Social Attitude Recognition in Multimodal Interaction with a Pedagogical Agent

Berardina De Carolis, Stefano Ferilli, Nicole Novielli, Fabio Leuzzi, Fulvio Rotella

Dipartimento di Informatica, University of Bari, Italy
{decarolis, ferilli, novielli}@di.uniba.it
{fabio.leuzzi, fulvio.rotella}@uniba.it

Conversational agents have been widely used in pedagogical contexts. They have the advantage of offering to users not only a task-oriented support, but also the possibility to relate with the system at social level. Therefore, besides endowing the conversational agent with knowledge necessary to fulfill pedagogical goals, it is important to provide the agent with social intelligence. To do so the agent should be able to recognize the social attitude of the user during the interaction in order to accommodate the conversational strategy. In this paper we illustrate how we defined and applied a model for recognizing the social attitude of the student in natural interaction with a Pedagogical Conversational Agent (PCA) starting from the linguistic, acoustic and gestural analysis of the communicative act.

for citations:
De Carolis B., Ferilli S., Novielli N., Leuzzi F., Rotella F. (2012), *Social Attitude Recognition in Multimodal Interaction with a Pedagogical Agent,* Journal of e-Learning and Knowledge Society, v.8, n.3, 141-151. ISSN: 1826-6223, e-ISSN:1971-8829

1 Introduction

Embodied Conversational Agents (ECAs) are used as a new metaphor of Human Computer Interaction in which the user has the feeling of cooperating with a partner, a companion rather than just using a tool (Reeves & Nass, 1996). For this purpose, an ECA must be able to: (i) recognize and answer to verbal and non-verbal inputs, (ii) generate verbal and non verbal outputs, (iii) handle typical functions of human conversations, with particular emphasis on social aspects (Cassell, 2001). In particular, when ECAs are applied in the pedagogical domain, besides handling these conversational functions, they have to fullfill pedagogical goals. In this case we talk about Pedagogical Conversational Agents (PCAs) (Johnson *et al.*, 2000). Due to these features, many learning environments have been integrated with PCAs in order to increase the learner's engagement and motivation (D'Mello *et al*., 2008; Baylor and Kim 2005). Jensen *et al.* (2012) show that PCAs can be used to improve learning and cognition in students of all ages and abilities. In particular, in their system agents are equipped with cognitive and social intelligence, personality, emotions and user awareness that increase their effectiveness and realism. In this view, Kim and Baylor (2008) argue that PCAs may improve the quality of the learning task by providing situated social interaction, that traditional computer-based learning environments often fail to provide. Thus, PCAs should be able not only to adapt their behaviour to the cognitive skills and capabilities of the learner, but also to tune their social behaviour for accommodating critical situations from the social point of view (e.g. closure and negative attitudes). To this aim, it is important to endow the agent with the capability of recognizing the learner social attitude in order to interleave a task and domain-oriented conversation with a more socially-oriented one by establishing a social relation with the students (Bickmore, 2003). To this aim, besides modeling cognitive ingredients of the user's mental state, a conversational agent should consider also extra-rational factors such as *empathy, engagement, involvement, sympathy* or *distance* (Paiva, 2004; Hoorn & Konijn, 2003).

This paper describes a study aimed at building a multimodal framework for the recognition of the social response of users to a PCA in the context of a system aimed at providing useful concepts about a correct nutrition. Since the combination of speech and gesture is a natural way for humans to interact, we propose a framework in which the agent is projected on the wall and the user interacts with it using an ambient microphone and Microsoft Kinect[1]. Then, the multimodal user input is analyzed from the linguistic, acoustic and gesture points of view. The underlying idea is that the combination of these different input modalities may improve the recognition of multimodal behaviours that

[1] http://kinectforwindows.org, (verified on July 24, 2012).

may denote the openness attitude of the users towards the embodied agent.

The framework was built as a Dynamic Bayesian Network (Jensen, 2001), due to the ability of this formalism in representing uncertainty and graduality in building an image of the user cognitive and affective state of mind.

In order to evaluate and refine the model, we designed an experimental setting to collect a corpus of multimodal conversations with a PCA in a Wizard of Oz simulation study. Then, after carefully examining our corpus and considering suggestions from the studies about verbal and non-verbal expression of social attitude, we annotated the user moves in the corpus according to the social attitude conveyed by users in each multimodal dialogue move. Then, we tested our model on the resulting dataset.

The paper is structured as follows: in Section 2 we provide a brief description of the conceptual framework; in Section 3 we describe the dynamic modelling approach used for integrating the results of the multimodal analysis; then, Section 4 provides an evaluation of the proposed framework. Finally, conclusions and future work directions are reported in Section 5.

2 Signals of Social Attitude

Since Reeves and Nass (1996) work on the media equation, in which they reported that people react to media as if they were social actors, there is a growing interest in studying the forms of anthropomorphic behavior of users towards technologies. For instance, Paiva (2004) explores the concept of empathy, Hoorn and Konijn (2003) address the concepts of *engagement, involvement, sympathy* and their contrary, *distance*. Cassell and Bickmore (2003) adopt the theory of *interpersonal relations*. In particular, in referring to the social response of users to PCAs, we distinguish warm/open from cold/close *social attitude*, according to the definition of interpersonal warmth in (Andersen & Guerrero, 1998).

The multimodal indicators of social attitude, that we employ in our approach, concern signals deriving from linguistic, acoustic and gesture analysis. The signals of social attitude in the linguistic part of the student's communicative act are recognized according to the approach described in (Novielli *et al.*, 2010). In this work, a taxonomy of signals for analyzing social communication in text-based interaction is defined. Indicators of *affect, cohesion* and *interaction* (i.e. talking about self, expressing feelings, using a friendly style, expressing positive or negative comments, and so on) are used as linguistic signals of social attitude.

However, according to several studies (Litman *et al.*, 2003; Sundberg *et al.*, 2011), linguistic analysis is not enough to properly interpret the real user's communicative intention and his attitude towards an embodied agent. For in-

stance, the user can pronounce the same sentence with different emotional attitudes in order to convey different meanings and to show a different attitude (Bosma & André, 2004). In order to classify the social attitude of the user from speech, we decided to use a bi-dimensional model that describes the affective space by two parameters: valence, indicating the hedonic value (positive vs. negative), and arousal, indicating the emotional intensity (from low to high; Russell, 2003). Recognising the value of only these two dimensions is justified since the valence indicates a failure/success in the achievement of the user's goal and, if related to the arousal, it allows to distinguish for instance a negative/cold attitude towards the agent from sadness related to a personal mental state. Therefore, a *positive* valence is a signal of positive feedback, comment, agreement towards the agent, while a *negative* one indicates a disagreement or a negative comment.

In order to classify the attitude of the user expressed through speech we used the corpus annotated in terms of valence and arousal used in (De Carolis *et al.*, 2012). In parallel with the annotation process, the audio files relative to the moves in the corpus were analyzed using Praat[2] functions in order to perform a macro-prosodic or global analysis and to extract from the audio file of each move features related to the variation of the fundamental frequency (f0), energy and harmonicity. Then we consider also the spectrum central moment, gravity centre, skewness, kurtosis and the speech rate. At present, our classifier exploits the NNge algorithm and recognizes the valence with an accuracy of 89%, evaluated on a dataset of 4 speakers and 748 user moves overall, and validated using a *10 Fold Cross Validation technique*.

In order to endow an embodied agent with the ability of recognizing the social attitude also from gestures, according to the literature, we considered those involving arms and hands position. Arms are quite reliable indicators of mood and feeling, especially when interpreted with other signals. For example, crossed arms act as defensive barriers, indicating closure; using an arm across the body denotes nervousness or a negative attitude. Conversely, arms in open positions (especially combined with open palms) indicate feelings of openness and security. However, since we perform gesture recognition using Microsoft Kinect, we had to consider only a subset of gestures compatible with the nodes in the skeleton that the Kinect SDK is able to detect.

Hands are also very expressive parts of the body as well, used a lot in signaling consciously or unconsciously feelings and thoughts. Since at present Kinect skeleton does not include nodes for detecting the position of fingers, we are able to recognize only simple hands gestures like hands picking nose (denoting social disconnection or stress), neck scratching (expressing doubt or disbelief), running hands through hair (indicating vexation or exasperation).

[2] http://www.praat.org/ (verified on February 24, 2007).

Even if the position of the legs cannot be considered as a part of gesture, in evaluating the social attitude we take into account whether the legs are crossed or not, to support the corresponding arms signals (in conjunction with crossed arms they indicate a strong closure or rejection or insecurity).

For more details, the reader is referred to (De Carolis *et al.*, 2012).

3 Modeling the User Social Attitude

The user modeling procedure integrates the results of language and prosodic analysis with gesture recognition into a Dynamic Belief Network (DBN) (Jensen, 2001). The DBN formalism is particularly suitable for representing situations which gradually evolve from a dialog step to the next one since time slices (local belief networks) are connected through temporal links to constitute a full model. The DBN (Figure 1) is used to infer how the user's social attitude evolves during the dialog according to signals expressed in the verbal and non-verbal part of the communication.

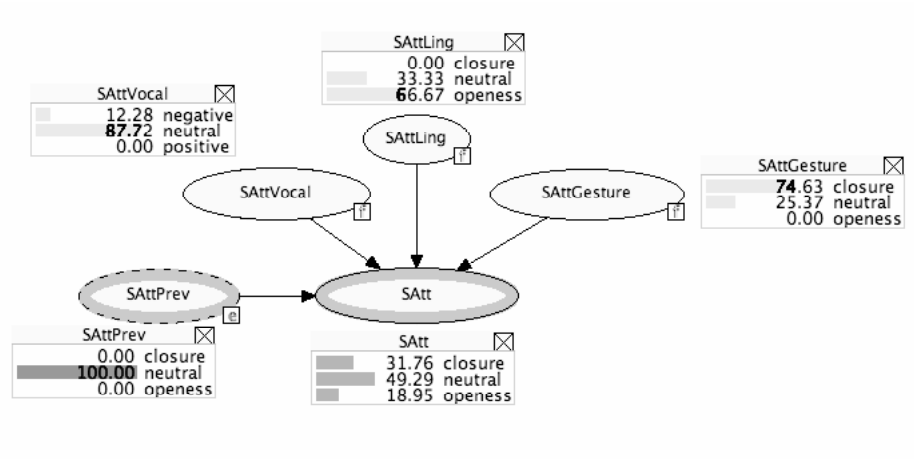

Fig. 1 -DBN modelling the user social attitude.

Social attitude (SAtt) is the variable we want to monitor, which depends on *observable* ones: the recognized significative signals in the user move deriving from the linguistic, acoustic and gestural analysis. These nodes may correspond to a simple variable, as in the case of SAttVocal, or to a nested belief network as in the case of the SAttLing and SAttGesture whose probability distribution is calculated by the corresponding belief networks.

For instance, Figure 2 shows the BN for recognising signals of social attitude from gestures. In particular, the gestures recognized by Kinect become

the evidence of the root nodes of this model. This evidence is then propagated in the net and the probability of the SAttGesture node is computed given the probabilities of intermediate nodes, Hands, Arms and CrossedLegs, denoting the social attitude expressed by each of them.

At the beginning of interaction, the model is initialized; at every dialog step, knowledge about the evidence produced by the multimodal analysis is entered and propagated in the network in order to compute the probabilities of the social attitude node. The probability of the social attitude node supports revising high-level planning of the agent behavior.

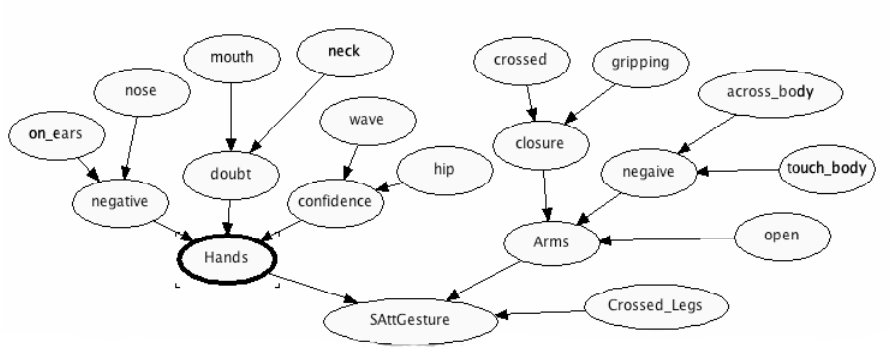

Fig. 2 - The BN corresponding to the SAttGesture node.

4 Evaluating the model

In order to perform an evaluation of the model, we started an experiment for collecting multimodal dialogues for tuning the probabilities of our model using a Wizard of Oz study. Participants involved in the study were 10 Italian students aged between 16 and 25, equally distributed by gender. They were divided in two groups, composed by 5 people each. We assigned to each group the same goal of information seeking: *Getting information about correct nutrition*.

To obtain this information, subjects could dialogue with the PCA playing the role of an expert in nutrition. Before starting the experiments we administered to each subject a simple questionnaire aimed at collecting some personal data (age and gender) and at understanding their background (department, year of course, Artificial Intelligence background). Subjects were told they had to both provide a final evaluation of the agent and to answer a series of questions about the degree of recalling, in order to test the effectiveness of the interaction with the PCA with respect to both the engagement in the interaction and the information provision task. When finished, the subject had to compile a questionnaire about the topic.

Following the described approach we collected a corpus of about 300 moves. Each move was recorded using a wireless microphone whose output was sent to the speech processing system. We assigned to each utterance a unique identifier, the corresponding transcript and the related 16-bit single channel, 8 kHz signal (in.wav format). For recording gestures, we used Microsoft Kinect and Dynamic Time Warping (DTW)[3] for the recognition of signals concerning arms, hands and legs.

Each move was annotated by three human raters (researchers in the field of human-computer interaction) with respect to the perceived user social attitude, conveyed by the turn. Each move received a final label for the social attitude using a majority agreement criterion. Then we automatically annotated the single moves using the modules defined for the language and speech analysis and the DTW for the gestures. This evidence constitutes the features describing each user move, according to the signals variables in the previously described models. According to the result of the annotation experiments (both manual and automatic) we conducted a preliminary evaluation for tuning the structure and the probabilities of our multimodal framework. Results of this experiment show that the model is quite accurate since it agrees with the human annotation at 89% in recognizing the overall social attitude.

4.1 An Example of Dialogue

In this section we show a sample dialogue (translated from Italian) from our corpus and the recognized social attitude for significant moves. In this dialogue the agent and user moves are denoted respectively by Vi (generated by our dialogue simulator) and Uj (performed by the user). The letters L, V and G denote signals of social attitude in the Language, Voice and Gestures respectively. Satt indicates the probability value of recognized overall social attitude in the user move. Where no annotation is provided, the model did not recognize any signal of social attitude. U0 denotes the initial attitude of the user before starting the interaction.

U0: (L:-, V: -, G: crossed arms) – Satt: 0.5
V1: Hi, my name is Valentina! I'm here to give you information about correct nutrition.
U1: Hi, my name is Carla. (L:friendly self introduction- V: positive valence, medium arousal - G: crossed arms) – Satt: 0.61
V2: Hi Carla, nice to meet you. How can I help you?
U2: I would like to learn more about a correct diet! (L: neutral – V: positive valence, medium arousal – G: open arms) – Satt: 0.8

[3] http://kinectdtw.codeplex.com, (verified on July 24, 2012).

V3: Good. Let's start with a pratical example. What did you eat for breakfast this morning?
U3: Coffee and some cookies.
V4: And for lunch?
U4: Just a sandwich and a pastry.
V5: What did you eat for dinner yesterday night?
U5: Meat and salad and an ice cream at the end.
V6: Well your diet is not so bad. However, like you, most of people forget about the
importance of eating more fruit and vegetables and reducing fat and sugar. International Research about foods proves that fruit and vegetables are very important in a healthy diet. They provide fibre and vitamins that are very important … (benefits of vitamins and fibers). Now, try to think again about your diet, is it healthy?
U6: Well, I should change something (L: talk about self – V: neutral valence, low
arousal – G: touching nose) - Satt: 0.48
V7: Good, let's try to do something then. Did you ever think to follow a correct diet?
U7: Yes, sometimes I did it and I did not get very good results (L: talk about self- V:
negative valence, medium arousal, G: using an arm across the body). – Satt: 0.21
V8: Why?
U8: I don't know. (L:- V: negative valence, low arousal- G: crossed arms). – Satt: 0.08
V9: OK I understand that sometimes talk about self can be difficult….

For instance, in the move U7, the linguistic analysis provides an evidence of Talk_about_Self to allow the recognition of a warm social attitude through linguistic analysis. However, the acoustic analysis classifies the valence as negative and the arousal as medium and the recognized gesture is touching nose, thus denoting a negative/cold attitude. Figure 3 shows the model results in this case.

Then the agent, in the move V8, asks which is the reason of this result. In the next move U8 the user says that she does not want to answer, with a negative prosody and crosses her arms. These signals provide evidences of a negative/cold social attitude. Then the agent, having recognized this attitude, tries to cope with the situation by changing dialog strategy.

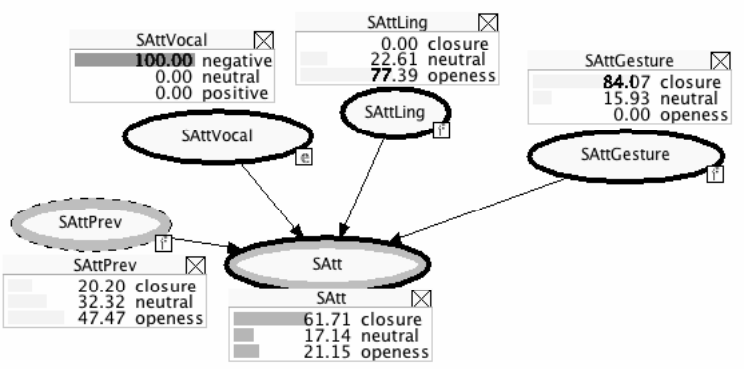

Fig. 3 - Recognition of the social attitude for move U7.

Conclusions

The use of ECAs in computer-based learning context has been widely investigated and there is not a strong evidence of pedagogical benefits deriving from their use (for more details: Gulz, 2004). Moreover, if the social agent's behavior is not properly designed and implemented there could be the risk of creating unrealistic expectations on the part of the users, and to lead to wrong mental models about the system's functionality and capacity (Brahnam & De Angeli, 2008). On the other hand, some studies report successful results on how socially intelligent agents can be employed as interaction metaphors in the pedagogical domain (Moreno *et al.*, 2001; Marsella *et al.*, 2003; Jensen *et al.*, 2012) where it is important to settle long-term relations with the user (Bickmore, 2003). In these successful experiences the agent is endowed with the capability of modeling not only the cognitive ingredients of the user's mental state (interests, preferences, beliefs), but also extra-rational factors such as *affect, engagement, attitude*. In this paper we presented a model for recognizing the social attitude from the analysis of signals regarding verbal and non-verbal communication: language, prosody and gesture in particular. The proposed model has been validated with satisfying results. In our future research we plan to improve the gesture recognition analysis. We are currently testing the proposed approach with the FUll Body Interaction (FUBI) framework (Kistler *et al.*, 2008) since it allows for hands recognition. However, we do not see this as a limitation of our approach since new devices, like the new Kinect 2, should allow for a better gesture recognition, thus improving the precision of the model. In the near future, we plan to perform more evaluation studies and we plan to test different strategies for recovering the dialogue when the social

attitude starts decreasing by extending the model with contextual features in order to use it in a prognostic way.

REFERENCES

Andersen P.A., Guerrero L.K. (1998), *Handbook of Communication and Emotions: Research, theory, applications and contexts*, Academic Press.

Baylor A., Kim Y. (2005), *Simulating instructional roles through pedagogical agents*, IJAIED, 15(1), 95-115.

Bickmore T. (2003), *Relational agents: Effecting change through human-computer relationships*, MIT, Media Arts & Sciences.

Bosma W.E., André E. (2004), *Exploiting Emotions to Disambiguate Dialogue Acts*, in: Conference on Intelligent User Interfaces, 85-92, ACM Press.

Brahnam S., De Angeli A. (2008), *Special issue on the abuse and misuse of social agents*, Interacting with Computers, 20(3): 287-291.

Cassell J. (2001), *Embodied Conversational Agents: Representation and Intelligence in User Interface*, AI Magazine, 22(3), 67-83.

D'Mello S.K., Craig S.D., Witherspoon A.W., McDaniel B.T., Graesser A.C. (2008), *Automatic Detection of Learner's Affect from Conversational Cues*, User Modelling and User-Adapted Interaction, 18(1-2), 45-80.

De Carolis B., Ferilli S., Novielli N. (2012), *Towards a Model for Recognising the Social Attitude in Natural Interaction with Embodied Agents*, in Proc. 5th Int'l Workshop on Intelligent Interfaces for HCI, 552-559, IEEE.

Gulz A. (2004), *Benefits of virtual characters in computer based learning environments: claims and evidence*. International Journal of Artificial Intelligence in Education, 14(3), 313-334.

Hoorn J. F., Konijn E. A. (2003), *Perceiving and Experiencing Fictitional Characters: An integrative account*, Japanese Psychological Research, 45, 250-268.

Jensen F. V. (2001), *Bayesian Networks and Decision Graphs*, Springer.

Jensen A. S., Jordine K., Sakpal R., Wilson D. M. (2012), *Using Embodied Pedagogical Agents and Direct Instruction to Improve Learning Outcomes for Young Children with Learning Disabilities*, Global Conference on Technology, Innovation, Media & Education.

Johnson W., Rickel J., Lester J. (2000), *Animated pedagogical agents: face-to-face interaction in interactive learning environments*, IJAIED, 11, 47-78.

Kim C., Baylor A. L. (2008), *A Virtual Change Agent: Motivating Pre-service Teachers to Integrate Technology in Their Future Classrooms*, Educational Technology & Society, 12(2), 309-321.

Knapp M., Hall J. (1992), *Non-verbal communication in human interaction*, Orlando, Holt, Rinehart & Winston.

Kistler F., Endrass B., Damian I., Dang C. T., André E. (2008), *Natural interaction with culturally adaptive virtual characters*. In: Journal on Multimodal User Interfaces,

Springer Berlin/Heidelberg, Vol. 1, No. 1.

Litman D., Forbes K., Silliman S. (2003), *Towards emotion prediction in spoken tutoring dialogues*, in Proc. of HLT/NAACL, 2, 52-54.

Marsella S. C., Johnson W. L., LaBore C. M. (2003), *Interactive pedagogical drama for health interventions*. In: U. Hoppe *et al*. (Eds.), Artificial Intelligence in Education: Shaping the Future of Learning through Intelligent Technologies, Amsterdam: IOS Press, 341-348.

Moreno R., Mayer R. E., Spires H., Lester J. (2001), *The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents?*, Cognition and Instruction, 19, 177-213.

Novielli N., de Rosis F., Mazzotta I. (2010), *User Attitude Towards an Embodied Conversational Agent: Effects of the Interaction Mode*, Journal of Pragmatics, 42(9), 2385-2397, Elsevier.

Paiva A. (2004), (Ed): *Empathic Agents*, in Workshop in AAMAS'04.

Polhemus L., Shih L. F., Swan K. (2001), *Virtual interactivity: the representation of social presence in an on line discussion*, in Annual Meeting of the American Educational Research Association.

Reeves B., Nass C. (1996), *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, Cambridge University Press.

Russell J. A. (2003), *Core affect and the psychological construction of emotion*, Psychological Review, 110, 145-172.

Sundberg J., Patel S., Björkner E., Scherer K. R. (2011), *Interdependencies among voice source parameters in emotional speech*, IEEE Trans. on Affective Computing, 2(3), 162-174.

Vogt T., Andre' E., Bee N. (2008), *EmoVoice - A Framework for Online Recognition of Emotions from Voice*, in Proc. 4th IEEE tutorial and research workshop on Perception and Interactive Technologies for Speech-Based Systems: Perception in Multimodal