MagiCster: Believable Agents and Dialogue

Colin Matheson (University of Edinburgh)
Catherine Pelachaud (Università degli Studi di Roma 'La Sapienza')
Fiorella de Rosis (Università degli Studi di Bari)
Thomas Rist (Deutches Forschungszentrum für Künstliche Intelligenz)

MagiCster is concerned with the development of believable conversational interface agents which make use of gaze, facial expression, gesture and body posture as well as speech in a synchronised fashion. The project is evaluating the use of conversational agents in laboratory conditions to determine which aspects are important for what types of human-computer interactions. The project also aims to develop and document the agent architecture and components to enable other research and development teams to prototype and evaluate new versions of the agent interface in new domains and for novel tasks.

1. Introduction

As people spend more and more time interacting with computers a number of questions arise about the design of communication interfaces. Speech is often seen as the most natural, but this is only part of the answer as there is evidence that it is the ability to engage in conversation, in particular face-to-face conversation, which allows for the most natural interactions. Conversation involves more than speech, and a wide variety of signals such as gaze, gesture, and body posture are used in natural interactions. Building speech-based interfaces which do not allow for these other signals will restrict the communication bandwidth and is likely to lead to interactions which break down easily and which are generally unsatisfactory for the user. The main goal of the project is thus the design of properly conversational interfaces, in particular those which make use of non-linguistic information. The MagiCster partners are the universities of Edinburgh, Bari, and Roma, the Deutches Forschungszentrum für Künstliche Intelligenz (DFKI), the Swedish Institute of Computer Science (SICS), and AvatarMe in the UK.

2. Objectives

In recent years, a number of research projects have deployed multiple characters as a new information presentation style and/or as a device to provide the user with a new interaction experience. It has been argued (Bates 1994) that it is a necessary requirement for the success of systems with embodied characters that the characters involved come across as socially believable individuals with their own distinct personalities and emotions. However, the term 'believability' is rather vague, and subjective in the sense that it is used differently across disciplines and even differently across researchers belonging to the same community. We simply assume that believability has many facets and levels, not all of which are equally relevant for a given application, and set out some specific contexts in which to investigate this concept and test our theories through the development of different prototypes.

Systems using embodied conversational characters often rely on settings in which the character addresses the user directly as if it were engaged in a face-to-face conversation with human beings. For example, a character may serve as a personal guide or assistant in information spaces, it can be a user's personal consultant or tutor, or it may represent a virtual shop assistant trying to convince a customer to buy. Such settings are appropriate for a number of applications that rely on a particular agent-user relationship. However, other situations exist in which the emulation of direct character-to-user communication is not necessarily the most effective way to present information to a user. Many TV programs demonstrate that information can be conveyed in an appealing manner by *multiple actors with complementary characters and role castings*. This presentation style is used heavily in advertisement clips and infotainment/edutainment where information presentation is combined with entertainment. Studies have shown that subjects who watch news and entertainment segments on different TV screens rate them higher in quality than news and entertainment segments shown on just one TV screen (Nass, Reeves, and Leshner 1996). Such effects may be reinforced if information were distributed onto several characters representing different specialists. The generation of performances using a computer system allows us to take into account the particular information needs and preferences of the individual user.

In this context, the main MagiCster objectives are:

Believable animation: to develop a 'realistic' agent which displays natural or believable movement. This includes modelling not only voluntary movement (such as walking) and involuntary movement (blinking, and so on), but also movements which correlate with communicative acts (limb gestures, body posture, facial expressions). Whereas research in computer animation has actively studied these different classes of human movement, it is still difficult to generate believable natural movements within the latter class. Our agent should perform conversational functions such as repair, feedback or turn-taking in a believable way, using verbal and non-verbal signals. Of particular concern here is the development of a formalism that allows us to synchronise the different information signals, rather than assume that a particular channel is always the dominant one.

Believable multiagent interaction: to develop a multiagent system in which several embodied agents with different personalities and social relationships may interact to find an agreement in some domain or in a story-telling scenario. While in the first case the user may interact with the system to set up the simulation conditions, in the second one the user is one of the actors of the conversation.

Believable user-agent conversation: to develop an agent which can deliver information and advice in a contentful and contextually appropriate way. This work builds on earlier work on the simulation of user-system dialogues, integrated with work on animation. To increase the believability of the agent, we are paying particular attention to the role of affective factors in the dialogue dynamics.

Throughout the project, various aspects of believability have been developed and instantiated in different combinations, in a series of prototypes which integrate increasingly sophisticated aspects of each partner's research: the following section outlines the focus at each stage in this series. The variety of prototypes developed allows us to evaluate the effects of different

approaches to believability in different contexts: because the technology itself is novel, making these evaluation studies requires setting a new methodology.

3. Animation

Our work on animation has centered on the Greta system (Pelachaud & Bilvi 2003), which handles both facial and full-body animation. Facial expression and speech are tightly synchronized by using the Festival speech synthesiser (Black & Taylor, 1997) to produce timing information which runs the animation. Communication is carried out via APML, an XML-based language which was designed to handle language and animation (De Carolis et al., 2002). In the facial animations, the APML tags correspond to the meaning of a given communicative function and the task is to convert the input markers into the corresponding facial signals. The following is a simple example of a 'sorry-for' affective tag which carries the relevant communicative function:

```
<affective type="sorry-for">
I'm sorry to tell you that your proposal has been rejected
</affective>
```

The meaning of the tag is looked up in a library file, and as a result it is associated with a combination of signals; including in this case, 'head aside' and 'oblique eyebrow'. The result is the expression in Figure 1.

Figure1. Greta with a 'sorry-for' expression

An interesting aspect of our research in this area concerns situations in which expressions are to be combined. In some circumstances, of course, expressions, or at least their components, are in conflict, and in these cases we attempt to resolve the conflict and display one of the relevant parameters. However, in other contexts the target expression is a combination of two or more components; for instance, we assume that in some circumstances it is necessary to combine an

expression of condolence (as in Figure 1) with the 'certain' expression. Existing approaches to this problem typically use additive rules, simply adding the co-occurring communicative functions, or display the signal which is assumed to have the highest priority. In our case we employ belief networks, as elsewhere in the project (see the full discussion of Avatar Arena in Rist & Schmitt 2002, and also Pelachaud & Poggi 2002), and as a result we can produce relatively natural mixed expressions, as exemplified in Figure 2.

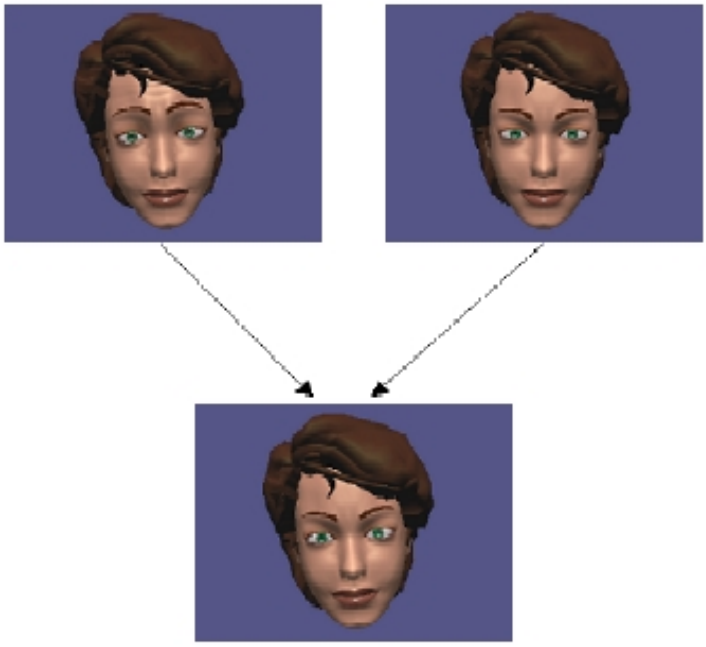

Figure 2. Combination of 'sorry-for' and 'certain' expressions using belief networks

We assume that the combination of expressions in Figure 2 would be appropriate in uttering a sentence such as "I'm sorry to tell you that you have been diagnosed as suffering from a mild form of what we call angina pectoris". To represent the necessary expressive and intonational information, we encode this in APML as shown in Figure 3.

```
<performative type="inform">
<theme belief-relation="gen-spec" affect="sorry-for" certainty="certain">
I'm sorry to <emphasis x-pitchaccent="LplusHstar">tell</emphasis> you
<boundary type="LH"/> </theme> <rheme>that you have been <emphasis x-
pitchaccent="Hstar"> diagnosed </emphasis> as <emphasis x-
pitchaccent="Hstar">suffering</emphasis> from a <emphasis x-
pitchaccent="Hstar" adjectival="small">mild</emphasis> <emphasis x-
pitchaccent="Hstar">form</emphasis> of what we call <topic-comment
type="comment"><emphasis x-pitchaccent="Hstar">angina</emphasis> <emphasis
x-pitchaccent="Hstar">pectoris.</emphasis></topic-comment> <boundary
type="LL"/> </rheme>
</performative>
```

Figure 3. APML example showing intonation and expression markup.

The main things to note in Figure 3 are the theme/rheme information and the various 'emphasis' specifications. The expressive information is associated with the theme as values of the 'affect' and 'certainty' attributes, which are interpreted by the Greta player as discussed above, and the emphasis statements carry the intonation information.

The general point here is that it has been shown that the intelligibility of speech and speech perception by hearers improve when visual signals are considered as well as audio (Risberg & Lubker 1978, Schwippert & Benoit 1997). Another important MagiCster goal in this area is thus to create a natural talking face with lip-readable movements. CNR in Padova provided us with real data extracted from a speaker with an opto-electronic system that applies passive markers to the speaker's face. We first approximated this data using a neural network model and then developed a computational model of lip movements and coarticulation effects using a logistic function. Our model is based on some phonetic-phonological considerations of the parameters defining the labial orifice, and on identification tests for visual articulatory movements.

4. Prototypes

Three prototypes were planned during the life of the project. The first built on work on DFKI's *Avatar Arena* (Rist & Schmitt, 2002), in which virtual characters negotiate over meeting appointments on behalf of human users. The second prototype is a testbed for simulating emotional dialogues. The final system is looking at the application of MagiCster technology in story-telling, game-playing scenarios. As we said, each prototype is thus based on different aspects of believability; in the first, several agents communicate among themselves; in the second, a single agent communicates with the user, and in the third several agents communicate both among themselves and with the user.

Avatar Arena

Where earlier versions of DFKI's Avatar Arena used the Microsoft Agent toolkit to render the agents, the MagiCster system now uses the Greta animation system described above, to represent the different agents, as shown in Figure 4.

Figure 4. Avatar Arena: three negotiating Greta agents with a human observer.

The main point here is that in the context of Avatar Arena a special focus of the MagiCster research interest is on a simulation of the dynamics of social relationships among affective characters. Our working hypothesis is that believability in this domain can be assessed by human observers along various dimensions, including:

Domain competence; the characters need to show some understanding of the subject matter domain (here meeting appointments);
Conversational skills; the characters need to adhere to the basic rules for participation in a group discussion/negotiation dialogue.
Affective behaviour; the characters should display affect in compliance with both assigned personality traits and changes of affective states and social relationships that may occur in the process of a negotiation dialogue.

One approach for assessing believability in negotiation dialogues is to show human observers several such interactions with virtual characters that differ in the degree of how well the above-mentioned dimensions are modelled. The Avatar Arena prototype provides such a simulation framework in that, before a meeting appointment negotiation dialogue is generated, the user can chose which criteria, such as a character's social position in the group, should be taken into account in the negotiation.

As an underlying basis for modelling the dynamics of social relationships among avatars we rely on the concept of a Cognitive Configuration which can either be balanced or unbalanced and which has its roots in socio-physiological theories of cognitive consistency, in our case especially in Heider's Balance Theory (Heider 1946, 1958) and Festinger's theory of cognitive dissonance (Festinger 1957). Heider's Balance Theory starts from the hypothesis that a good deal of interpersonal behaviour and social preception is determined by simple cognitive

configurations which are either balanced or unbalanced. Together with the hypothesis that people tend to avoid unbalanced configurations or cognitive dissonances, the theory makes predictions about how a certain person might behave in certain social situations. In Festinger's approach imbalanced cognitive configurations are viewed as cognitive inconsistencies and a general tendency for individuals to seek consistency among their beliefs and opinions is assumed. In the case of an inconsistency between attitudes or behaviours, there is a tendency to eliminate the dissonance. In particular, a discrepancy between an attitude and behaviour may cause a change in attitude so that the attitude accommodates the behaviour.

While Heider's and Festinger's work forms the foundations of our approach to modelling changes in interpersonal relationships during a negotiation process, we need to combine concepts such as balance and dissonance with a model of interpersonal communication. Fortunately, moves in this direction have already been made by some psychologists, such as Newcomb (1953), who considers the case where a person performs a communicative act to give another person information about some particular subject matter. The Congruity Theory of Osgood and Tannenbaum (1955) has also shaped our research in this area as it addresses attitude changes brought about by means of communication, and augments the original balance concept in that it considers not only polarisation but also the intensity of relations. Generally, Congruity Theory makes an assertion about the impact of a message on a hearer, and we have attempted to model some aspects of this in the prototype; to illustrate this, Figure 5 shows how the 'liking' relationship between characters was set up in a version of Avatar Arena using Microsoft Agents.

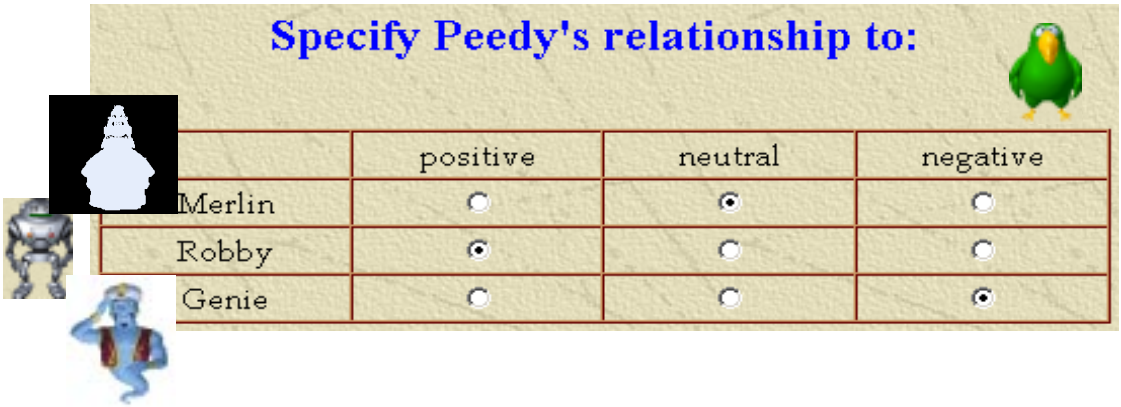

Figure 5. Setting the 'liking' relationships between the Peedy character and three other avatars.

In early versions of Avatar Arena the affective states of the agents were not modelled. In the MagiCster context, however, we started to integrate the dynamic belief networks developed by the University of Bari for modelling affective states and emotion triggering, and this feature is an important focus of our research, being the central aspect of the second prototype.

Emotional Dialogue Testbed

This work is based on research at the University of Bari on modelling context and personality-dependent activation of emotions. The ability to simulate feeling of emotions and to adapt

behaviour accordingly is quite universally recognized as one aspect of believability. While the affective component is particularly important in some application domains, like tutoring or entertainment, there is practically no application in which at least some elementary form of affect can be excluded (Paiva, 2000). Affect is a generic term that includes various aspects of 'extra-rational' mind and reasoning in humans: personality (a combination of various traits) refers to long-term features, while emotions refer to short-term ones; the notion of 'attitude' is more blurred and controversial (see Lisetti, 2002 for a detailed discussion of these terms). The three parameters are closely interrelated: some personality traits influence the level of emotional reaction to a given event. Others affect the consequent behaviour, at a shallow or an inner level: the tendency to show or to hide an emotion, decision process, or conversational attitude, to mention a few (Picard, 2000).

In our affective dialogue testbed, emotion triggering is the result of a combination of several factors, including the physical context, the agent's individual characteristics (such as personality and temperament), and the social context in which the event occurs. One cause alone does not, usually, activate an emotion, and it is quite widely accepted that context also plays a critical role in the actual expression of the emotion.

We model the cognitive aspects of emotion activation: how the agent interprets a user move in terms of a consistent combination of beliefs and goals and which is the combination of beliefs and goals that triggers multiple emotions, with varying intensities. The formalism we employ in representing this process is that of dynamic belief networks. By applying the theory developed by Oatley and Johnson-Laird (1987), we model the activation of emotions produced by the belief that a high-level goal will be achieved or threatened. 'Personality traits' in the agent correspond to the assignation of weights to these high-level goals (as suggested by Carbonell; see for example Carbonell 1982), while 'temperament' is modelled as influencing the 'threshold' below which emotions are not triggered, and the time decay of emotions. The social context also affects emotion intensity, especially as far as 'fortune-of-others' emotions (Ortony et al, 1988) are concerned. Event-based emotions (ibid) have also been modeled; for more details on the emotion simulation method, see (Carofiglio et al, in press).

Affective factors influence the character behaviour in several ways:

- at a *shallow* level, they are displayed in the character's face. Fine-grained models of emotion expression are based on the idea that there are many more facial expressions than those reflecting the few 'basic' ones (such as joy, fear, and anger). These expressions are the result of a large number of 'appraisal dimensions', which combine dynamically to produce cumulative changes in the face. The consequence is that prototypical expressions occur very infrequently in spontaneous interactions: natural expressions depend on the situational context and the inter-individual variability is very high. On the other hand, a given facial expression may convey several meanings at the same time: smiles and frowns may be employed as speech regulation signals, speech-related signals, means for signalling relationship, indicators for cognitive processes, and so on. Our agent combines individual and multiple emotion expressions with the display of other 'meanings' (performatives, rhetorical relations, turn-taking and so on);
- At a deeper *inner* level, affect influences the dialogue dynamics. The agent has several goals it must try to achieve during the dialogue: the priorities of these goals are a function of

its personality, of the presumed characteristics of the user (such as age and cooperative attitude) and of the social context in which the conversation occurs (friendly or unfriendly, and suchlike). When the agent 'feels' an emotion as a result of a user move, the goal priorities are revised: for instance, if the emotion belongs to the 'fortune of others' category, the agents activates a plan which aims to show its 'participation' in the user's problems. Other emotions activate a different behaviour strategy, according to the positive or the negative effect that showing them may play on the main agent goal (to establish a cooperative and trusting relationship with the user).

The domain-independent testbed integrates the various components of the dialogue simulator and enables us to test its behaviour in several situations. Emotion triggering is modelled by the Mind module (de Rosis et al, 2003); the dialogue manager is implemented with the Trindikit (Larsson et al, 2000) and the system is coupled to Greta via an APML file. A graphical interface enables the selection of an application domain and the setting of the simulation conditions: the user may follow the course of the dialogue both in natural language and with the animated agent. The user moves are recognized and interpreted by Mind in terms of the affective reaction they activate in the agent. The variation of the emotional state of the agent is displayed, in graphical form, in the interface. Figure 6 shows this interface with an example of dialogue in the domain of advice about eating disorders.

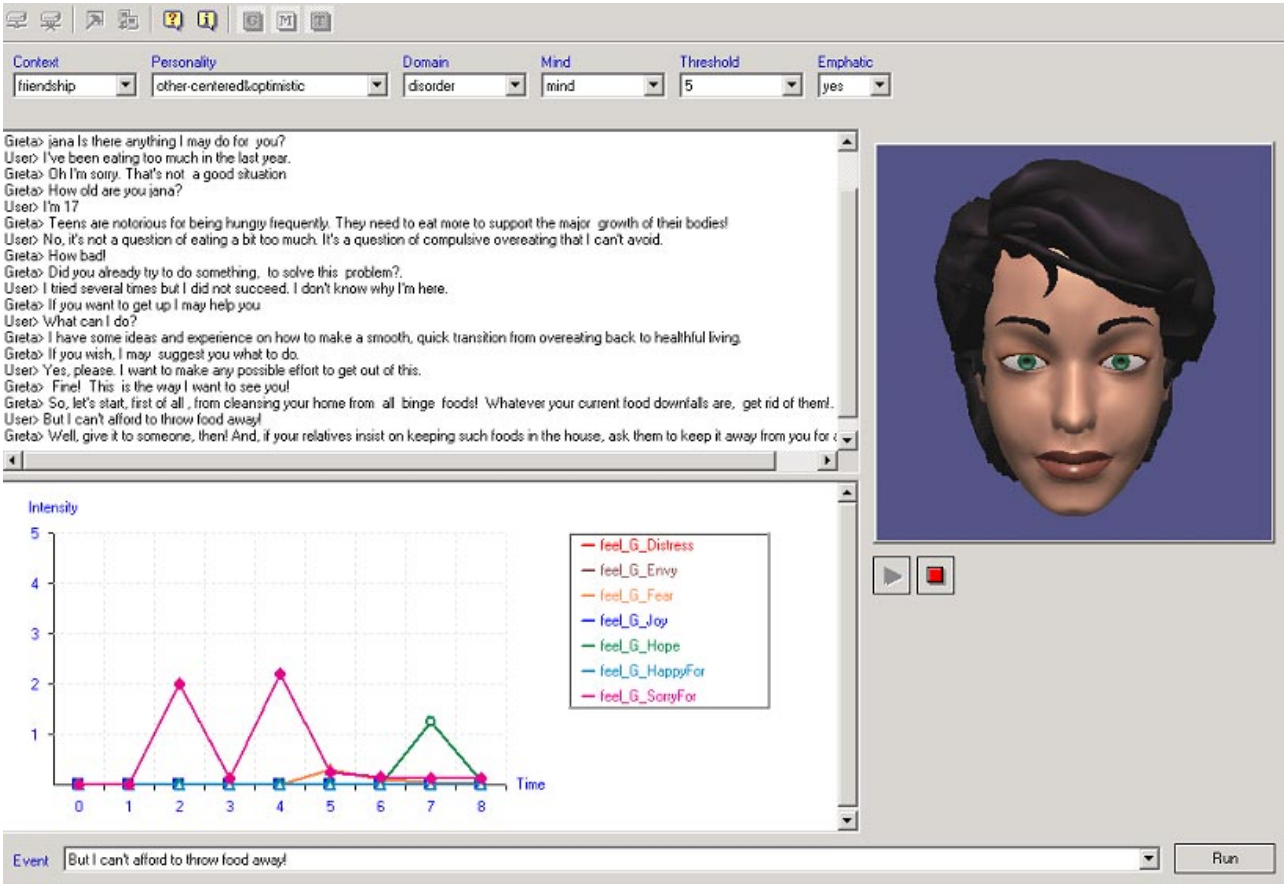

Fig 6. Graphical Interface of Prototype 2 (the Affective Dialogue Simulator)

Story Telling

The final prototype is working on adapting aspects of all the MagiCster research for use in a story telling scenario in which, as mentioned above, agents interact both with each other and with the user. We worked initially on generalising the concept of a Dialogue Manager to encompass not only dialogue actions, but also other types of actions. Having an action manager allows agents to choose which communication channel to use in various contexts, and we believe that this is important both for expressing personality and mood as well as for achieving believability. For instance, a character who is angry may be less inclined to communicate verbally and may instead opt to use only facial expressions or gestures. In addition, some personality types, for instance shy persons, may be less talkative than others but may nevertheless have a wide communicatory repertoire using other channels.

We thus adopt an agent-based approach to story telling where characters in the guise of artificial agents and users are the primary vehicles for conveying dramatic action and progress along different story arcs. For the final prototype we will continue to enhance the Greta player, for example by investigating whether it is possible for several agents to share a sound device, thus making it possible to blend the speech output of multiple agents making squabbles, interruptions, rude behaviour and other dramatically interesting interactions possible. Also, inspired by the new believability factors we will explore an interface based on a cartoon metaphor. While agents will be animated using existing photorealistic technology, the presentation will be divided into dynamically changing frames, illustrating the linear progression of the story line. The size and style and dynamicity of the frames will be used to enhance and clarify the agents' emotions and dramatic turns in the story, and we aim to integrate the Bari emotional model and the DFKI social model.

5. Ongoing Work

In the remaining part of the Project, we plan to complete, refine and evaluate the three Prototypes in several ways:

- The University of Edinburgh's DIPPER dialogue system architecture (Bos et al, 2003) is being tested as the hub to make all the prototypes more flexible and open to integration of new components;
- The role of affective factors in enhancing the strength of persuasion (in Prototype 2) through user-tailored appeal to emotions is being studied;
- In the final version of the conversational system the user will interact with Greta by means of a portable device.

Finally, a set of evaluation studies are being performed in cooperation with the University of Reading in the domain of advice about healthy eating. These studies have so far involved more

than 200 subjects: they are aimed at a comparative assessment of the effects of the Greta system in terms of the user's cognitive performance (memory and persuasion) and in terms of subjective evaluation of the agent's message (sincerity, credibility, easiness to follow, and so on) and quality (intelligence, likability, reliability, etc). Several versions of Greta are being used and compared with text and human videos: special attention is paid to the concept of consistency in the expression of the emotional state and how this affects the performance measures being considered.

6. References

Bates, J. (1994). The role of emotion in believable agents. Communications of the ACM, 37, 7, 1994.

Black, A. & Taylor, P. (1997). Festival Speech Synthesis System: system documentation (1.1.1). Human Communication Research Centre Technical Report HCRC/TR-83.

Bos, J., Klein, E., Lemon, O., & Oka , T. (2003). DIPPER: Description and formalisation of an information-state update dialogue system architecture. Submitted to 4th SIGdial Workshop on Discourse and Dialogue, 5-6 July 2003, Sapporo, Japan.

Carbonell J. G. (1982). Where do goals come from? In proceedings of the 4th Annual conference of the Cognitive Science Society, pages191-194. Ann Arbor, Michigan.

Carofiglio, V., de Rosis, F. & Grassano, V. (2002). Dynamic models of mixed emotion activation. In D. Canamero and R Aylett (Eds), Animating expressive characters for social interactions. John Benjamins, in press.

De Carolis, B., Carofiglio, V., Bilvi, M., and Pelachaud, C. (2002). APML, a mark-up language for believable behaviour generation. Workshop on Embodied conversational agents - let's specify and evaluate them! In conjunction with AAMAS 02, Bologna, Italy.

de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., & De Carolis, B. (2003). From Greta's mind to her face: Modelling the dynamics of affective states in a conversational embodied agent. International Journal of Human-Computer Studies. Special Issue on Applications of Affective Computing in HCI. E. Hudlicka and M. McNeese (Eds). 59, 81-118.

de Rosis, F., De Carolis, B., Carofiglio, V., and Pizzutilo, S. (in press). Shallow and inner forms of emotional intelligence in advisory dialog simulation. To appear in H Prendinger and M Ishizuka (Eds): Life-like Characters. Tools, Affective Functions and Applications", Springer.

Festinger, L. (1957). A Theory of Cognitive Dissonance. Evanston, Illinois: Row Peterson & Co.

Heider, F. (1946). Attitudes and Cognitive Organization. Journal of Psychology, 21, pages 107-12.

Heider, Fritz. (1958). The Psychology of Interpersonal Relations. NY: Wiley. Chapter 7: Sentiment, pages174-217.

Lisetti, C. L. (2002). Personality, Affect and Emotion Taxonomy for Socially Intelligent Agents. FLAAIRS Conference 2002, pages 97-401.

Nass, C., Reeves, B., & Leshner, G. (1996). Technology and roles: A tale of two TVs. Journal of Communication, 46, pages 121-127.

Newcomb, T. M. (1953). An approach to the study of communicative acts, Psychological Review, 60, pages 393-404.

Oatley, K. J., & Johnson-Laird, P.N. (1987). Towards a cognitive theory of emotions. Emotion and Cognition, 1, pages 29-50.

Ortony, A., Clore, G. L., & Collins, A. (1988). The cognitive structure of emotions. Cambridge University Press.

Osgood, C. E. & Tannenbaum, P. H. (1955). The principle of congruity in the prediction of attitude change, Psychological Review, 62, pages 42-55.

Paiva, A. (2000). Affective Interactions: toward a new generation of computer interfaces, LNAI- State-of-the Art Surveys, LNAI 1814, Springer.

Pelachaud, C. & Bilvi, M. (2003). Computational Model of Believable Conversational Agents, in Communication in MAS: background, current trends and future, Marc-Philippe Huget (Ed), Springer-Verlag, 2003.

Pelachaud, C. & Poggi, I. (2002). Subtleties of facial expressions in embodied agents, Journal of Visualization and Computer Animation.

Picard, R. W. (2000). Towards computers that recognize and respond to user emotion? IBM Systems Journal, Volume 39, Nos. 3 & 4.

Risberg, A. & Lubker, J. L. (1978). Prosody and speechreading. Quarterly Progress and Status Report 4, Speech Transmission Laboratory, KTH, Stockholm, Sweden, 1978.

Rist, T. & Schmitt, M. (2002). Avatar Arena: An attempt to apply socio-physiological concepts of cognitive consistency in avatar-avatar negotiation scenarios. In Proceedings of AISB Symposium, London, April 2002.

Schwippert, C., & Benoit, C. (1997). Audiovisual intellegibility of an androgynous speaker. In C. Benoit and R. Campbell (Eds), Proceedings of the ESCA Workshop on Audio-Visual Speech Processing, Rhodes, Greece, September 1997.

Larsson, S., Berman, A., Bos, J., Grönqvist, L., Ljunglöf, P., & Traum. D. (2000). TRINDIKIT 2.0 Manual, TRINDI Deliverable 5.3.