A Persona is Not A Person: Designing Dialogs With ECAs After Wizard of Oz Simulations

A.Cavalluzzi, G. Clarizio, B. De Carolis, F. de Rosis

Intelligent Interfaces, Department of Informatics, University of Bari

{cavalluzzi, clarizio, decarolis, derosis}@di.uniba.it

http://www.di.uniba.it/intint/

INTRODUCTION

After an enthusiastic effort towards developing increasingly refined systems, research on Embodied Conversational Agents (ECAs) is experiencing a stage of critical reflection on the role this new form of interaction may play in real applications. Question issues concern which are the domains in which ECAs may be of demonstrated utility, how the aspect and the behaviour of the agent which best suit a particular domain may be selected, which agent's features affect its effectiveness, and many more. To answer these questions, several evaluation studies have been performed, with different aims and methods (see [25] for a review of the most recent of them). These studies were focused, initially, on analysis of individual images, with a comparison between 2D and 3D agents, evaluation of the 'convincingness' of emotional expressions [4] and similars. More recently, evaluation employed dynamic images in which global aspects of the agent performance were compared with those of other media (text, speech or human video: [5, 18]). The object of evaluation (a specific expression, the agent, the message or the interaction in its entirety) varies across these studies, as well as the evaluation criteria applied. As even the meaning of 'usability' is not yet clear for this interaction style, a large variety of terms has been employed to denote the aspects to investigate: naturalness, engagement, satisfaction, effectiveness, pleasantness, usefulness, attractiveness, and the most popular of them, 'believability'. The importance of these studies in orienting research efforts towards aspects to develop with particular care is unquestionable. However, this kind of studies does not yet provide clear suggestions on which *conversational style* is appropriate for ECAs, in every application domain. In addition, very little is known about how real users will behave when interacting with animated characters and how much they will empathize with them [20]. In the large majority of the available studies, the object of evaluation is a monolog, not a dialog: as a consequence, the user attitude is mainly *observational, passive*. Therefore, these studies do not contribute to defining *which interaction models with an ECA* should be implemented in every application context. For instance, which type of initiative is preferable (system-driven or user-driven?), which forms of empathy should be simulated in the agent, which expressions should be recognized in the user and, more in general, how the desired level of communication and engagement between agent and user may be established.

Which are the landmarks to which we may refer, in designing effective interactions with ECAs? On one hand, the 'media equation' proposed by Nass and Colleagues, according to which in interaction with technology (and, in particular, with ECAs), the same social rules of interaction between humans would hold [17]. By applying this theory literally, one may argue that human dialogues should be regarded as a norm and that a dialog with an ECA should resemble a dialog between humans as much as possible. A corpus of 'natural dialogs' should then be collected, from which to infer the rules to reproduce in the artificial conversational system, with the graduality that the difficulty of existing methods implies. On the other hand, however, various studies concur to reduce the optimism of the Stanford group, by claiming (and proving) that the behaviour of users interacting in natural language with a system is not exactly the same as that adopted in interacting with other humans: goals are reduced, language is simplified, the structural complexity of the dialog decreases. As Gruen experienced, the more users were asked to pretend they were dealing with a computer assistant, the more they seemed to restrict themselves to low-level commands rather than expressing their high-level goals [14]. The phenomenon is so remarkable, that Dahlback et al concluded a long set of studies by saying that "*goals in some dialogue research in computational linguistics, such as 'getting computers to talk like you and me' or developing interfaces that allow the user to 'forget that he is questioning a machine' are not only difficult to reach. They are misconceived."* [10]. If this view is accepted, natural dialogs should not be taken as the ideal situation that artificial ones should aim at reproducing.

Where is the truth, between these two extremes? As always, probably in the middle. A good reference on which to ground the design of ECAs is probably a corpus of human-machine dialogs which is enough articulate to offer examples of a variety of user attitudes and reactions. High-fidelity simulation environments, such as Wizard of Oz (WoZ) tools, provide the means for such an empirical work. They have been employed elegantly and systematically to collect corpora of dialogs in natural language, to evaluate speech-based interactions [6] and, occasionally, to evaluate the usability of ECAs [2,7,19]. In these studies, the general idea of WoZ tools was tailored to the particular needs of the studies to perform. In designing a tool which enables performing empirical simulation studies of ECAs in a variety of situations, several forms of flexibility should be considered:

a. Adaptation in the definition of the *context* in which interaction occurs;

b. Adaptation in the choice of the *agent's appearance* (age, gender, ethnicity, dressing);

c. Adaptation in the agent's *personality and culture* and, hence, in the way its comunicative acts are rendered;

d. Adaptation in the *interaction style*: user-driven, system-driven or mixed, with user input text or sppech-based;

e. Adaptation in the *evaluation criteria.*

In this paper, we describe an ongoing research which is aimed at implementing this tool and collecting a corpus of dialogs with ECAs in different contexts. We will start with a short summary of our background in the domain of ECAs to then describe the prototyping tool we implemented to perform WoZ studies in various situations. Finally, we reflect critically on our experience to illustrate its validity and its limits.

BACKGROUND

Rather than on the graphical aspects of embodied characters, our interest in the domain of ECAs is focused on the affective factors of interaction; in particular:

- how dialogs with the agent are influenced by the affective situation of both the agent and the user (their

personality and their dynamically evolving emotional states and attitudes);

- which dialog modeling methods are suited to represent adaptivity to the mentioned affective states (in particular, in advisory dialogs about problem health behaviors);
- how a solid foundation for interaction design may be created, by means of evaluation methods.

We implemented, in the last few years, several versions of a prototype dialog system [9, 11, 13]. Its peculiarity is to employ two models (of the agent and of the user) which integrate 'cognitive' aspects of the mental state (beliefs, goals, plans and their relationships) with 'affective' ones (personality, emotions and attitudes). In this prototype, we formalized how every move affects mental state and behaviour of the interlocutor and how the situation evolves dynamically during the dialog. To establish, in particular, which forms of expressions of our ECA are most effective in achieving information and persuasion objectives in the domain of our interest, we performed a set of controlled evaluation studies [5]. In these studies, which involved about 350 subjects, we compared various communication media: the agent in various modes, its voice only, a human video and (as a reference) a text. Subjective evaluations of various aspects of the agent and of the message were measured with a Likert scale, as well as 'objective' data about the recall of information items in the message. In spite of the considerable efforts in designing and performing such a large evaluation study, the results we obtained provided us only some partial answers to our original questions; at the same time, they raised some new questions. In particular, although the agent's appearance was very refined and highly realistic [22], apparently its persuasion power was a bit lower than that of a text or a human video. We concluded this study with the opinion that ECAs should be evaluated in 'interactive' contexts and settings, which try to reproduce at their best the ideal situation of 'interacting with a companion' (or an expert) for which they were originally proposed. Considering the difficulty of implementing complete prototypes to evaluate, we decided to equip us with a WoZ dialog simulation tool.

THE WOZ TOOL

The main knowledge source of our tool is a database (Study-DB) which describes the studies to perform. We will illustrate the various components of this DB by examining how the flexibility requirements mentioned in the Introduction are implemented in the tool:

a. *Adaptation in the definition of the context in which interaction occurs.* The context is defined by a text which describes the *scenario* in which the subjects involved in the study will find themselves: the application domain, the goals the agent and the subject will try to achieve with the dialog, the degrees of freedom of the subjects in their interaction with the agent. A scenario is stored in a text file when a new study is designed, becomes part of the Study-DB and is displayed to the subject when interaction begins.

b. *Adaptation in the choice of the domain and the agent's appearance.* A study is defined by a set of variables: the application domain, the social relationship between the ECA and the user, the attitude of the ECA towards the user's affective state (empathic vs non empathic) and the agent's appearance. These variables are associated with a repertory of agent's moves in Study-DB. The wizard selects a particular study to perform by setting the values of these variables on her interface as the first task of the simulation (Figure 1). We adopted the cast of characters of a commercial software (Haptek: see website), which provides a gallery of personas whose traits are diversified in ethnicity, age, gender and dressing. A particular tool of Haptek (FrameMaker) enables creating new expressions by manipulating individual portions of the character's face (eyes, eyebrows, mouth etc): we employed this tool to create new animations with varying degrees of intensity: small, medium and large smile; small, medium and large eye aperture and so on.

c. *Adaptation in the agent's personality and culture.* The idea behind this form of adaptation is that humans of different personality and culture behave differently in the same situations, in that they express differently the same cognitive and affective state [12]. Some classical examples are the differences in duration and direction of gaze between Asian and Western cultures [3] or the difference between introverts and extroverts in body movements, where extroverts *"tend to make wider movements and to approach others more freely in space"* [17]. To endow our agents with this form of adaptation, we implemented a *wrapper* which translates utterances labelled in terms of 'meanings' into agent's animations. The markup language employed to label the agent moves is APML [11]: this languages associates a meaning with entire sentences or their parts (up to individual words). For instance:

```
=========================================================
<apml><turnallocation type="take-turn"><performative type="inform">
<certainty="uncertain">As far as <topic-comment type="comment">vitamins
</topic-comment>are concerned,
</performative></turnallocation><performative type="announce">
<topic-comment type="comment"> research </topic-comment> has shown
that <affective type="happyfor">eating
<topic-comment type="comment"> the recommended levels of
vitamin A and C </topic-comment>
<certainty="uncertain">can have
<adjectival type="large">beneficial effects </adjectival>
for your appearance and health.</certainty></affective>
</performative></apml>
=========================================================
```

In this example, meanings of 'turn-taking' and 'inform' are associated with the clause 'As far as vitamins are concerned'. The sentence which follows is an 'announce', whose main topic is 'research'; an emotion of 'happy-for' is associated with the description of the benefits vitamins may produce to those who eat them, with an adjectival of 'large' to emphasize the terms 'beneficial effects' and an 'uncertain' label to denote that these effects are not warranted.

A *meaning-signal table* establishes how every meaning will be rendered by the agent, through facial expressions and speech parameters. For instance:

```
=========================================
<Tagging> <setup>
<agent name="sally.haptar"/> <volume start="80"/> <speed start="0"/>
<voice type="female"/></setup>
...
<performative type="announce">
<animation> <switchON>M_HeadNod</switchON>
<switchON>M_EyeAperture</switchON>
</animation></performative>
...
<affective type="happyfor">
<animation><switchON>M_Smile</switchON>
</animation></affective>
...
<adjectival type="large">
<animation><switchON>M_EyeAperture</switchON>
</animation></adjectival>
...
</setup></Tagging>
=========================================
```

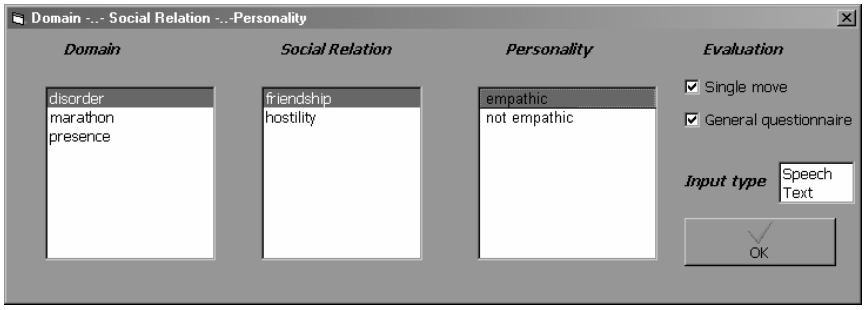

Figure 1: simulation setting (wizard's view)

The meaning of 'announce' is rendered with a head nod and a eye aperture of medium intensity; the emotion of happy-for with a smile, and the 'large adjectival' with a medium eye aperture [21]. Adaptation to the agent's personality and culture may be implemented by building, for every situation, an appropriate meaning-signal table. For instance: tables with smiles of higher intensity and wider gestures may be employed for agents representing personas from Southern Italy, than for those from Northern Italy; signals of 'smile' may be employed to express an 'embarrassment' in characters with Asian appearance, and so on. We employ two different text-to-speech synthesizers to produce the agent's voice: Microsoft TTS for the English version and Loquendo (see website) for the Italian version. In both cases, analysis of duration of phonemes drives synchronization of facial expressions (Haptek visemes) with the agent voice. Some speech parameters (volume, rate, pitch and emphasis) may be tuned, to give the voice an intonation consistent with the face expression. However, so far we did not yet introduce this kind of adaptation in our studies.

d. *Adaptation in the interaction style:* when an evaluation study is designed, a hierarchically organized set of XML files of agent moves, labelled with the APML language, is created for every condition. Moves are organized into small and clearly recognizable 'categories', such as: *Questions, Short Comments, Suggestions, Encouragements, SmallTalk* etc. Move categories may also be defined according to the agent's plan they enable implementing (as in figure 2).

When the study is applied to a subject, once the wizard has read the user move in the right top textfield of figure 2 or has heard it (if interaction is in the speech mode), he or she must select the subsequent agent move to transmit to the user. The system reads from the XML database the set of candidate sentences and shows them in a set of menus in the wizard's window, after clearing all APML tags (see again figure 2). A move may combine several sentences; for instance, a comment with a suggestion, an encouragement or a small talk. We therefore enable the wizard to select several sentences from the menus, which will be sequenced in a unique agent move. This increases the repertory of moves that may be pronounced by the agent and reduces the risk of being repetitive. However, the wizard cannot be left totally free in defining the dialog dynamics: to insure that he or she follows a well defined and consistent logic through all the study, the *dialog plan* that the wizard should apply in every study condition is specified in a paper document. In our advisory dialogs, we adopt the Stage of Change theory by Prochaska, Di Clemente and Colleagues [24]. According to this theory, addictive health-related behaviors (such as smoking, drinking, eating) evolve gradually through some recognizable stages (precontemplation, contemplation, preparation, action, maintenance and, possibly, relapse). The plan to apply in a given phase of the dialog is a function of the stage in which the subject is presumably situated in that phase [9]. Formulation of the plans the wizard is invited to follow during the dialog simulation is part of the study design. Let us assume, for instance, that the wizard presumes, from the dialog history, that the subject is in a *precontemplation phase*; that is, he does not believe that his behavior is incorrect and is not following - and not even formulating - a plan to change it. The wizard will apply a dialog plan which is aimed at achieving the following communicative goals:

Validate lack of readiness: Verify whether the subject is really not intending to take action in the forseeable future

Clarify: decision is yours: Explain that an effective change of behavior requires an intentional change

Encourage re-evaluation of current behavior: Try to reduce the subjects' resistance to think and talk about their risk behavior

Encourage self-exploration: Promote the subjects' reflection on their living style and the reason why they are adopting it

Explain and personalize risk: Inform the subject about short and long term effects of their behavior on their health, by adapting this analysis to their goals and priorities.

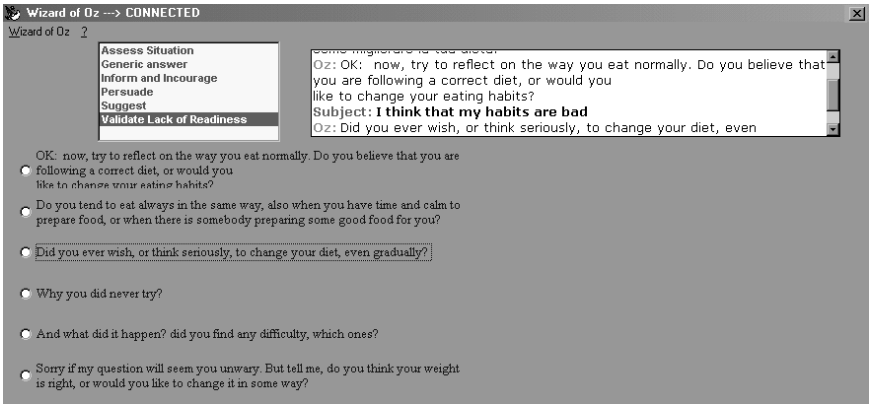

Figure 2: Move selection (wizard's view)

The plan will be applied gradually and will be revised when new information acquired will produce a revision of the presumed stage of the user.

e. *Adaptation in the evaluation criteria.* We enable three sorts of evaluation:

- Credibility of *individual moves*. This is linked to at least two factors: how appropriate was the agent move in the dialog context and how appropriate were the expressions the agent showed when pronouncing it. To avoid overloading the evaluation task during the dialog (with the risk of distracting the subject from the dialog flow), we do not ask the subject to make separate evaluations for the two aspects and leave this form of evaluation as facultative. In addition, we adopt a more 'natural' measuring method than a form compilation, which is based on selecting an 'emotional icon' from a set of alternatives: for instance, in figure 3 (rightside), the topmost icon (which is green) stands for 'good move', the second one (with a question mark) for 'unclear move' and the third one (which is red) for 'bad move'.

Figure 3: Individual move evaluation (subject's view)

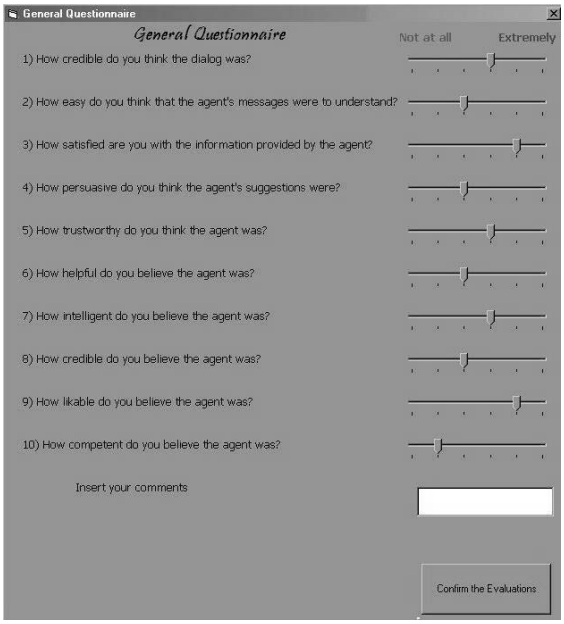

Figure 4: Final questionnaire (subject's view)

- *Effectiveness of interaction* in achieving its goal. A questionnaire which may include questions of various kinds (Yes/No, Likert scales, free text etc) is defined in the study design phase. This questionnaire is stored (again, as an XML file) in Study-DB and is displayed as a form on the subject's side, at the end of interaction. Figure 4 shows an example of questionnaire which includes subjective evaluations of the message and of the agent.

- *Relationship between agent and user behavior.* This kind of evaluation employs the corpus of subject-system interactions collected with a study or a set of studies, to investigate the effects, on the subject's attitude, of varying the agent behavior in a controlled way. Various methods have been employed for recording the user attitude during the dialog: videos of facial expressions speech or a combination of various media. Our goal is to recognize the users' attitude by means of a linguistic analysis of their behaviour [1,23]. We therefore collect, in a text file, a log of the dialog, in which we store the subject's and the agent's moves, the evaluations of individual moves (when they exist) and the results of the final questionnaire.

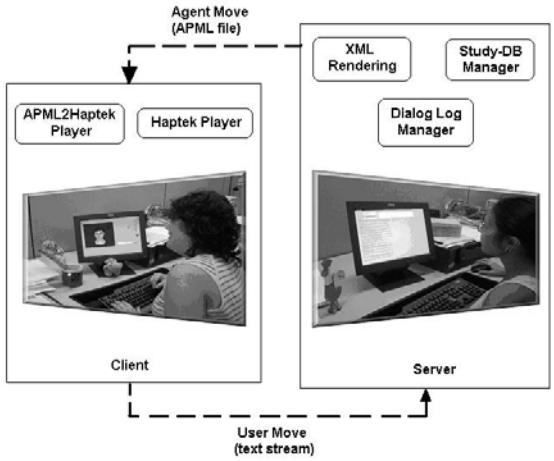

Figure 5: System architecture

A few details on the system implementation. The tool is based on a distributed architecture (figure 5) in which the subject's component is on the client side and the wizard's one is on the server side. The subject's interface accepts the user input, transmits it to the server and displays the subsequent agent move. This module includes: i) the APML-to-Haptek wrapper and ii) the character player. The server side displays the subject's input and enables the wizard to select the agent move. It includes: i) the manager of Study-DB, which selects XML files from the DB and renders them as a form and ii) a module for storing the log. Connections between the two components occur via socket. To speed up the wizard's answer, each character in the subjects input is sent to the server side while they are still typewriting it. The agent stays in the 'idle' state on the client side until it receives an APML file to pronounce: in this state, expressions are generated randomly by the Haptek player. When the dialog is concluded, the questionnaire is displayed on the client side and, once completed, the simulation log is stored on the server side.

DISCUSSION

Several Wizard of Oz studies have been performed to create a foundation for text generation design (see, for instance, [26]). In the simulation of dialogs with ECAs, some additional problems come into play. First, the difficulty to forecast, in the study design phase, the course every subject will give to the dialog: this produces the risk that the dialog is not really natural, when goals and plans pre-defined for the agent do not correspond to

the information needs of the subject. To overcome at least in part this limit, we see several alternatives:

- to simulate only system-driven dialogs, with a limited freedom of subjects in driving the dialog towards a desired direction;
- to simulate mixed initiative dialogs, by pre-defining a large number of plans and moves for the agent, so as to insure a wide variability in the agent's behavior.

The first method may be applied to compare alternative project designs in general terms. However, it does not produce a rich corpus which may be employed for quantifying the prevalence of linguistic phenomena in the system users. With the second method, an optimal equilibrium should be found between - on one hand - the number of options available for the agent moves (and therefore the wizard's choice) and - on the other hand – the response time of the wizard and the credibility of the simulation. In fact, as far as the range of agent's moves increases, the wizard will find more difficulty in warranting consistency of her behaviour across the simulations performed.

A final consideration about the *envelope and emotional feedback* [1]. In our simulations the characters make idle animations while the subjects compile their moves. Implementing intra-move behaviours which respond immediately to what the subject is writing (like quick nods of the head, glances towards or away from the subjects, immediate emotional reaction to what they are saying or short verbal expression of 'involvement') would require a real time interpretation of the subject move which is out of the scope of a reasonably simple simulation. As a matter of fact, interpretation of the user moves is a research issue – and a potential subject of simulation studies – in itself. Our agent may only show some form of participative reaction to the subject's move *during its turn*. However, as the WoZ tool does not include the 'emotion modelling' module of our dialog simulator [13], the emotions the agent will show will not be necessarily and perfectly linked to the previous subject move. We must acknowledge that this is a limit of our simulations. In spite of these limits, we claim that our tool provides the opportunity to shift the focus of evaluation from the ECAs to the users, by analyzing with a subtle grain size how their reaction to the agent depends on the agent characteristics, on the application domain and on the context in which interaction occurs.

REFERENCES

1. Ahrenberg, L., Dahlback, N., and Joensson. A. Coding Schemes for Studies of Natural Language Dialogue. *AAAI Spring Symposium* 1995.

2. Andersson, G., Hook, K., Mourao, D., Paiva, A. and Costa, M. Using a Wizard of Oz study to inform the design of SenToy. *Proceedings of the Conference on Designing Interactive Systems.* ACM Press 2002.

3. Argyle, M. and Cook, M: *Gaze and mutual gaze*. Cambridge University Press, London. 1976.

4. Bartneck, C. How convincing is Mr Data's smile: affective expressions of machines. In F.de Rosis (Ed): User Modeling and Adaptation in Affective Computing. Special Issue of *User Modeling and User-Adapted Interaction.* 2, 4, 2001.

5. Berry, D.C., Butler, L.T. and de Rosis, F. Evaluating a realistic agent in an advice-giving task. *International Journal of Human-Computer Studies*, in press.

6. Bretan, I., Ereback, A.L., MacDermid, C., and Waern, A. Simulation-Based Dialogue Design for Speech-Controlled Telephone Services. *Proceedings of CHI'95*. 1995.

7. Buisine, S. and Martin, J.C. Experimental Evaluation of Bi-Directional Multimodal Interaction with Conversational Agents. *Proceedings of Interact'03.* 2003.

8. Cassell, J., and Thorisson, K.R. The power of a nod and a glance. The envelope vs emotional feedback in Animated Conversational Agents. *Applied Artificial Intelligence,*1999.

9. Cavalluzzi, A., Carofiglio, V. and de Rosis, F. Affective Advice-Giving Dialogs. In E.André, L Dybkjaer and P Heisterkamp (Eds): *Affective Dialogue Systems.* Springer LNAI 3068. 2004.

10. Dahlback, N., Joensson, A. and Ahrenberg, L. Wizard of Oz Studies – Why And How. *Proceedings of the Int Workshop on IUI*, 1993.

11. De Carolis, B. Pelachaud, C., Poggi, I., and Steedman, M. APML, a Mark-Up Language for Believable Behavior Generation. In H Prendinger and M Ishizuka (Eds): *Life-Like Characters: Tools, Affective Functions and Applications.* Springer, 2003.

12. de Rosis, F., Pelachaud, C. and Poggi, I: Transcultural believability in embodied agents: a matter of consistent adaptation. In R.Trappl and S Pays (Eds): *Agent Culture. Designing Human-Agent Interaction in a multicultural world.* Laurence Erlbaum Ass Inc, 2004.

13. de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V. and De Carolis, B. From Greta's mind to her face: Modeling the Dynamics of Affective States in a Conversational Embodied Agent. *International Journal of Human-Computer Studies*, 59, 1-2, 2003.

14. Gruen, D., Sidner, C., Boettner, C. and Rich, C.: A Collaborative Assistant for Email. *CHI'99*.1999.

15. Haptek website: http://www.haptek.com

16. Loquendo website: http://www.loquendo.com/

17. Nass, C., Steuer, J. and Tauber, E. Computers are social actors. *Proceedings of the Conference on Human Factors in Computing Systems* (CHI'94). ACM Press, 72-78, 1994.

18. Nass, C., Isbister, K. and Lee, E.J: Truth is beauty: Research in Embodied Conversational Agents. In J Cassell, J Sullivan, S Prevost and E Churchill: *Embodied Conversational Agents.* The MIT Press, 2000.

19. Oviatt, S. and Adams, B.: Designing and Evaluating Conversational Interfaces With Animated Characters. In J Cassell, J Sullivan, S Prevost and E Churchill: *Embodied Conversational Agents.* The MIT Press, 2000.

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

21. Pelachaud, C. and Poggi, I. Subtelties of facial expressions in embodies agents. *Journal of Visualization and Computer Animation,* in press.

22. Pelachaud, C. and Bilvi, M. Computational models of believable conversational agents. In M-P Huget (Ed): *Communication in Multiagent Systems: background, current trends and future.* Springer LNCS 2650, 2003.

23. Poggi, I. and Magno-Caldognetto, E. Il parlato emotivo. Aspetti cognitivi, linguistici e fonetici. *Proceedings of the Conference "Il Parlato italiano".* D'Auria, Naples, 2003

24. Prochaska, J., Di Clemente, C. and Norcross, H. In search ofhow people change: aplications to addictive behavior. *American Psychologist,* 47, 1992.

25. Ruttkay, Z. and Pelachaud, C. (Eds). *From Brows Till trust: Evaluating Embodied Conversational Agents.* Kluwer, in press.

26. Whittaker, S. Walker, M. and Moore, J. Fish or Fowl: A Wizard of Oz evaluation of dialogue strategies in the restaurant domain. *Language Resources and Evaluation Conference.* 2002.

ACKNOWLEDGEMENTS

This work was financed, in part, by HUMAINE, the European Human-Machine Interaction Network on Emotion (EC Contract 507422). The TTS in Italian was kindly provided by Loquendo, in the scope of an agreement with our Research Group. We thank Dianne Berry (University of Reading) for formulating the 'rational' vs 'emotional' persuasion texts, Gianluigi Del Vecchio for cooperating in developing the wrapper for Haptek Agents, and Haptek Inc. for assisting us in the application of their software.

[1] *"The nonverbal (and, occasionally, verbal) behaviors that exist in face-to-face conversations… that the animated agent produces in response to the user's communicative actions".* [8].