent through its affect expressions remains a difficult task,

essentially because there is no unique correct emotion and attitude that correspond to an input affective state. One of our perspective is to extend this evaluation by comparing the agent's expressed affects with annotated corpus of dialogical interactions (e.g. in the context of a job interview).

References

[1] C. Adam, A. Herzig, and D. Longin. A logical formalization of the OCC theory of emotions. *Synthese*, 168(2):201–248, Feb. 2009.

[2] K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, H. Jones, M. Ochs, C. Pelachaud, K. Porayska-Pomsta, R. Paola, and N. Sabouret. The TARDIS framework: intelligent virtual agents for social coaching in job interviews. *Proceedings of the Tenth International Conference on Advances in Computer Entertainment Technology (ACE-13). Enschede, the Netherlands. LNCS 8253*, page in press, 2013.

[3] E. André and C. Pelachaud. Interacting with embodied conversational agents. In *Speech technology*, pages 123–149. Springer, 2010.

[4] R. Aylett and S. Louchart. If I were you: double appraisal in affective agents. In *Proceedings of the 7th international joint conference on Autonomous Agents and MultiAgent Systems*, pages 1233–1236, 2008.

[5] S. Baron-Cohen. *Mindblindness: An essay on autism and theory of mind*. MIT press, 1997.

[6] M. Belkaid and N. Sabouret. A logical model of theory of mind for virtual agents in the context of job interview simulation. In *Proc. Second International Workshop on Intelligent Digital Games for Empowerment and Inclusion*, 2014.

[7] E. M. A. Blijd-Hoogewys, P. L. C. Van Geert, M. Serra, and R. B. Minderaa. Measuring theory of mind in children. Psychometric properties of the ToM storybooks. *Journal of Autism and Developmental Disorders*, 38(10):1907–1930, 2008.

[8] C. Castelfranchi. Modelling social action for AI agents. *IJCAI'97 Proceedings of the Fifteenth international joint conference on Artifical intelligence - Volume 2*, 103(1):1567–1576, 1997.

[9] M. Courgeon, C. Clavel, and J.-C. Martin. Appraising emotional events during a real-time interactive game. In *Proc. International Workshop on AffectiveAware Virtual Agents and Social Robots AFFINE 09*, pages 1–5. ACM Press, 2009.

[10] M. Courgeon, J.-C. Martin, and C. Jacquemin. Marc: a multimodal affective and reactive character. In *Proceedings of the 1st Workshop on AFFective Interaction in Natural Environments*, 2008.

[11] M. Dastani and E. Lorini. A logic of emotions : from appraisal to coping. In *Proceedings of the 11thiInternational conference on Autonomous Agents and Multiagent Systems*, pages 1133–1140, 2012.

[12] P. Ekman and W. V. Friesen. Facial action coding system. 1977.

[13] M. Harbers. Explaining agent behavior in virtual training. *SIKS dissertation series*, 2011(35), 2011.

[14] M. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. Picard. MACH: My Automated Conversation coacH. In *Proc. 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM Press, 2013.

[15] S. Marsella and J. Gratch. EMA: A computational model of appraisal dynamics. *European Meeting on Cybernetics and Systems Research*, 2:1–5, 2006.

[16] A. Ortony. On making believable emotional agents believable. *Trappl et al.(Eds.)(2002)*, 2002.

[17] A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobreperez, S. Woods, C. Zoll, and L. Hall. Caring for agents and agents that care: Building empathic relations with synthetic agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 194–201, Washington, DC, USA, 2004. IEEE Computer Society.

[18] C. Peters. Foundations of an Agent Theory of Mind Model for Conversation Initiation in Virtual Environments. *Virtual Social Agents*, page 163, 2005.

[19] D. V. Pynadath, N. Wang, and S. C. Marsella. Are you thinking what I'm thinking? An Evaluation of a Simplified Theory of Mind. In *Intelligent Virtual Agents*, pages 44–57. Springer, 2013.

[20] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. F. Valstar, and M. Wöllmer. Building autonomous sensitive artificial listeners. *T. Affective Computing*, 3(2):165–183, 2012.

[21] D. R. Traum and J. F. Allen. Discourse obligations in dialogue processing. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics, 1994.

[22] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André. The social signal interpretation (ssi) framework-multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain*, 2013.

Perception, langage, curiosité : éléments clés pour la conceptualisation de connaissances en robotique interactive

Christophe Sabourin[1] Kurosh Madani[1]

[1] Laboratoire Images, Signaux et Systèmes Intelligents (LISSI) - EA 3956
Université Paris-Est Créteil (UPEC)

christophe.sabourin@u-pec.fr

Résumé

La réalisation d'un robot compagnon pouvant s'adapter "naturellement" au contexte de son environnement nécessite qu'il soit capable d'interpréter la diversité et la complexité des informations perçues. L'objectif de cet article est de montrer comment à partir d'un ensemble d'observations un robot peut progressivement conceptualiser des connaissances. Ces observations sont le résulat d'une interaction homme-robot lui permettant d'associer des informations perçues avec des éléments du langage naturel qui sont donnés par un "tuteur". Nous montrons aussi le rôle majeur joué par la "curiosité" dans ce processus d'acquisition.

Mots Clef

Perception, langage, curiosité, interaction homme-robot.

Abstract

The design of a companion robot able to adapt naturally to a real environment implies that this robot should be able to interpret the complexity and the diversity of the perceived informations. The goal of this paper is to show how, with a set of observations, a robot can progressively conceptualize knowledges. These observations are the result of an human-robot interaction allowing him to associate perceived information from environment and utterances given by a "tutor". In addition, we show that the curiosity plays a key role in this acquisition process.

Keywords

Perception, langage, curiosity, human-robot interaction.

1 Introduction

La conception de robots compagnons capables d'assister à domicile des personnes en perte d'autonomie est un enjeu important mais force est de constater que les performances des prototypes actuels sont encore loin d'être satisfaisantes. Au-delà des barrières psychologiques qu'il sera nécessaire de faire tomber pour que ces robots soient acceptés et utilisés par les êtres humains, il reste en effet encore à lever des verrous technologiques et scientifiques majeurs. Parmi les nombreuses difficultés à surmonter pour réaliser un robot capable de s'adapter "naturellement" au contexte de son environnement, l'une d'entre elles concerne ses capacités à acquérir des nouvelles connaissances et créer de nouveaux concepts lui permettant d'interpréter la diversité et la complexité de son environnement. Par conséquent, cette capacité d'adaptation doit donc aller bien au-delà du simple fait d'apprendre à reconnaître, identifier et nommer des objets comme cela est généralement le cas en robotique (voir par exemple [1]).

Un enfant, dès l'âge de deux ans, possède déjà une habilité fondamentale qui consiste à pouvoir regrouper ou classer les propriétés, les objets voire les événements [2]. Ce processus de catégorisation permet ainsi à un être humain de réduire la diversité et la complexité du monde réel qui l'entoure. Par exemple, bien que l'œil humain soit théoriquement capable de discriminer plusieurs millions de couleurs, l'homme utilise généralement moins d'une quinzaine de mots pour les classer. Cette classification dépend d'ailleurs du langage et de la culture [3] [4]. En effet, comme cela a déjà été montré dans [6] [5], des éléments de langages simples facilitent ce processus de catégorisation des objets, mais aussi des concepts comme les formes ou les couleurs. On peut donc supposer que chez l'homme, cette association entre perception et langage permet d'enclencher des processus de catégorisation. Et en donnant du sens à ce que l'on perçoit, ces éléments de langages permettent aussi à l'être humain de communiquer et de partager sa connaissance. Par ailleurs, certains scientifiques émettent aussi l'idée que le langage naturel a émergé du grand nombre de catégories et de concepts que l'homme a formé [7], et que par conséquent, le langage est avant tout un outil cognitif permettant aux êtres humains d'acquérir des connaissances. Le langage joue donc un rôle fondamental dans l'interprétation de la perception et l'acquisition de connaissances. Partant de ce constat, l'acquisition autonome de connaissances en intelligence artificielle, et plus particulièrement en robotique, doit se fonder sur des processus d'apprentissage intégrant la perception et le langage via une interaction homme-robot, comme pourrait le faire un jeune enfant en interagissant avec ses parents.

Un autre aspect très important est de comprendre et modéliser les processus liés à la curiosité qui sont chez l'être humain une source de motivation dans l'acquisition de nouvelles connaissances [8]. Dans "Theory of human curiosity" [9], Berlyne propose de diviser cette curiosité en deux catégories distinctes qui sont la curiosité perceptuelle et la curiosité épistémique. La curiosité perceptuelle permet de sélectionner les informations importantes qui sont perçues par les voies sensorielles. La curiosité épistémique, quant à elle, correspond au désir d'apprendre de nouvelles connaissances, de résoudre des problèmes [10]. Cette curiosité épistémique est aussi liée à la mémorisation à long terme [11]. C'est sur ces bases que dans [12], nous avons proposé l'implémentation d'un système de curiosité artificielle décomposé en deux parties qui sont la curiosité perceptuelle et la curiosité épistémique. En stimulant la collecte d'informations, le système de curiosité artificielle contribue aussi à l'amélioration des interactions entre l'homme et le robot. L'interaction homme-robot étant bidirectionnelle, le robot peut décider d'interagir avec l'homme afin d'acquérir de nouvelles connaissances.

L'objectif de cet article est donc de montrer comment, à partir d'un ensemble d'observations, un robot peut progressivement conceptualiser des connaissances. Chaque observation est le résulat d'une interaction homme-robot permettant au robot d'associer les informations de l'environnement qu'il perçoit avec des éléments du langage naturel donnés par un "tuteur". Et l'implémentation d'un système de curiosité artificielle permet de stimuler l'acquisition de nouvelles observations. Dans la section 2, nous commencerons par faire une présentation synthétique des fondements de notre système de curiosité artificielle. Dans les sections 3 et 4, nous montrerons comment le robot peut progressivement, à partir d'un ensemble d'observations (section 3), conceptualiser des connaissances (section 4). La section 5 nous permettra de décrire les moyens expérimentaux mis en œuvre ainsi que présenter quelques résultats. La conclusion et les perspectives de ce travail feront l'objet de la dernière section.

2 Curiosité artificielle

La curiosité artificielle n'est bien sûr pas un concept nouveau et a déjà fait l'objet de travaux en robotique (voir par exemple [13]). Cependant, ces travaux se sont généralement focalisés sur l'exploration sensori-motrice du robot, le but étant que le robot découvre par lui-même sa capacité à interagir physiquement avec son environnement. L'originalité de notre approche, quant à elle, est que nous avons proposé de décomposer la curiosité en deux niveaux cognitifs [12]. Le premier niveau, qui correspond à la curiosité perceptuelle, regroupe un ensemble de Fonctions Cognitives Inconscientes (FCI). Ces FCI permettent d'extraire, à partir des données sensorielles (images, sons, etc..), les informations qui paraissent essentielles (pertinentes). Un exemple de curiosité perceptuelle est sans aucun doute l'attention visuelle (visual saliency) qui permet d'extraire d'un flux visuel les informations qui nous semblent pertinentes. Le deuxième niveau, qui correspond à la curiosité épistémique, regroupe un ensemble de Fonctions Cognitives Conscientes (FCC). Les FCC ont pour but de stimuler l'acquisition de nouvelles observations pour résoudre un problème donné. La figure 1 illustre cette approche.

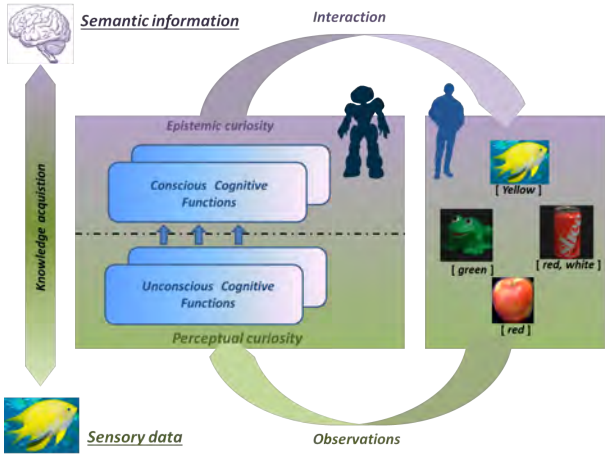


FIGURE 1 – Illustration de l'influence de la curiosité sur la perception et l'acquisition de connaissances.

La curiosité perceptuelle permet d'acquérir de manière inconsciente des informations sensorielles. La curiosité épistémique a alors pour rôle de stimuler des interactions homme-robot afin de transformer les informations sensorielles perçues en informations sémantiques sur la base d'un ensemble d'observations. Par conséquent, ce système de curiosité artificielle permet de conceptualiser progressivement des connaissances "sémantiques" sur la base d'acquisitions sensorielles (les observations).

3 Observations

Comme nous l'avons écrit dans l'introduction, une des caractéristiques fondamentales de l'homme est sa capacité à regrouper, classifier les objets ainsi que leurs propriétés (formes, couleurs, sons, etc..) et ainsi créer des concepts permettant de "simplifier" la diversité et la complexité du monde dans lequel il évolue. Cette catégorisation, ou plus généralement cette conceptualisation, est le résultat d'une interaction homme-environnement et d'un processus cognitif permettant d'interpréter un ensemble d'observations $O = \{o_1, o_2, ..., o_k\}$. Où chaque observation $o_k = \{I, U\}$ est définie comme une association d'un ensemble d'informations sensorielles (features) $I = \{i_1, i_2, ..., i_j\}$ et d'un ensemble de termes linguistiques (utterances) $U = \{u_1, u_2, .., u_i\}$. Dans la suite de cet article, nous nous appuierons sur l'exemple de la catégorisation des couleurs, mais il est important de garder à l'esprit que l'approche que nous proposons peut s'inscrire dans un cadre plus général. Cependant, une catégorisation, comme c'est le cas pour la reconnaissance des couleurs, ne peut se limiter à un simple problème d'apprentissage supervisé. Prenons par exemple

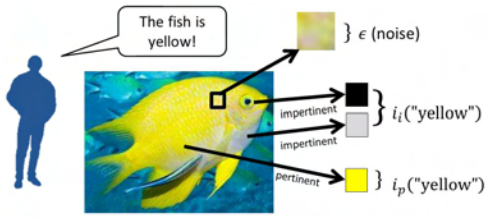


FIGURE 2 – Un homme à qui on demanderait de décrire le poisson répondrait probablement qu'il est jaune bien qu'il ne soit pas complètement jaune. Les symboles i_i, i_p représentent les informations des couleurs extraites des différents segments. Ces informations sont décompsés en informations pertinentes (i_p) et informations non pertinentes (i_i).

le cas de la figure 2. Si nous demandions à une personne de décrire le poisson, une des réponses serait très probablement qu'il est jaune. Bien que cela nous semble être une évidence, il n'en est pas de même si nous faisons une analyse plus fine de cette image. Car d'un point de vue du signal, le codage de l'image du poisson fait apparaître une représentation numérique beaucoup plus complexe (nombre de pixels, nombre de couleurs). Même une segmentation "grossière" nous amènerait à identifier des segments de couleurs légèrement différentes (nuances de jaunes), assimilables à des informations pertinentes, et des segments de couleurs très différentes (couleur noir de l'œil), qui correspondraient cette fois à des informations non pertinentes.
Dans ce contexte, la problématique est : comment différencier les informations pertinentes (la ou les nuances de jaunes) des informations non pertinentes afin de les associer à la catégorie qui est représentée par un mot prononcé par le tuteur ? L'objectif de la section suivante est donc de montrer comment, à partir d'un ensemble d'observations, un robot peut progressivement conceptualiser des connaissances. Ces observations étant le résulat d'une interaction homme-robot lui permettant d'associer les informations de l'environnement qu'il perçoit avec des éléments du langage naturel prononcés par un "tuteur".

4 Interprétations des observations

La figure 3 schématise le processus d'acquisition et d'interprétation des informations perçues lors de 3 observations ($o_1 = \{i_1, i_2, i_3, i_4, green\}, o_2 = \{i_5, i_6, white, red\}$, $o_3 = \{i_7, i_8, i_9, green, white\}$). Chaque observation correspond à un ensemble de données hétérogènes regroupant des informations extraites ($i_1, i_2, i_3, i_4, etc..$) des signaux sensoriels (vision dans notre cas) et des mots prononcés par un tuteur ("green", "white", "red"). Pour chaque observation, les informations sensorielles $I = \{i_1, i_2, ..., i_j\}$ sont extraites via une segmentation de l'image et codées sous forme numérique (par exemple codage RGB).
Le problème consiste maintenant à interpréter ces observations en associant les couleurs des différents segments identifiés à leur classe respective (figure 4). En définissant :

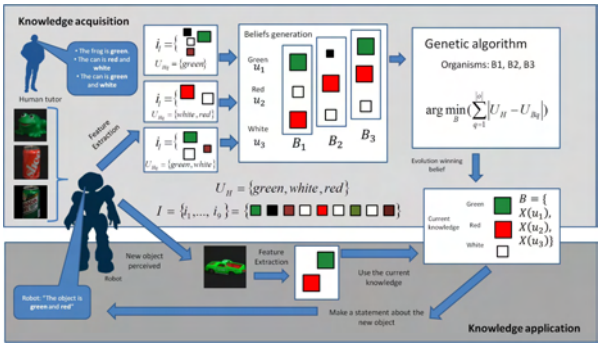


FIGURE 3 – Le robot effectue une série de plusieurs observations : $o_1 = \{i_1, i_2, i_3, i_4, green\}, o_2 = \{i_5, i_6, white, red\}$ et $o_3 = \{i_7, i_8, i_9, green, white\}$. Après une phase d'apprentissage, le robot est capable de déduire que la voiture est rouge et verte.

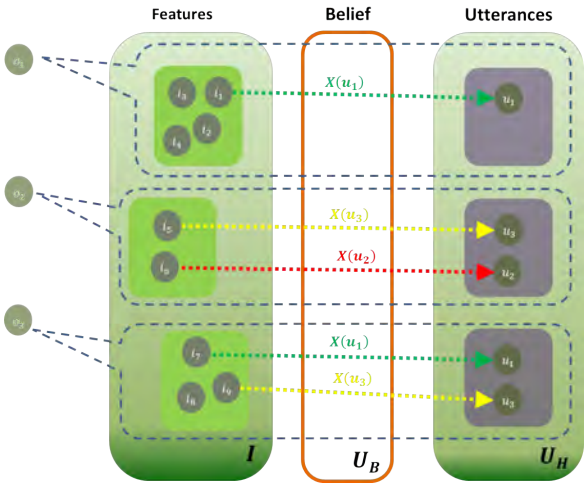


FIGURE 4 – Interprétation des observations en associant les couleurs des différents segments : $X(u_1) = \{u_1, i_1, i_7\}$, $X(u_2) = \{u_2, i_6\}$, $X(u_3) = \{u_3, i_5, i_9\}$.

- I (équation 1) comme l'ensemble des informations pertinentes i_p, des informations non pertinentes i_i ainsi que le bruit ϵ,
- $X(u) = \{u, I_j \subseteq I\}$ une interprétation de u permettant d'associer à cette catégorie u les perceptions adéquates (par exemple $(X(green) = \{green, i_1, i_7\})$) ,
- et $U_B = \{X(green), X(white), X(red)\}$ la croyance qui permet de déterminer l'ensemble des interprétations $X(u)$.

$$I = \bigcup i_p(u) + \bigcup i_i(u) + \epsilon \tag{1}$$

Il est alors possible de trouver la croyance U_B qui se rapproche le plus possible de la réalité via le processus d'optimisation suivant :

$$\arg\min_{B} \left(\sum_{q=1}^{|O|} |U_{Hq} - U_{Bq}| \right) \tag{2}$$

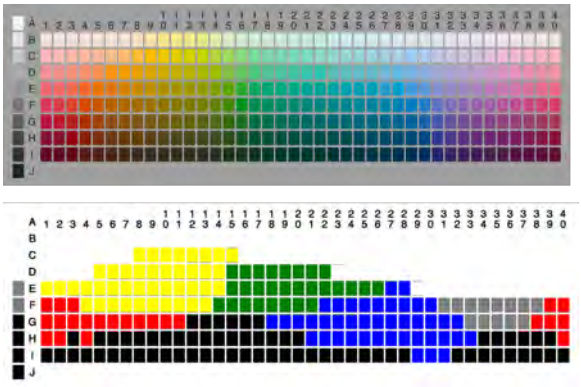


FIGURE 5 – Représentation de la table des couleurs d'après "world color survey database" (haut de la figure), et représentation de la classification des couleurs après apprentissage (bas de la figure).

FIGURE 6 – Comparaison des couleurs réelles des objets et celles perçues par le système après apprentissage.

Où pour chaque observation $o_q \in O$, U_{Hq} représente les mots prononcés par le tuteur et U_{Bq} correspond aux interprétations générées par le robot. Une solution à ce problème d'optimisation (2) repose sur l'utilisation d'algorithme génétique où chaque organisme correspond à une croyance (U_{Bq}), cette croyance représentant un ensemble d'interprétations $X(u)$. Le problème se résume alors, via un processus évolutif, à trouver la croyance U_B qui se rapproche le plus possible de la réalité U_H. Pour plus de détail concernant la solution à ce prolème d'optimisation, nous invitons le lecteur à consulter les références [15] [16].

Avant d'utiliser la méthode que nous venons de décrire dans un environnement réel, nous avons réalisé une première évaluation dans un environnent virtuel. Pour cela, nous avons utilisé la base de données de "Columbia Object Image Library Database" [17] qui contient des images d'une centaine d'objets. Chacun de ces objets ont été décrit à l'aide de une ou deux couleurs. Le choix des couleurs étant limité à : Noir, gris, blanc, rouge, vert, bleu et jaune. Les figures 5 et 6 montrent les résultats que nous avons obtenus. La figure 5 représente une comparaison entre les couleurs comme définies dans "world color survey database" [17] et la classification des couleurs après apprentissage. La figure 6, quant à elle, permet de montrer les couleurs perçues de quelques objets de cette base de données.

5 Validation expérimentale sur la plateforme robotique Nao

Après avoir vérifié notre approche en simulation, nous avons réalisé différentes expérimentations à l'aide de la plateforme robotique Nao. Nous avons notamment imaginé plusieurs scénarios afin de valider les capacités du robot à acquérir, mais aussi restituer, des concepts nouveaux. Dans la suite de cette section, nous présenterons plus spécifiquement des résultats concernant la catégorisation des couleurs mais nous invitons le lecteur à consulter les vidéos en ligne suivantes [18] [19] ainsi que les références [15] et [16] qui donnent un bon aperçu de l'ensemble des résultats que nous avons obtenus. Mais avant de présenter ces résultats, nous commencerons par décrire succinctement l'architecture logicielle développée pour cette plate-forme robotique expérimentale.

5.1 Architecture logicielle

Le robot Nao [20] est un robot humanoïde programmable de 58 cm équipé notamment de plusieurs caméras, microphones et haut-parleurs. Cet équipement lui permet donc de percevoir et communiquer avec son environnement. L'ensemble des programmes ont été codés en c# et exécutés sur un ordinateur distant, l'objectif ici n'étant pas d'embarquer les programmes sur Nao mais de contrôler le robot à distance afin de valider nos résultats.

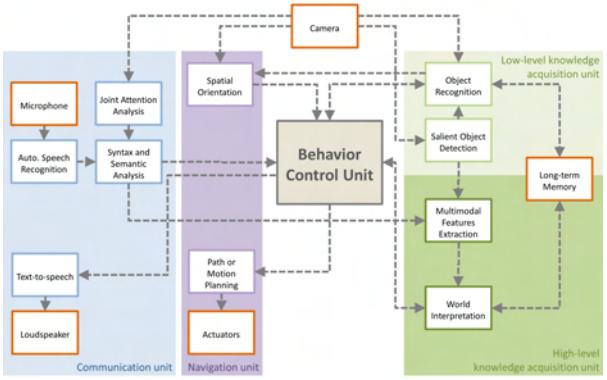


FIGURE 7 – Description de l'architecture logicielle de la plate-forme robotique expérimentale.

La figure 7 schématise l'architecture logicielle qui peut être décomposée en 5 unités :

- L'unité de communication pour gérer l'interaction homme-robot.
- L'unité de navigation qui permet au robot de se positionner et de se diriger dans son environnement.
- L'unité de perception (The Low-level Knowledge Acquisition Unit) assurant la collecte des informations. Cette unité utilise notamment des algorithmes d'attention visuelle et de reconnaissance d'objet.
- L'unité d'acquisition et de conceptualisation de connaissance (High-level Knowledge Acquisition Unit).

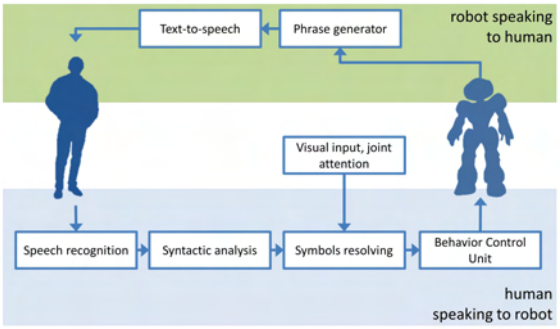


FIGURE 8 – Schéma descriptif des interactions entre le tuteur et le robot.

– L'unité de contrôle qui permet de coordonner l'ensemble des autres unités.

La figure 8 schématise plus spécifiquement le principe d'interacion homme-robot. La détection des objets est notamment améliorée via la détection de la main du tuteur. Concernant l'interaction vocale homme-robot, nous avons utilisé l'outil de reconnaissance vocale du robot ainsi que le logiciel "TreeTagger" [21] afin de déterminer les informations données par le tuteur. Pour l'interaction vocale robot-homme, nous avons utilisé l'outil "text-to-speech" du robot Nao.

5.2 Résultats

Dans le cadre de ces expérimentations, nous avons commencé par rassembler 25 objets de la vie courante (voir Fig. 9). L'ensemble des objets, en les nommant, ont été présentés un à un au robot par le tuteur. Ce prérequis est indispensable pour que le robot puisse par la suite reconnaître et nommer les objets. Cet ensemble d'objets a ensuite été divisé, de manière aléatoire, en deux groupes, le premier étant destiné à la phase d'apprentissage, le deuxième à la phase de validation. Le tuteur présente ensuite au robot, un à un, les objets du premier groupe en les décrivant, par exemple, "le livre est noir". Le robot, après avoir identifié l'objet, peut alors extraire, à l'aide d'algorithmes de segmentation standard, les couleurs de l'objet désigné.

Suite à cette phase d'apprentissage qui permet au robot de catégoriser les couleurs grâce à un ensemble d'observations, il est alors nécessaire de valider cette acquisition en utilisant les objets du deuxième groupe. En utilisant un processus d'interaction similaire, le tuteur demande alors de nommer et de décrire un objet. Les figures 10 et 11 illustrent quelques résultats des expériences qui viennent d'être décrites. Sur la figure 10, le tuteur attire l'attention du robot sur un objet et lui demande de le décrire. Le robot répond alors au tuteur que la boîte est jaune. La figure 11 illustre quant à elle la perception des couleurs du robot pour deux objets : une pomme et une boîte de lait.

La dernière expérience permet de montrer comment le robot est capable d'utiliser cette acquisition des connaissances, la catégorisation des couleurs dans notre cas, pour différencier des objets semblables. Dans cette expérience,

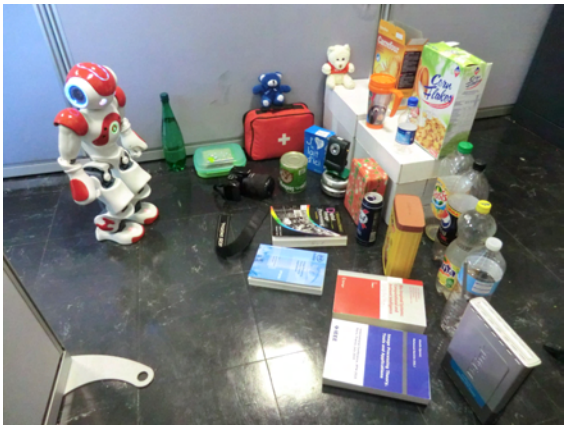


FIGURE 9 – Le robot Nao devant les objets utilisés lors des validations expérimentales.

FIGURE 10 – Exemple d'expérience ou le tuteur demande au robot de lui décrire l'objet pointé du doigt. Le robot répond alors au tuteur que cet objet est jaune.

le tuteur commence par demander au robot de se diriger vers un livre. Après observation, le robot découvre trois objets qu'il assimile à des livres, il demande alors au tuteur plus de précisions. Le tuteur lui indique de se diriger vers le livre rouge. Après analyse des différentes couleurs des livres, le robot se dirige vers le livre rouge.

6 Conclusion et perspectives

Dans cet article, nous avons décrit une technique qui permet à un robot de progressivement créer des concepts. Cette méthode repose, d'une part, sur l'acquisition d'un ensemble d'observations, et d'autre part, sur un algorithme évolutionniste permettant d'interpéter les observations. Une observation est définie comme une association d'un ensemble d'informations sensorielles et d'un ensemble de termes linguistiques. L'objectif de l'algorithme

FIGURE 11 – Illustration de la perception des couleurs de deux objets par le robot.

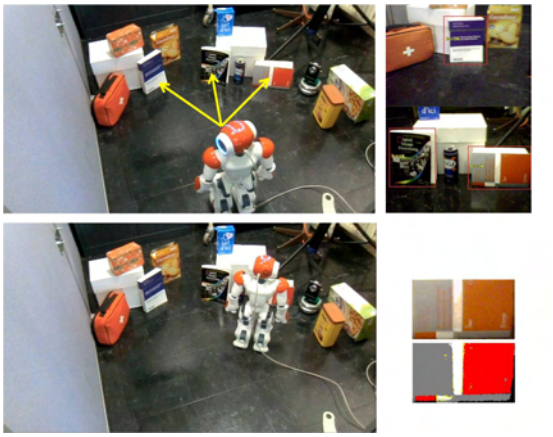


FIGURE 12 – Exemple de résultat expérimental où le robot utilise les connaissances acquises pour différencier des objets similaires.

évolutionniste est alors de trouver une croyance qui se rapproche le plus possible de la réalité via un processus d'optimisation. Cette approche a été validé expérimentalement pour le cas de la catégorisation de couleurs, premièrement dans un environnement purement virtuel, et deuxièmement, dans un environnement réel à l'aide du robot Nao. Dans ce contexte, nous avons aussi montré l'intérêt du rôle joué par la curiosité perceptuelle et la curiosité épistémique.

Ces premiers résultats très prometteurs permettent d'envisager de nombreuses perspectives à moyen terme. La première consiste à étendre cette approche à des concepts autres que la couleur, comme les formes, les représentations spatiales ; en proposant des algorithmes de co-évolution permettant d'apprendre simultanément les formes et les couleurs. La deuxième consiste aussi à généraliser cette approche à d'autres sens (ouïe, odorat, toucher, etc..). Sur le long terme, l'objectif est bien sûr de concevoir un robot doté d'une très grande autonomie en ce qui concerne l'acquisition et la conceptualisation de connaissances pouvant répondre aux exigences d'un robot compagnon.

Références

[1] S. Coradeschi, A. Saffiotti. An introduction to the anchoring problem. Robotics and Autonomous Systems, vol. 43(2-3), pages 85-96, 2003.

[2] D. Poulin-Dubois. Le déloppement cognitif de lenfant, chapitre : Le déloppement cognitif de 0 à 2 ans : les fondements du déloppement ulteur, pages 9 - 34. De Boeck Université 2007.

[3] P. Kay, T. Regier. Color naming universals : The case of Berinmo. Cognition, vol. 102(2), pages : 289-298, 2007

[4] A. Majid, M. Bowerman, S. Kita, D. B.M. Haun, S.C. Levinson. Can language restructure cognition ? The case for space. Trends in Cognitive Sciences, vol. 8(3), pages :108-114, 2004.

[5] K. Plunkett, J. Hu, L.B. Cohen. Labels can override perceptual categories in early infancy. Cognition, vol. 106(2), pages :665-681, 2008.

[6] F. Xu. The role of language in acquiring object kind concepts in infancy. Cognition, vol. 85(3), pages :223-250, 2002.

[7] Reboul, A. Cognition et langage, in Dortier, J-F. (ed.) Le cerveau et la pensée, Editions Sciences Humaines, 2011.

[8] G. Loewenstein. The psychology of curiosity : A review and reinterpretation. Psychological bulletin, vol. 116(1), pages :75-98, 1994.

[9] D.E. Berlyne. A theory of human curiosity. British Journal of Psychology, vol 45, pages :180-191, 1954.

[10] J.A. Litman. Interest and deprivation factors of epistemic curiosity. Personality and Individual Differences, vol. 44(7), pages :1585-1595, 2008.

[11] M.J. Kang, M. Hsu, I.M. Krajbich, G. Loewenstein, S.M. McClure, J.T.T. Wang, C.F. Camerer. The wick in the candle of learning : epistemic curiosity activates reward circuitry and enhances memory. Psychological Science, vol. 20(8), pages : 963-973, 2009.

[12] D.M. Ramik, K. Madani, C. Sabourin. From visual patterns to semantic description : A cognitive approach using artificial curiosity as the foundation. Pattern Recognition Letters, vol 34(14), pages : 1577-1588, 2013

[13] P.Y. Oudeyer, F. Kaplan, V. Hafner. Intrinsic Motivation Systems for Autonomous Mental Development IEEE Transactions on Evolutionary Computation, vol. 11(2), pages : 265-286, 2007.

[14] J. Gottlieb, , P.Y. Oudeyer, M. Lopes, A. Baranes. Information Seeking, Curiosity and Attention : Computational and Neural Mechanisms Trends in Cognitive Science vol. 17(11), pages : 585-593, 2013.

[15] D.M. Ramik, C. Sabourin, K. Madani. Autonomous Knowledge Acquisition based on Artificial Curiosity : Application to Mobile Robots in Indoor Environment. Robotics and Autonomous Systems, Vol 61(12), pages : 1680-1695, 2013.

[16] D.M. Ramik, C. Sabourin, R. Moreno, K. Madani. A machine learning based intelligent vision system for autonomous object detection and recognition Applied Intelligence, vol 40(2), pages : 358-375, 2014.

[17] `http://www.icsi.berkeley.edu/wcs/data.html`

[18] `http://youtu.be/Y_JM0KfJb8Q`

[19] `http://youtu.be/W5FD6zXihOo`

[20] `http://www.aldebaran.com/fr`

[21] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

Adaptation in an Interactive Model designed for Human Conversation and Music Improvisation: a preparatory outline

Kevin Sanlaville[12] Frédéric Bevilacqua[1] Catherine Pélachaud[2] Gérard Assayag[1]

[1]UMR STMS (Ircam, CNRS, UPMC)
1, place Stravinsky
75004 Paris
[2]Telecom-Paristech CNRS-LTCI
37/39 rue Darreau
75014 Paris

kevin.sanlaville@ircam.fr

Résumé

A travers une brève description de notre champ d'étude et de questions ouvertes nous essaierons d'esquisser l'évolution souhaitée de notre these de troisième cycle.

Mots Clef

Adaptation, multimodalité, non-verbal, musique_informatique,agent_conversationnel_animé

Abstract

Through a brief description of our research field and of open questions we will try to draw an outline of our PhD project and of its future evolution

Keywords

Adaptation, multimodality, non-verbal. computer_music, embodied_conversational_agents

1 Introduction

The SeNSE project of the LabEx SMART focuses on emotional social signals, from their detection to their exploitation, including also their interpretation and their modeling[1].

In this PhD project, we aim at developing a system that covers both both Embodied Conversational Agents (ECA) and computer agents able to perform spontaneous musical improvisation (such as the OMaX agent[3], see in part 3.1). Such a system must obviously be sufficiently generic to describe these rather different kinds of agents. We hypothesize that such a challenge might lead us to formalize an interaction model that should avoid shortcomings of current systems.

In particular, our goal in this context is to design an interaction model using group dynamics, covering different possible cases: several people interacting with one agent, one person interacting with several agents, or several people interacting with several agents.

[1] http://www.smart-labex.fr/index.php?perma=SeNSE

These different configurations should be able to handle the various temporal dimensions of interaction (e.g. a conversation turn, an entire dialogue, repeated interaction, etc.). This model should also be able to take into account the different modalities of interaction, especially the non-verbal communication and how it is linked to verbal communication.

In the next section, we will describe how we aim to perform interaction in our system and what should be useful to attain it.

In Section 3, we will discuss the notion of adaptation and how we intend to model it in our system. In Section 4 we will present the directions we aim to follow and the concepts we acknowledge in our PhD project. In section 5 we will recall the notion of non-verbal communication. In section 6, we will detail the previous researches that we intend to build on.

In the last section, we will conclude by giving a brief summary of the similarity of the models presented in section 6 and how that will influence our approach.

2 Interaction

In the setting of our PhD Project, we intend to model an interactive and communicative task with multiple interactants. These interactants could be human or virtual agents, through the exchange of multi-modal communicative signals.

The agents, virtual and human alike, may be either involved in a conversational or musical improvisation task. We hypothesize that this task is performed cooperatively, meaning that agents must cooperate towards its fulfillment, even if we consider that on a local scale agents may try to compete to occupy or retain a particular role (more on that later).

The interactions between the different agents should be defined as the way two or more interactants influence each other, hence our proposition to model interaction as a shared construct between the interactants. If we aim to model Human-Machine Interaction, for us these interactions should simulate Human-Human interactions to increase the engagement of the human interactants, and therefore the efficiency of the system[14]. In our

PhD project we are interested in two aspects of interaction: synchronization and emergence.

2.1 Synchronization

The aim of synchronization is to induce mutual reactions between interactants[19] and to negotiate the role of each participant in a continuous communicative interaction. To assure the success of synchronization, several mechanisms are necessary that are located at different levels of abstraction.

We are considering three levels of synchronization, the lowest being the use of backchannel. These backchannels are multi-modal signals emitted during an interaction by listeners to indicate their reaction to what is being performed (i.e. a turn in conversation or a musical solo)[2]. This information is very important to assess information concerning communicative interaction: attention, perception, comprehension and internal reactions [15].

Another level of synchronization is the use of Turn-Taking analysis. The way turn of a communicative interaction is taken is relevant to understand interaction in two aspects: first, it relies on low-level cues (e.g. in conversation throat clearing, mouth opening, etc.) to regulate and give indications about the current state of the interaction (e.g. in conversation again, overlapping turns may indicate conflicts[18]).

Another important aspect of the synchronization that occurs at a higher level is the existence of a Theory of Mind. Defined by Simon Baron-Cohen as "the ability to attribute the full range of mental space (both goal states and epistemic states) to ourselves and to others, and to use such to make sense and predict behavior "[5]. The use of a Theory of Mind is therefore a way to attain consciousness not only of other interlocutors but also of ourselves and hence is essential to reach synchrony with the other interactants of a communicative interaction.

In the next section, we will try to present another essential aspect of our PhD thesis and a way to enforce the believability of the interaction: adaptation.

2.2 Adaptation

Although there is no generally accepted definition of adaptation, we could assert that adaptation is a process or a set of processes devised to augment the fitness of an object to a purpose.

In our particular case of an interactive system oriented towards a communicative task, we reckon that adaptation could be divided in three interconnected levels.

The first level of adaptation concerns low-level interaction and reflex responses to external stimuli. It concerns processes like imitation. It also copes with the low-level responses to these stimuli.

The second level of adaptation is devoted to the building of meaning. Through the interpretation of multi-modal signals, this level is able not only to construct meaning but also to use the constructed meaning to provide meaningful response to higher-level stimuli.

The last level of adaptation deals with cognition and aspects such as planning, self-assessing and world representation.

2.3 Emergence

In the sub-section 2.1 , we have defined a few elements that are part of synchronization. Synchronization is an essential element contributing to the performance of a believable interaction. However, we still need to define the method to reach such interaction. In our opinion, this believability of the communicative interaction can be reached through emergence.

There is no real consensus about the definition of emergence, but according to Teo, Meng and Szabo, "a system is said complex if it exhibits designed properties that can be derived from the system specifications as well as properties that are irreducible from knowledge of the interconnected components. These irreducible properties are defined as emergent properties or emergence" [17].

So we could say that emergence is a phenomenon that allows a system to use the interaction of its component to develop properties that were not designed in them. For us, this will occur at three distinct though interconnected levels.

The first level is the emergence of behavior. From the low-level interactions of a group of interactants, each participant will adapt its behavior not only considering its own set of basic reactive rules but also those of their co-interactants.

The second level is the emergence of organization. In an organization (i.e. a music band or a group of people), each participant occupies a role, like the conductor of an orchestra or the chairman of an assembly. From the behaviors of the interactants, each participant will adopt a role conditioned by the dynamics of the present interaction.

The last level of emergence is the emergence of control. This last level will allow interactants to become aware of their own organization and to be able to influence it through their behavior to best fit to their task. It represents the reciprocal influence between the system and its environment [5]. For instance, the members of the interacting group could isolate a perturbing element to maintain the accomplishment of the common task.

3 Our objective: computational model

In this section, we will define the computational model that we intend to elaborate in our PhD Thesis. We will then present our comprehension of the links between these fields. We will at last complete this description with two use cases clarifying the desired application of the model to our defined settings.

3.1 General Definition

Our objectives are to devise a computational model of interaction that encompasses the different adaptation levels of interaction through the various levels of emergence.

In a complementary approach to the use of group dynamics, we propose to use the multi-modality of interaction to enrich our interaction. This approach will notably use the non-verbal modalities of communication to complete interaction with paralinguistic signals and expressive non-verbal communication. We will define non-verbal communication in a following subsection.

Our agents, both human and artificial, will dispose of an ensemble of similar features, including but not limited to cultural background, a representation of the task, a representation of the role of the agent in the defined task, a form of experience, etc. If we do not intend to alter some of this knowledge like its cultural background, we think that in order to enhance its efficiency the system should be able to augment and amend its experience, and from what it had learned try to adapt its role, or event its task, to better fit to its designed purpose.

3.2 Music and Conversation

As we stated above, our model was planned to model two different although similar situations: musical improvisation and human conversation.

We consider that music follows implicitly or explicitly a system of syntactic rules akin to spoken language [7]. In this aspect and in many others, we state that music is in itself comparable to linguistic communication. Generally, it is based on a set of sounds, abstracted as notes, that are combined to build sequences, themselves organized in higher level structures such as phrases.

Just as in natural language conversation non-verbal expression adds to expressivity, in music some form of "non-instrumental" expression bears the same role When for instance the soloist of a small group of musician turns towards another musician to give him the floor, he is reproducing mechanisms similar to those who regulate turn taking in natural language conversation. The expressive deviations of interpretation (tempo, rubato, accent etc.) also convey meaningful information just as prosody does in language.

3.3 Use Cases

The first use case that we will present is the case of the Musical Improvisation. In this setting, a musician or a group of musician will play an improvised melody following the principle of free improvisation (i.e. with no a-priori explicit and common rules)
The agents of the system will listen to the instrumentalist, analyze it, and play a melody following the musical characteristics of the played improvisation (i.e. rhythm, tempo, even style and genre).

The second use case that we will present is the case of the Human-like conversation. In this setting, a human or a group of humans will converse with an agent or a group of agents.

We have not already defined whether this conversation should follow a specific purpose (e.g. the completion of a task) or if it should have the only purpose of appearing believable.

4 Non-Verbal Communication

In the previous sub-section, we have evoked our objective of building a believable interaction, and our intention of using non-verbal communication to allow the fulfillment of this objective. In this section, we will present a brief history of non-verbal communication and of its main research topics.

Non-verbal communication researches focus on various modalities and functions [11], such as turn taking[18], paraverbal signal emission and the way they fit in their context[2], or concentrated on specific signals like head nod in conversation[10].

From these researches emerged certain constants. First, these non-verbal signals obey a dynamic evolution[18] and like verbal communication they hugely depend on their emission context and sometimes of past interactions to be deciphered, sometimes even perceived. An example of that phenomenon might be irony, in which case information carried by prosody contradict the actual meaning of the verbal communication[8]. From this approach emerged the idea of non-verbal communication as communication signals in themselves, like facial expressions or offensive gestures can be by themselves a support for information and as such must be considered as communication acts[1].

More, if non-verbal communications varies between individuals[1] [18], it also shows many forms of cultural variations ; the head nod, for instance[10], however present in almost every culture, does not transmit the same informations and is not used with the same frequency around the world (e.g. in the hellenistic world and in the Balkans, nodding is considered as a sign of refusal).

At last, and considering again the head nod, regarding its emission context it could mean an acceptance, a greeting, an emphasis (on what is being said) or a

deictic gesture. Alike verbal communication, non-verbal communication is heavily dependent on the context in which it happens[1].

5 Context of our PhD Project

Our PhD project relies on existing researches. We will first present OMAx, an improvising agent developed at Ircam, and GRETA-VIB, an ECA developed in LTCI-Telecom-Paristech.

5.1 OMAx

OMAx[3] is a continued program of improvising musical agent conceived in the Représentation Musicale (RepMus) team at Ircam and developed in Max/MSP[2].

OMaX agents are able to listen to a performer, learn on the fly and play musical phrases inspired from his/her performance. This is performed thanks to Factor Oracles [1], formal structures close to factor automata and suffix trees, built incrementally during the listening process, that have been designed in the formal languages community. These oracles capture statistical sequence models of the input streams and give a mapping of phrases and patterns for different musical dimensions (pitch, rhythm, timbre, harmonic textures etc.)

The "motivic" structure of the musical piece is therefore parsed and reconstructed in quasi-real-time (with a few dozen of milliseconds delay). This memory model can be walked through by the agent's "improvisation" module to play phrases that sound appropriately according to the context of what has already been played [4]. At the time of the redaction of this article, a human operator performs such choices as deciding to play or to be silent, activating particular regions of the past memory etc.. It should also be noted that at this point the OMaX system has been proven efficient confronted to a single instrumentalist but less adapted when confronted to a band.

Although the input stream can be an audio signal, the memory model is entirely symbolic, so it implements a multi-scale, symbol to signal (discrete / continuous) integration scheme. Studies have shown the importance of symbolic representation in the human minds, and their exploitation allow rapid and measurable results[16].

5.2 GRETA-VIB

GRETA-VIB[13] is an ECA conceived in the Multimedia team of the LTCI-Telecom-Paristech and was developed using the Java language.

The GRETA-VIB system is a virtual embodied character that uses a modular architecture independent of the agent's embodiment . This architecture follows the SAIBA framework that specifies three modules: the intent planner, the behavior planner and the behavior realizer. These modules communicate with each other through XML-based languages (namely FML[9] and BML[12]).

The modularity is at the center of the GRETA-VIB architecture. In addition to the three modules implementing the SAIBA framework, each designer can provide the program with independent module attached to these "backbone" modules and that could specify the characteristic of the ECA, notably its behavior, independently from the way it is embodied.

In addition to these modules, vocabulary lexicon made of pairs (behaviors, meaning) has been also created. The production of the lexicon can be modulated by expressivity parameters; it allows simulating virtual agents with various expressive styles that could be linked to their emotional states, personality, etc.

It must also be noted that the platform can control different embodiments, like to control virtual agents or humanoid robots, to be displayed in a 3D environment or to appear in a web-based application.

5.3 Building from Use Cases Similarities

Both architectures related to our 2 use cases show similarities in their underlying structures.

GRETA-VIB and OMAx are both interacting agents, and also have a retroactive loop that allows them to update their representation from what is being said or played by the human interactant. However, they both possess a second retroactive loop, which is listening to their own interaction (see Fig. 1 and Fig. 2).

[2] http://cycling74.com/products/max/

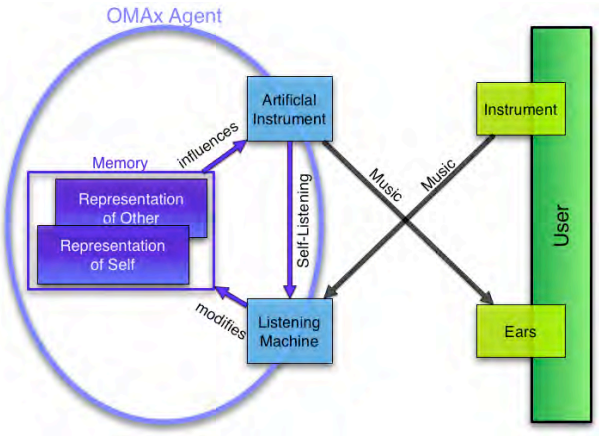


Fig. 1 General ideal scheme of OMAx

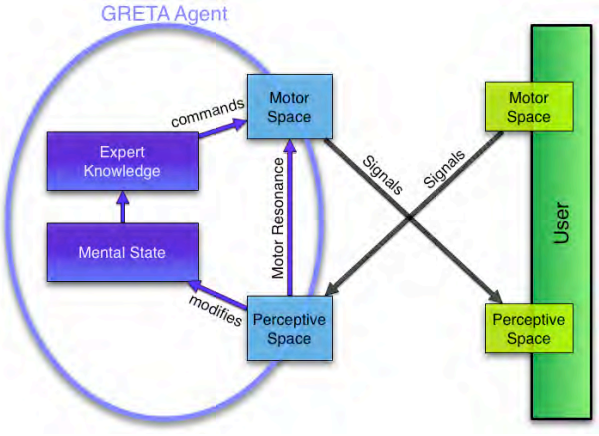


Fig. 2 General Architecture of GRETA

Using these loops, the interacting agent is able to learn not only from its interlocutor or fellow improviser but also from its own abilities, leading it to amend its representation in a more believable way, since it takes into account both interactants.

Since these models are similar in their conceptualization, we propose that our model be conceived in a top-down approach; by reifying the concepts at stakes in both models and their embodiment in each field of research. Trough this approach will we be able to model our goal of adaption of interaction.

6 Conclusion

To perform the cooperative and communicative interactions that are multimodal conversation and musical improvisation, these interactions need to reach a certain level of synchronization. To provide a believable interaction, we aim to use emergent properties of the interaction between our agents.

In this paper, we have defined what aspects of adaptation that are relevant in our context. We showed through use cases how we intend to exploit this closeness in two specific contexts.

Acknowledgments

This work was performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02.

Bibliographie

[1]Allauzen, Cyril, Maxime Crochemore, and Mathieu Raffinot. "Factor oracle: A new structure for pattern matching." SOFSEM'99: Theory and Practice of Informatics. Springer Berlin Heidelberg, 1999.

[2] Allwood, Jens. "Linguistic communication as action and cooperation." Gothenburg monographs in linguistics 2 (1976): 637-663.

[3] Assayag, Gérard, and Shlomo Dubnov. "Using factor oracles for machine improvisation." Soft Computing 8.9 (2004): 604-610.

[4]Assayag, Gérard, and Georges Bloch. "Navigating the oracle: A heuristic approach." International Computer Music Conference. Vol. 7. 2007.

[5] Baron-Cohen, Simon. "The evolution of a theory of mind." The descent of mind: Psychological perspectives on hominid evolution (1999): 261-277.

[6] Clair, Gael, Frédéric Armetta, and Salima Hassas. "Self-adaptive tuning of dynamic changing problem solving: a first step to endogenous control in Multi-Agents Based Problem Solvers." ICAS 2011, The Seventh International Conference on Autonomic and Autonomous Systems. 2011.

[7]Donnay, Gabriel F., et al. "Neural Substrates of Interactive Musical Improvisation: An fMRI Study of 'Trading Fours' in Jazz." PloS one 9.2 (2014): e88665.

[8] Fujie, Shinya, et al. "Spoken dialogue system using prosody as para-linguistic information." *Speech Prosody 2004, International Conference*. 2004.

[9] Heylen, D., et al. "Why conversational agents do what they do? Functional representations for generating conversational agent behavior." The First Functional Markup Language Workshop. Estoril, Portugal. 2008.

[10]Ishi, Carlos Toshinori, Hiroshi Ishiguro, and Norihiro Hagita. "Analysis of relationship between head motion events and speech in dialogue conversations." *Speech Communication* 57 (2014): 233-243.

[11]Knapp, Mark, Judith Hall, and Terrence Horgan. Nonverbal communication in human interaction. Cengage Learning, 2013.

[12]Kopp, Stefan, et al. "Towards a common framework for multimodal generation: The behavior markup language." Intelligent virtual agents. Springer Berlin Heidelberg, 2006.

[13] , Radosław, et al. "Cross-media agent platform." Proceedings of the 16th International Conference on 3D Web Technology. ACM, 2011.

[14]Novielli, Nicole, Fiorella de Rosis, and Irene Mazzotta. "User attitude towards an embodied conversational agent: Effects of the interaction

mode." Journal of Pragmatics 42.9 (2010): 2385-2397.
[15]Peters, Christopher, et al. "Engagement capabilities for ecas." AAMAS'05 workshop Creating Bonds with ECAs. 2005.
[16]Sun, Ron. "Autonomous generation of symbolic representations through subsymbolic activities." Philosophical Psychology 26.6 (2013): 888-912.
[17]Teo, Yong Meng, Ba Linh Luong, and Claudia Szabo. "Formalization of emergence in multi-agent systems." Proceedings of the 2013 ACM SIGSIM conference on Principles of advanced discrete simulation. ACM, 2013.
[18]Thórisson, Kristinn R. "Natural turn-taking needs no manual: Computational theory and model, from perception to action." Multimodality in language and speech systems. Springer Netherlands, 2002. 173-207.
[19]Vinciarelli, Alessandro, et al. "Bridging the gap between social animal and unsocial machine: A survey of social signal processing." Affective Computing, IEEE Transactions on 3.1 (2012): 69-87.

MyBlock/AgentSlang : une plateforme pour le déploiement d'ACA

Ovidiu Şerban[1] Alexandre Pauchet[2]

[1] ISR Laboratory, University of Reading, United Kingdom : o.serban@reading.ac.uk
[2] LITIS, INSA de Rouen, France : alexandre.pauchet@insa-rouen.fr

Résumé

Cet article décrit MyBlock, une plate-forme générique dédiée à la conception et au déploiement de systèmes interactifs comme les Agents Conversationnels Animés. La plate-forme est basée sur une approche par composants, avec une structure flexible permettant de faciliter l'organisation des différents éléments et permettre une intégration facilitée de systèmes déjà existants. AgentSlang consiste en une collection de composants originaux et de composants encapsulant des algorithmes et logiciels existants, intégrés à MyBlock. La plate-forme intègre un système d'échange de messages efficace dont les performances lui permettent de surclasser les plate-formes existantes de la littérature. Elle propose de plus la vérification d'intégrité des données ou encore la confirmation de réception de messages.

Mots clefs : Systèmes Interactifs, Plate-forme d'ACA, Benchmark de systèmes

1 Introduction

La conception d'environnements virtuels, avec lesquels les utilisateurs peuvent interagir de manière naturelle, reste un problème non résolu. Dans ce cadre, les Agents Conversationnels Animés (ACA, ou ECA -Embodied Conversationnal Agent - en anglais [Ogan et al., 2012]), en raison de leur apparence hyper-réaliste, déçoivent souvent les attentes des utilisateurs quant à leurs capacités réelles. Ce phénomène est appelé "*uncanny valley*" [Mori, 1970]. En particulier, à cause de la complexité à mettre en place un système de dialogue complet, les ACA n'intègrent que des processus très basiques de gestion du dialogue, comme les systèmes par mots-clefs ou des approches dîtes "frame-based" (voir par exemple SEMAINE [Schröder, 2010]). Ainsi, la gestion du dialogue reste à ce jour assez inefficace dans les ACA [Swartout et al., 2006].

Permettre une interaction riche et multimodale, issue de multiples entrées, tout en conservant un traitement rapide et fiable des données, est le challenge que nous proposons de résoudre. L'interaction Homme-Machine pose des problèmes à différents niveaux comme, de manière non exhaustive, la retranscription et la détection de caractéristiques vocales, la reconnaissance d'expressions faciales ou de postures, le traitement de la langue naturelle, la gestion du dialogue et d'émotions, la génération de la parole ou encore l'animation réaliste d'un personnage virtuelle. Chacune de ces étapes a déjà été traitée indépendamment ou conjointement, mais à notre connaissance aucun système interactif n'intègre l'intégralité de ces éléments.

Cet article est structurée de la manière suivante : nous présentons tout d'abord un bref état de l'art sur les systèmes interactifs (section 2). La section 3 est dédiée à la plate-forme MyBlock et à ses différents composants AgentSlang, suivie section 4 d'une évaluation de ses performances. Nous concluons cet article par une courte discussion.

2 Systèmes interactifs à base d'ACA

Un système interactif, comme tout système complexe, comporte de multiples composants. Ceux en charge du traitement des entrées utilisateur se formalisent comme des Extracteurs d'Information, un Gestionnaire du Dialogue s'occupe de l'interaction et enfin des composants de sortie sont des Générateurs de Comportement Multimodal associés à des Players. Un player peut être une simple interface vocale, voire textuelle, aussi bien qu'un ACA ou un robot. Dans cette section, nous nous focalisons sur les plate-formes dont le player est suffisamment générique pour être adapté à n'importe quel scénario et environnement.

2.1 Multiplatform

Les premières plate-formes de systèmes interactifs ont été conçues autour du protocole PVM [Geist et al., 1994]. Ce protocole a été intégré au projet Multiplatform [Herzog et al., 2004] afin de permettre le développement de deux projets célèbres : Verbmobil [Wahlster, 2000] et Smartkom [Wahlster, 2006]. Malheureusement, le projet n'est actuellement plus maintenu.

2.2 Psyclone

Une seconde génération de systèmes s'appuie sur le l'intergiciel Psyclone [CMLabs, 2007], qui intègre un protocole de communication par tableau noir. Ce choix permet une intégration facilitée mais ralentit considérablement les échanges d'information entre composants, comme le montre le test de performance effectué dans le projet Semaine [Schröder, 2010].

Parmi les systèmes utilisant Psyclone, on citera Mirage [Thórisson et al., 2004], dont le code source n'est malheureusement pas accessible publiquement, et GECA [Huang et al., 2008]. En particulier GECA, pour Generic Embodied Conversational Agent framework est

dédié à la conception d'ACA capables de traiter des entrées verbales et non verbales, de générer de la voix, des gestes et des postures et de réaliser des fonctions dialogiques basiques.

2.3 Companions

Companions [Cavazza et al., 2010] n'est pas qu'un simple ACA mais plus un compagnon artificiel engagé dans une interaction à long terme afin de forger une relation empathique avec l'utilisateur. Compagnons est construit autour d'un scénario de type *"How was your day ?"*, qui permet un dialogue relativement ouvert autour de ce thème. La plate-forme est malheureusement relativement fermée en raison d'une technologie propriétaire.

2.4 Semaine

Semaine [Schröder, 2010] est conçu afin de permettre ce que l'on peut qualifier d'interaction affective. Le projet est construit autour d'un personnage virtuel capable de reconnaître les émotions de l'utilisateur au travers d'un dispositif de captation multimodale, et d'y répondre en fonction des émotions perçues. La réponse en elle-même n'est pas forcément une réaction directe puisqu'une certaine planification du comportement de l'agent est pris en compte.
Plusieurs personnages aux personnalités variées sont proposés, avec différents modèles de réaction aux émotions perçues. Le système est géré par une architecture à base de composants, dans lequel chaque élément fonctionne indépendamment des autres. La détection de l'émotion de l'utilisateur est une fusion des caractéristiques de la voix de l'utilisateur extraites à l'aide d'OpenSMILE [Eyben et al., 2010] et d'expressions faciales détectées en utilisant iBug [Soleymani et al., 2012]. Le comportement de l'agent est traité par deux composants : un premier, vocal, synthétise une voix : MaryTTS [Pammi et al., 2010]. Le second convertit gestes et expressions faciales planifiés en code BML jouable par Greta [Poggi et al., 2005].

2.5 VHToolkit

Virtual Human Toolkit (VHToolkit) [Hartholt et al., 2013] est une plate-forme générique permettant la conception d'ACA. Elle a été utilisée avec succès dans diverses applications allant du e-learning à de l'entraînement militaire, démontrant ainsi sa généricité.
VHToolkit propose une collection de composants dédiés à toutes les étapes indispensables d'un système interactif : retranscription, génération de voix à partir de texte, gestion du dialogue (en utilisant NPCEditor [Leuski and Traum, 2011]), un générateur de postures [Lee and Marsella, 2006] et une couche de perception formalisée autour de PML [Scherer et al., 2012]). Enfin, le projet intègre SmartBody Embodiment [Shapiro, 2011] comme interpréteur de comportement BML.

2.6 Discussion

Les fonctionnalités clefs pour une plate-forme de systèmes interactifs performants sont, de notre point de vue : un gestionnaire de dialogue, un composant pour la gestion du comportement affectif (détection et génération) ou boucle de rétroaction affective ainsi qu'une architecture performante pour l'échange de messages entre les différents composants. Dans ce cadre, seuls les projets Semaine et Companions proposent une gestion des émotions, alors qu'un système de dialogue est intégré par VHToolkit, Companions, Mirage et GECA. De plus, du point de vue de l'échange de messages, tous ces systèmes ne proposent que des échanges soit par messages XML lourds ralentissant le système, soit par chaînes de caractères propriétaires empêchant toute extension future.
Dans la suite de cet article, nous décrivons AgentSlang, notre proposition de plate-forme, qui essaye d'intégrer les aspects positifs de chacun de ces travaux tout en en améliorant les performances. Nous nous concentrons ici principalement sur la plate-forme en elle-même, en proposant décrivant uniquement des composants basiques pour gérer le dialogue et détecter l'émotion de l'utilisateur.

3 MyBlock et AgentSlang

AgentSlang est une collection de composants, intégrés à l'intergiciel MyBlock qui permet de construire des systèmes distribués rapides et riches. MyBlock assure la communication entre les différents composants, en proposant un système d'échange de messages efficace. Il propose une couche de communication transparente pour les composants d'AgentSlang, facilitant ainsi la conception de systèmes interactifs.
Le protocole de transport utilisé dans MyBlock est ZeroMQ [Hintjens, 2013], qui permet diverses méthodes de connexion, supporte plusieurs patrons de communication et offre un système totalement décentralisé, sans broker, composant source de perte de robustesse et ralentissant habituellement les architectures.

3.1 Conception orientée données

Même si les formats de données n'ont pas à être exactement identiques entre chaque couple de composants échangeant des messages, une compatibilité minimale doit être assurée. Deux grands choix sont possibles :

1. des types de données conçus spécialement pour avoir une taille de transfert minimale,
2. une représentation générique, utilisant un format standard (i.e. JSON ou XML par exemple)

Les représentations ad-hoc sont très populaires car elles permettent des performances optimales. Cependant, la maintenance et l'extension de tels types de données est très difficile. D'un autre côté, l'utilisation de standards de formalisation pour l'échange de données permet de faciliter la maintenance et l'extensibilité des systèmes mais tend à augmenter la taille des données à transporter, réduisant ainsi les performances des systèmes les utilisant.
Google Protocol Buffers [Google, 2012] présente par exemple une comparaison entre leur propre format de sérialisation et un représentation XML et conclut un gain

de performance de 10 à 100 en sérialisation/désérialisation de données. MsgPack [Sadayuki, 2012] poursuit le même type d'approche en offrant des performances encore améliorées par rapport à celles de Google. Ainsi, la standardisation des types de données nous parait une approche plus en adéquation avec le déploiement par exemple de services web à large échelle, plutôt que pour l'échange de données entre composants d'un ACA. En effet, pour un système interactif, le temps de traitement de certains algorithmes peut déjà être assez long (par exemple pour un composant de retranscription) et nécessite donc de minimiser les temps d'échange d'information entre composants.
Nous proposons donc de définir notre propre format de données, optimisé pour un système interactif. Notre système de représentation des données est actuellement orientée objets, et ses différents éléments sont extensibles et indépendants de la sérialisation, afin d'en changer par la suite si nécessaire. La sérialisation est effectuée avec MsgPack [Sadayuki, 2012], le système le plus efficace actuellement. Ainsi, l'empreinte mémoire de nos données est minimale, sérialisable/désérialisable rapidement tout en conservant les avantages d'une représentation objets native.

3.2 L'intergiciel MyBlock

Sur la base d'une conception orientée objets pour la représentation des données, d'une sérialisation/désérialisation par MsgPack et d'une couche de transport assurée par ZeroMQ, l'intergiciel MyBlock a été conçu. MyBlock est ainsi peu gourmand en mémoire, flexible et rapide pour permettre l'échange d'information en quasi temps réel.
MyBlock, comme la plupart des plate-formes modernes, s'articule sur trois niveaux afin de faciliter le déploiement : composants, services et architecture.

Les composants MyBlock. Un composant MyBlock est la structure atomique minimale pour la plate-forme. Chaque composant d'une chaîne de traitements MyBlock traite un ensemble de données typées en entrée avant d'en renvoyer le résultat aux éléments suivants de la chaîne.
Chaque composant peut être soit réactif (i.e. la sortie est uniquement une fonction de l'entrée), ou actif s'il produit lui-même une donnée en sortie sans forcément recevoir pour autant de données en entrée. Certains éléments spéciaux ne peuvent que recevoir des données (Sink) ou en produire (Source).
Au niveau Composant, les types de données ont une importance particulière puisqu'ils régissent les échanges d'information entre composants. Un composant MyBlock est ainsi défini formellement par des préconditions d'entrée et de sortie en terme de types de données. Ainsi, toute architecture intégrant ce composant doit en respecter les préconditions d'entrée et de sortie pour être fonctionnelle.
Le protocole de communication entre deux composants est un simple publish-subscribe, supporté nativement par ZeroMQ, afin de faciliter les échanges d'information.

Les Services MyBlock. Similairement au niveau Composants, nous définissons un niveau Services. Un Service MyBlock est conçu pour répondre à des requêtes en provenance de n'importe quel autre composant ou service.
Le protocole de communication entre deux services est un request-reply synchrone, là encore supporté par ZeroMQ.

Les Architectures MyBlock. Au dernier niveau, celui de définissant l'Architecture d'un système, la chaîne de traitements est définie en configurant les différents paramètres des composants et services, ainsi qu'en liant dynamiquement les éléments entre eux. Bien évidemment, cette architecture doit respecter les préconditions d'entrée et de sortie de chaque composant.
Le caractère dynamique de ce niveau permet de modifier aussi bien les paramètres des différents composants et services aussi bien que la manière dont ils s'architecturent, sans avoir à recompiler les différents éléments.

3.3 Les Composants AgentSlang

Tous les principes énumérés ci-avant pour MyBlock restent valables pour AgentSlang puisqu'il ne s'agit que d'une bibliothèque de composants MyBlock, permettant de concevoir des systèmes interactifs.
La plupart des composants basiques d'AgentSlang proposés actuellement sont basés sur des librairies existantes : Google Speech API [Google, 2013] pour la reconnaissance automatique de la parole, Cereproc Voice [CereProc,] pour la synthèse vocale, un ensemble de taggers de catégorie grammaticale (SENNA [Collobert et al., 2011], TreeTagger [Schmid, 1995]) et M.A.R.C. [Courgeon et al., 2008] comme personnage virtuel animé.
AgentSlang inclut également plusieurs composants dédiés au traitement automatique de la langue, la gestion du dialogue et la détection d'émotions, ces éléments étant considérés comme indispensables pour un système interactif. La figure 1 résume les différents éléments proposés.

Composant TAL : Syn!bad . Ce composant a été conçu spécialement pour supporter un langage d'expressions rationnelles appelé Syn!bad , permettant d'extraire des informations à partir d'énoncés. Syn!bad , dont l'acronyme signifie *Synonyms [are] not bad*, est construit autour du concept de synonymes. Syn!bad est basé sur une structure d'expressions rationnelles POSIX [Alfred, 1990], étendue afin de pouvoir inclure reconnaissance de synonymes, restrictions grammaticales et gestion de variables.
Les synonymes sont des structures regroupant des ensembles de mots selon leur signification. La base de synonymes la plus connue est WordNet [Miller, 1995]. WordNet consiste en un ensemble de classes de mots regroupés selon leur sens et leur catégorie grammaticale. Ces classes, appelées synsets, sont caractérisées par un identifiant unique permettant de les identifier et retrouver.
Dans un système interactif, le processus de reconnaissance et d'extraction d'information est souvent ralenti par la complexité des règles décrivant le concept à reconnaître. Une alternative à cette complexité est l'association de d'expressions rationnelles à une structure de variables. Par exemple, la phrase

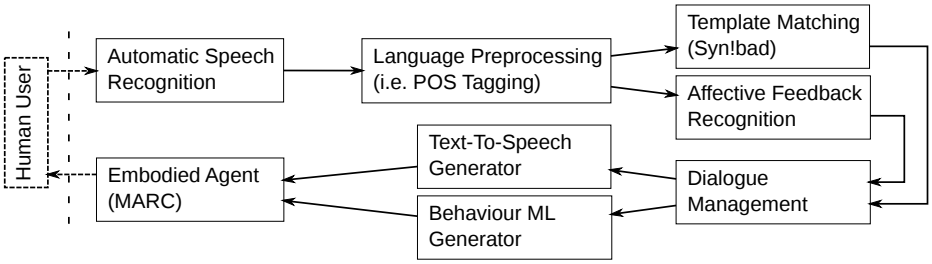


FIGURE 1 – The current status of the AgentSlang platform

`Bob do you have water` peut être reconnue par l'expression `<name> do you <verb> <object>`. L'extraction de variables dans des expressions rationnelles existe dans de nombreuses implémentations. Le problème est cependant davantage complexe quand certaines restrictions, par exemple grammaticales, sont ajoutées dans les contraintes (par exemple `<verb>` et `<object>` dans notre exemple). À notre connaissance il n'existe pas de langage permettant de reconnaître une expression rationnelle avec restriction grammaticale à part Syn!bad .

L'exemple suivant d'expression Syn!bad permet d'introduire au mieux les caractéristiques du langage :

```
$name <#*>? do you <VB*>* [some|RB*]
          [water#object]
```

pour laquelle : 1. `$name` représente un marqueur sans restriction qui accepte n'importe quel mot simple et le retourne dans la variable *name*. 2. `<#*>?` est un marqueur, ici optionnel, reconnaissant les signes de ponctuation. 3. *do* et *you* sont des mots simples. 4. `<VB*>*` est un marqueur de type "0-plusieurs" portant sur la reconnaissance de verbes. 5. `[some|RB*]` reconnaît les mots synonymes de *some*. Une restriction supplémentaire y est appliquée pour ne reconnaître que les adverbes. 6. Le marqueur `[water#object]` ne reconnaît que les synonymes de *water* et les instancie dans la variable *object*.
La précédente expression Syn!bad reconnaîtra alors la phrase suivante : `Bob, do you want any aqua`, et produira le résultat suivant : $\$name \leftarrow Bob$ et $\#object \leftarrow aqua$, tandis que `<#*>?` reconnaîtra la virgule, `<VB*>*` le verbe *want* et *any* sera reconnu par `[some|RB*]`.

3.4 Composant de gestion du dialogue et de génération de langue naturelle

Dans cette version d'AgentSlang, le gestionnaire de dialogues est un simple automate à états fini déterministe. Sa structure est définie par une suite de motifs d'interactions dialogiques déterminées lors d'une expérimentation précédente [Serban et al., 2014]. Ce composant prend en entrée une expression Syn!bad et une série de variables, et génère une action dialogique.
Une action dialogique correspond ici à un énoncé à prononcer, associé si nécessaire à une expression faciale et/ou à un geste pouvant être interprété par le player. La réponse permet, comme en détection, des substitutions de variable similaires à celle du langage Syn!bad : `item1 item2 $variable1 item3`, où `item1,item2,item3` sont des mots de la réponse et `$variable1` est une variable pouvant être instanciée, par exemple, en fonction d'un élément collecté dans l'expression Syn!bad d'entrée.

3.5 Composant de reconnaissance d'émotion

Le composant de reconnaissance de l'émotion de l'utilisateur s'appuie sur une version étendue de notre système de reconnaissance d'émotion dans des énoncés, basé sur une fusion de dictionnaires affectifs contextualisés [Serban et al., 2013].
Ce composant prend en entrée une phrase annotée grammaticalement. La valence de cette phrase est alors calculée en fonction de la proximité de la phrase avec un ensemble de contextonymes annotés par une valence [Serban et al., 2013]. Un contextonyme est une structure représentant des ensembles de mots ayant été utilisés dans un contexte commun (une phrase). Les contextonymes permettent une reconnaissance de mots en contexte facilitée en comparaison de la relation plus classique de synonymie.

4 Évaluation d'AgentSlang

Une évaluation des performances d'AgentSlang a été réalisée en comparaison des principales plate-formes existantes. La machine de test est équipé d'un chipset I7 Intel cadencé à 1.6 GHz par coeur et de 3.9 Gb de RAM. L'OS est un Linux Ubuntu 12.04, avec Oracle Java 1.7 installé.
Le dispositif envoie une série de messages aléatoires, de taille fixe, d'un composant à un autre. Nous avons choisi d'envoyer des séquences aléatoires afin d'éviter de favoriser les systèmes proposant une gestion du cache. 100 messages sont envoyés successivement et la moyenne est calculée.
La vitesse de transmission entre composants est alors calculée par le nombre de messages transmis en une seconde. La figure 2 représente les résultats de cette mesure pour les principales plate-formes, en fonction de la taille des messages et sur une échelle logarithmique.
AgentSlang, intégrant MyBlock, est présenté dans deux versions : en version sécurisée (Automatic System Feedback - ASF) et en version simple (non-ASF). L'ASF permet à AgentSlang d'envoyer une confirmation lorsqu'une

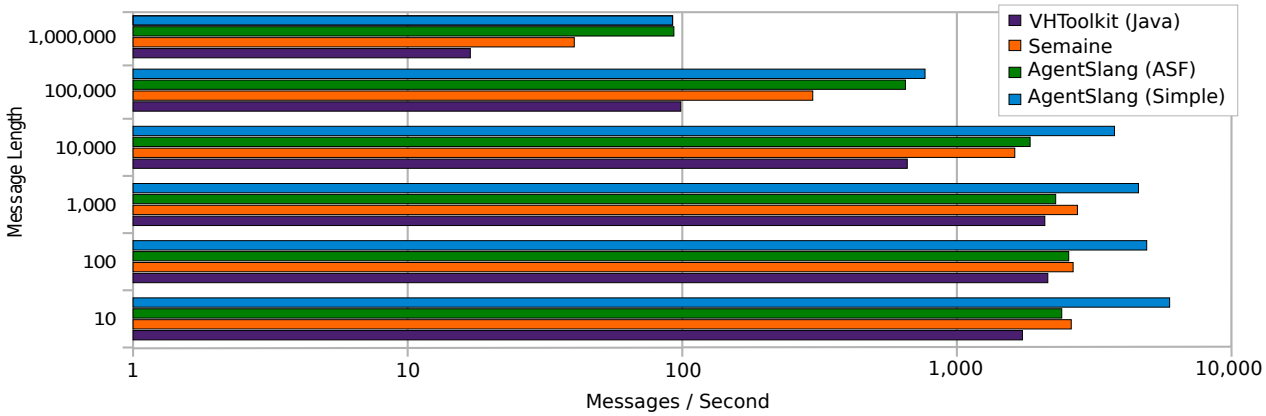


FIGURE 2 – The performance comparison between SEMAINE, VHToolkit and AgentSlang, for message throughput representation. The representation is done on a logarithmic scale.

action a été réalisée avec succès. Cette confirmation est effectuée à chaque envoi/réception de message. SEMAINE et VHToolkit ne proposant pas de mécanisme similaire, la version non-ASF correspond à un fonctionnement équivalent à ces deux plate-formes.

La conclusion de cette évaluation est qu'AgentSlang fonctionne toujours plus rapidement que SEMAINE et VHToolkit, en version non-ASF. Par ailleurs, pour des messages d'une longueur supérieure à 10,000 caractères, la version ASF ne dégrade plus les performances d'AgentSlant, pour des performances équivalentes pour les deux versions pour des messages d'une taille d'un million de caractères. Puisqu'AgentSlang vise à être utilisé avec un large éventail de types de données, chacune des deux versions peut être utilisées. Pour des performances optimales, AgentSlang simple (non-ASF) est un bon choix. Pour une version sécurisée garantissant qu'un message a bien été traité, AgentSlang ASF est actuellement la seule solution. En conclusion, les deux versions d'AgentSlang surclassent les versions actuelles de SEMAINE et de VHToolkit, et le choix entre ASF et non-ASF doit se faire selon le scénario (performance vs. sécurité).

5 Conclusion et perspectives

Dans cet article, nous avons présenté AgentSlang, une bibliothèque de composants permettant de concevoir et déployer des systèmes interactifs comme les ACA. AgentSlang est construit autour de l'intergiciel MyBlock afin de permettre les meilleurs performances possibles en terme de vitesse de transmission des messages et propose un ensemble de composants soit issus de l'encapsulation de systèmes et algorithmes existants, soit conçus spécialement dans l'objectif de traiter les éléments d'une interaction homme-machine évoluée.

De notre point de vue, une telle plate-forme doit intégrer au minimum trois composants principaux : un gestionnaire de dialogue, un composant permettant une boucle de rétroaction affective et des extracteurs d'information à partir d'énoncés. AgentSlang propose une architecture minimale autour de ces trois composants. La fonctionnalité minimale de ces composants est rendue possible par le langage Syn!bad , permettant la reconnaissance d'expressions rationnelles supportant la synonymie, la restriction grammaticale et l'utilisation de variables.

En ce qui concerne les performances, nous avons montré qu'AgentSlang surclasse Semaine et VHToolkit et propose une fonctionnalité supplémentaire en version ASF : la confirmation d'exécution.

Cependant, AgentSlang est toujours en phase active de développement et l'ensemble des composants proposés peut être amélioré. Les éléments de génération de langue naturel, par exemple, sont très sommaires et pourraient inclure des techniques permettant de créer des contenus plus naturels et non répétitifs. À l'heure actuelle, le comportement non verbal généré est attaché de manière fixe à un comportement verbal, alors qu'il devrait être créé dynamiquement, par exemple en fonction de la personnalité de l'agent, du contexte de l'interaction, etc. Le gestionnaire de dialogue est également basique puisqu'il n'inclut qu'une mémoire très limitée et un système à base d'automates. Une approche mixant jeux de dialogues, apprentissage et planification est pressentie. Enfin, la boucle de rétroaction affective comporte un élément de détection de l'émotion de l'utilisateur uniquement basé sur le texte. Plusieurs autres éléments d'entrée pourraient être pris en compte, par fusion, comme par exemple la prosodie de l'utilisateur, ses attitudes ou expressions faciales.

Remerciements

Cette démonstration fait partie du projet NARECA, financé par le programme CONTINT de l'ANR (ANR-13-CORD-0015).

Références

[Alfred, 1990] Alfred, V. (1990). Algorithms for finding patterns in strings. *Handbook of Theoretical Computer Science : Algorithms and complexity*, 1 :255.

[Cavazza et al., 2010] Cavazza, M., de la Camara, R. S., and Turunen, M. (2010). How was your day ? : a companion eca. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems : volume 1-Volume 1*, pages 1629–1630. International Foundation for Autonomous Agents and Multiagent Systems.

[CereProc,] CereProc. Cerevoice sdk. `http://www.cereproc.com/`.

[CMLabs, 2007] CMLabs (2007). Psyclone. `http://www.mindmakers.org/`.

[Collobert et al., 2011] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12 :2493–2537.

[Courgeon et al., 2008] Courgeon, M., Martin, J.-C., and Jacquemin, C. (2008). Marc : a multimodal affective and reactive character. In *Proceedings of the 1st Workshop on AFFective Interaction in Natural Environments*.

[Eyben et al., 2010] Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile : the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM.

[Geist et al., 1994] Geist, A., Beguelin, A., Dongarra, J., Jiang, W., Manchek, R., and Sunderam, V. (1994). *PVM : Parallel virtual machine : a users\'guide and tutorial for networked parallel computing*. MIT Press.

[Google, 2012] Google (2012). Protocol buffers. `https://code.google.com/p/protobuf/`.

[Google, 2013] Google (2013). `http://developer.android.com/reference/android/speech/`.

[Hartholt et al., 2013] Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.-P., and Gratch, J. (2013). All together now : Introducing the virtual human toolkit. In *International Conference on Intelligent Virtual Humans*, Edinburgh, UK.

[Herzog et al., 2004] Herzog, G., Ndiaye, A., Merten, S., Kirchmann, H., Becker, T., and Poller, P. (2004). Large-scale software integration for spoken language and multimodal dialog systems. *Natural Language Engineering*, 10(3-4) :283–305.

[Hintjens, 2013] Hintjens, P. (2013). *Zeromq : Messaging for Many Applications*. O'Reilly Media.

[Huang et al., 2008] Huang, H.-H., Cerekovic, A., Tarasenko, K., Levacic, V., Zoric, G., Pandzic, I. S., Nakano, Y., and Nishida, T. (2008). Integrating embodied conversational agent components with a generic framework. *Multiagent and Grid Systems*, 4(4) :371–386.

[Lee and Marsella, 2006] Lee, J. and Marsella, S. C. (2006). Nonverbal behavior generator for embodied conversational agents. In *6th International Conference on Intelligent Virtual Agents*, Marina del Rey, CA.

[Leuski and Traum, 2011] Leuski, A. and Traum, D. (2011). NPCEditor : a tool for building question-answering characters. In *International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

[Miller, 1995] Miller, G. (1995). WordNet : a lexical database for English. *Communications of the ACM*, 38(11) :39–41.

[Mori, 1970] Mori, M. (1970). The uncanny valley. *Energy*, 7(4) :33–35.

[Ogan et al., 2012] Ogan, A., Finkelstein, S., Mayfield, E., D'Adamo, C., Matsuda, N., and Cassell, J. (2012). Oh dear stacy ! : social interaction, elaboration, and learning with teachable agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 39–48. ACM.

[Pammi et al., 2010] Pammi, S., Charfuelan, M., and Schröder, M. (2010). Multilingual voice creation toolkit for the mary tts platform. *Proc. LREC. Valettea, Malta : ELRA*.

[Poggi et al., 2005] Poggi, I., Pelachaud, C., Rosis, F., Carofiglio, V., and Carolis, B. (2005). Greta. a believable embodied conversational agent. *Multimodal intelligent information presentation*, pages 3–25.

[Sadayuki, 2012] Sadayuki, F. (2012). Msgpack. `http://msgpack.org/`.

[Scherer et al., 2012] Scherer, S., Marsella, S. C., Stratou, G., Xu, Y., Morbini, F., Egan, A., Rizzo, A., and Morency, L.-P. (2012). Perception markup language : Towards a standardized representation of perceived nonverbal behaviors. In *The 12th International Conference on Intelligent Virtual Agents (IVA)*, Santa Cruz, CA.

[Schmid, 1995] Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer.

[Schröder, 2010] Schröder, M. (2010). The semaine api : towards a standards-based framework for building emotion-oriented systems. *Advances in Human-Computer Interaction*, 2010 :2–2.

[Serban et al., 2014] Serban, O., Bersoult, A., Alès, Z., Lebertois, E., Chanoni, E., Rioult, F., and Pauchet, A. (2014). Modélisation de dialogues pour personnage virtuel narrateur. *Revue d'Intelligence Artificielle (RIA) - Numéro spécial Affects, Compagnons Artificiels et Interaction*, 28(1) :101–130.

[Serban et al., 2013] Serban, O., Pauchet, A., Rogozan, A., and Pecuchet, J.-P. (2013). Modelling context to solve conflicts in sentiwordnet. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 393–398. IEEE.

[Shapiro, 2011] Shapiro, A. (2011). Building a character animation system. In *The Fourth International Conference on Motion in Games*, Edinburgh, Scotland.

[Soleymani et al., 2012] Soleymani, M., Pantic, M., and Pun, T. (2012). Multimodal emotion recognition in response to videos. *Affective Computing, IEEE Transactions on*, 3(2) :211–223.

[Swartout et al., 2006] Swartout, W. R., Gratch, J., Jr., R. W. H., Hovy, E. H., Marsella, S., Rickel, J., and Traum, D. R. (2006). Toward virtual humans. *AI Magazine*, 27(2) :96–108.

[Thórisson et al., 2004] Thórisson, K. R., Benko, H., Abramov, D., Arnold, A., Maskey, S., and Vaseekaran, A. (2004). Constructionist design methodology for interactive intelligences. *AI Magazine*, 25(4) :77.

[Wahlster, 2000] Wahlster, W. (2000). *Verbmobil : foundations of speech-to-speech translation*. Springer verlag.

[Wahlster, 2006] Wahlster, W. (2006). *SmartKom : Foundations of Multimodal Dialogue Systems (Cognitive Technologies)*. Springer-Verlag New York, Inc.

Curiosity-driven learning and development: How robots can help us understand humans

Pierre-Yves Oudeyer
Inria and Ensta ParisTech, France

A great mystery is how human infants develop: how they progressively discover their bodies, how they learn to interact with objects and social peers, and accumulate new skills all over their lives. Constructing robots, and building mechanisms that model such developmental processes, is key to advance our understanding of human development, in constant dialogue with human and living sciences.

I will present examples of robotics models of curiosity-driven learning and exploration, and show how developmental trajectories can self-organize, starting from discovery of the body, then object affordances, then vocal babbling and vocal interactions with others. In particular, I will show that the onset of language spontaneously forms out of such sensorimotor development.

Dr. Pierre-Yves Oudeyer is Research Director at Inria and head of the Inria and Ensta-ParisTech FLOWERS team (France). Before, he has been a permanent researcher in Sony Computer Science Laboratory for 8 years (1999-2007). He studied theoretical computer science at Ecole Normale Supérieure in Lyon, and received his Ph.D. degree in artificial intelligence from the University Paris VI, France. After working on computational models of language evolution, he is now working on developmental and social robotics, focusing on sensorimotor development, language acquisition and life-long learning in robots. Strongly inspired by infant development, the mechanisms he studies include artificial curiosity, intrinsic motivation, the role of morphology in learning motor control, human-robot interfaces, joint attention and joint intentional understanding, and imitation learning. He has published a book, more than 80 papers in international journals and conferences, holds 8 patents, gave several invited keynote lectures in international conferences, and received several prizes for his work in developmental robotics and on the origins of language. In particular, he is laureate of the ERC Starting Grant EXPLORERS. He is editor of the IEEE CIS Newsletter on Autonomous Mental Development, and associate editor of IEEE Transactions on Autonomous Mental Development, Frontiers in Neurorobotics, and of the International Journal of Social Robotics. He is also working actively for the diffusion of science towards the general public, through the writing of popular science articles and participation to radio and TV programs as well as science exhibitions. Web: http://www.pyoudeyer.com and http://flowers.inria.fr