Interactive Information Presentation by an Embodied Animated Agent

B. De Carolis, F. de Rosis and V. Carofiglio
Dipartimento di Informatica,
Università di Bari
{decarolis, derosis, carofiglio}
@di.uniba.it

C. Pelachaud
Dipartimento di Informatica e Sistemistica,
Università di Roma, "La Sapienza
cath@dis.uniroma1.it

I. Poggi
Dipartimento di Educazione
Università di Roma Tre
poggi@uniroma3.it

1 Abstract

This paper illustrates the architecture of a multimodal believable agent, provided with a personality and a social role, aiming at providing information to users engaging them in a natural conversation. To achieve this aim, we provide our agent with a mind, a dialogue manager and a body: a) the mind, according to the agent's personality, the events occuring and the user dialog move, triggers, if appropriate, an emotion; b) the dialog manager, according to an overall dialog goal and the corresponding plan to be pursued, selects the appropriate dialog move to be performed by the agent; c) the body is a combination of a 3D face model and a speech synthesizer. The facial model is capable of expressing the nonverbal communicative functions foreseen for our conversational agent.

Keywords

Conversation agents, believable agent architectures, personality and emotions in agents.

2 Introduction

Natural conversation involves more than speech. Humans communicate using language and a lot of other signals, in combination with speech: body posture, gestures (pointing at something, describing object dimensions, …), facial expression, gaze (making eye contact, looking down, up, to a particular object,…) and so on [2, 3, 13]. While communicating, people exhibit a behaviour that is consistent with their personality, their goals, their affective state and the context in which the conversation takes place.

In the context of the EU project MagiCster[1], we developed the first prototype of an Embodied Agent that combines appropriately verbal and nonverbal signals when delivering information, to establish a natural communication with the User. Our agent shows a rich expressiveness during the conversation, by showing the communicative functions that are typically used in human-human dialogs; for instance: syntactic, dialogic, meta-cognitive, performative, deictic, adjectival and belief relation functions [12].

However, simply providing a 2D or 3D character, (being a full-body virtual human or a face with multimodal communicative functions) is not enough for achieving a believable behaviour. The definitions of believability that have been proposed involve several dimensions: personality, affect, social intelligence and, in particular, consistent behavior [1, 9, 10]. Our Agent, that is named *Greta*, is embodied in a 3D talking head that shows a personality, social intelligence and, in addition, has the capability of reacting emotionally to events occurring in the environment, consistently with the context in which the conversation takes place and with its goals. This paper aims at giving an overview of the current state of the system development by describing its architecture and by showing some examples of dialog between the agent and the user. Although the whole system is domain-independent, we will show an example in the medical domain.

3 The agent's architecture

The type of conversations we simulate at present are *information-giving* dialogs, in which the main function of Greta is to provide some kind of information to the User, in a given domain. MagiCster is a 'mixed-initiative' system: the User can therefore ask questions after Greta's 'giving turns'; this opens a *question-answering subdialog*, after which Greta revises, if needed, her discourse plan according to the User's request.

The architecture of Greta, shown in **Figure 1,** includes the following main components: a manager of the Agent's *Mind*, a *Dialog Manager*, an *Enricher* of the dialog move (Midas) and a generator of the Agent's *Body*.

When the dialog starts, a dialog goal in a particular domain is set and passed to the Dialog Manager (DM). From this goal, an overall discourse plan is produced for the Agent. This is done by retrieving an appropriate 'recipe' from a plan library; this plan represents the way in which the agent will try to achieve the specified communicative goal during the conversation. Indeed, the generation of the Agent's reply (including verbal and nonverbal behavior) to a given user request depend also on the social context of the conversation [5]: when talking with somebody we consider our discussion partner (what is our relation with her, what

[1] IST project IST-1999-29078, partners: University of Edinburgh, Division of Informatics; DFKI, Intelligent User Interfaces Department; Swedish Institute of Computer Science; University of Bari, Dipartimento di Informatica; University of Rome, Dipartimento di Informatica e Sistemistica; AvartarMe, UK.

are her intellectual capacities an so on), the environment in which the conversation takes place, as well as how are related to a particular event or object that we may be referred in the conversation. The Social Context describes the Agent's personality as well as the role and the relationship existing between the user and the agent.

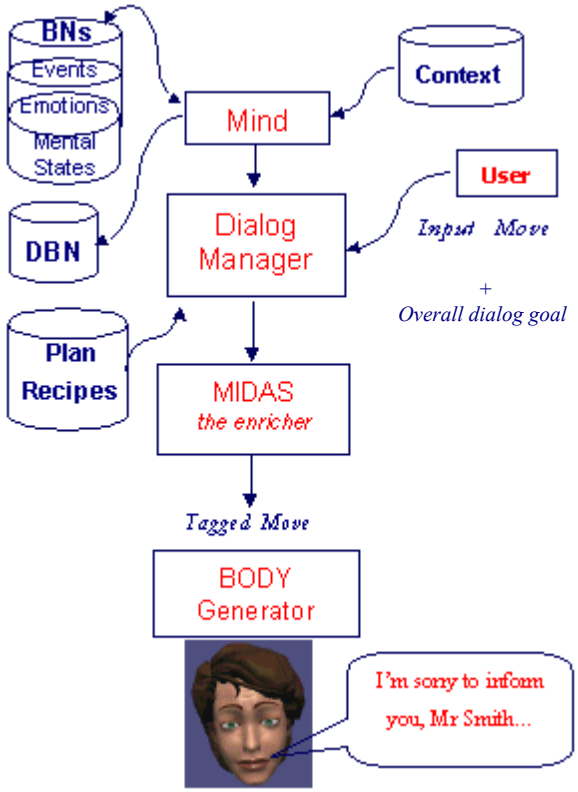

Figure 1. The Architecture of our Conversational Agent.

Let us now describe the architecture components of our Agent in more detail:

a. *Mind* is responsible for updating the Agent's mental state, by deciding whether a particular affective state of the Agent should be activated and with which intensity; it decides, as well, whether the felt emotion should be displayed and how, according to the contextual variables. Mind is based on a dynamic belief network (DBN), that combines a network representing the agent's mental state at time T with a network representing its mental state at time T+1 and a network that monitors the triggering of emotions in the interval ([T, T+1]). Three kinds of nodes can be found in the Agent's mental state: 'belief' nodes, 'goal' nodes and 'goal-achievement' nodes. A weight is associated with goal-achievement nodes, as a function of the agent's personality. The Belief Network (BN) at time (T+1) is generated according to the network at time (T) and to the events occurred in the interval [T,T+1]. These events are modeled, as well, by belief networks.

b. the *Dialogue manager* (DM) is built on the top of the TRINDI toolkit [14], which is an engine for computing dialog moves. Our DM controls the dialog flow by iterating the following steps:

- after an 'overall dialog goal' has been specified, an appropriate discourse plan is selected then by the library of plan recipes and the first move is generated according to the first step of the plan; this 'overall dialog goal' becomes the main topic of the conversation;
- at the end of this first move, the initiative is passed to the User, that can make questions to the agent on any subjects among the main topics under discussion;
- the User move is translated into a symbolic communicative act (through a simplified interpretation process) and is passed to the DM;
- the DM decides "what to say next" by selecting the sub-plan to execute.
- the DM goes on by cycling on these three last steps, until the user leaves the conversation.

In both cases, for the first dialog move and for the subsequent ones, a symbolic representation of the selected dialogue act is passed to Mind, that updates Greta's mental state and possibly returns the name of an 'affect' that should be associated with the communicative act.

c. The Midas module has the role of translating the symbolic representation of the dialog move into an Agent's behaviour specification. In order to overcome integration problems between the mind and body components and to allow their independence and modularity, we based the specification language of Midas on XML and followed the directions of the Human Markup Language initiative (HML, [8]) The aim of HML is to "develop Internet tools and repository systems which will enhance the fidelity of human communications". Our Affective Presentation Markup Language (APML, see Figure 2 for its DTD specification) specifies how to markup the verbal part of a dialog move so as to add to it the 'meanings' that the graphical and the speech generation components of Greta need, to produce the required expressions. So far, we only defined the meanings that may be translated into face expressions: rhetorical relations between communicative acts, deictic [9] or adjectival components, certainty values, metacognitive or turn-taking expressions. In the future, we plan to extend our language, to enable representing meanings that may be expressed with other nonverbal signals (for instance: gestures).

APML – Affective Presentation Markup Language

```
<!ELEMENT APML (turn-allocation+, perfomative*, meta-cognitive*)>
<!ENTITY %TA-TYPE "(take |give)">
<!ENTITY %PA-TYPE "(inform|ask|greet|request|…)">
<!ENTITY %affect "(joy|sorry-for|distress|…)">
…
<!ELEMENT turn-allocation (performative+)>
<!ATTLIST turn-allocation %TA-TYPE #REQUIRED>
<!ELEMENT adjectival (#PCDATA)><!ATTLIST adjectival %ADJ-TYPE
#REQUIRED>
<!ELEMENT belief-relation (performative+)><!ATTLIST belief-relation
%BR-TYPE #REQUIRED>
<!ELEMENT performative (belief-relation*, adjectival*, deicitic*)>
<!ATTLIST performative %P-TYPE #REQUIRED %affect #IMPLIED
%certainty #IMPLIED>
<!ELEMENT meta-cognitive(#PCDATA)>
<!ATTLIST meta-cognitive %MC-TYPE #REQUIRED>
<!ELEMENT deictic (#PCDATA)><!ATTLIST deictic obj CDATA
#REQUIRED>
…
```

Figure 2 . The APML DTD.

Compared to HML, APML can be seen as a finer-grained language whose schema definition may refer to HML types such as intentions, emotions and so on.

The input of Greta's Body is a 'dialog move', represented as an APML string. This string is automatically generated, from MIDAS, from the output of the 'discourse planner', that is from a 'discourse plan'. This plan may include just one 'primitive' communicative act (for instance: a greet, a thanks, an inform, a request) or may be more complex (for instance: 'Describe an object with its properties'). In both cases, it is represented as an XML-tree structure, according to a "Discourse Plan Markup Language" (DPML) whose DTD is shown in **Figure 3** [4].

```
<!ELEMENT d-plan (node+)>
<!ATTLIST d-plan  name CDATA #REQUIRED>
<!ELEMENT node (node*, info*)>
<!ATTLIST node
   name CDATA #REQUIRED
   goal CDATA #REQUIRED
   role (root | nucleus | sat) #REQUIRED
   RR CDATA #IMPLIED
>
<!ELEMENT info EMPTY>
<!ATTLIST info
   focus CDATA #REQUIRED
   compl (H | M | L) #REQUIRED
>
```

Figure 3 . The DPML DTD.

The algorithm applied by Midas translates this DPML-based tree-structure into a APML-based structure, through a set of transformation rules that depend on the information attached to nodes in the discourse plan: rhetorical relation name and type, communicative goal, discourse focus and so on. When Mind establishes that an emotion is felt by the Agent in correspondence with the whole dialog move or with part of it and that this emotion has to be displayed, the corresponding text is annotated by Midas with an appropriate tag. The result of this transformation is then represented as a valid APML string.

For example, let's suppose that, after a user request of getting information about a problem (his disease) and its properties (severity), the DM decides to answer by selecting the following complex dialog act (subplan expressed according to DPML):

```
<node      name="n1"      goal="Describe(has(U,
angina))"            role="sat"focus="angina"
RR="ElabObjAttr">
  <node      name="n2"      goal="Inform(has(U,
  angina))"                  role="nucleus"
  focus="?has(angina)" RR="null"/>
  <node                              name="n3"
  goal="Inform(severity(angina))"   role="sat"
  focus="?severity(angina)" RR="null"/>
</node>
```

Every dialog turn of the Agent starts with the turn-allocation function that indicates that the agent has the initiative. Therefore, after the root tag <APML>, the <turn-allocation> container tag is generated by setting up its type attribute equal to "take". Each RR attribute in a node is transformed into a <belief-relation> whose type attribute is set with the name of the RR; each leave node is transformed into a <perfomative> element of APML. At this point, if Mind establishes that an emotion is felt by the Agent in correspondence with the current node and that this emotion has to be displayed, the affect attribute of the perfomative tag is set to be equal to that emotion.

The surface realization of the leave node, corresponding to the text within the <performative> tags, is made by a generation function that, besides producing the verbal part of the speech act, includes, if needed, two other types of tags: the <adjectival> one, when the argument of the current communicative goal is a quantitative attribute of the current discourse focus; the <deictic> one when the argument of the current communicative goal is described in the domain knowledge base as 'referrable through its coordinates'. In this case "severity" is a quantitative property of angina, which is the discourse focus, and then the <adjectival> tag is generated around the attribute-word.

The following APML string shows what is generated for the previously shown subplan:

```
<APML><turn-allocation type=”take”>
<performative type="inform" affect="sorry-
for" certainty=”certain”> I'm sorry to tell
you that you have been diagnosed as suffering
from what we call angina
pectoris,</performative><belief-relation
type=”eoa”> which <performative type="inform"
certainty=”uncertain”> appears to be
<adjectival type="small">mild. </adjectival>
</performative> </belief-relation></turn-
allocation></APML>
```

d. The *Body Generator* module interprets the APML-tagged dialog move and decides which signal to convey on which channel for each communicative act. The Body we use is a combination of a 3D face model compliant with the MPEG-4 standard [11] and a speech synthetiser [7]. The facial model is capable of expressing the nonverbal communicative functions foreseen for our conversational agent.

To connect the various components (Mind, MIDAs, Greta, and the DM itself), we use a Java class (**jcontroller)**, which controls activation, termination and information exchange for the various processes involved in the dialogue management, via socket.

4 An Example

To give an idea of how the system works, we show a simple example of the output; in this dialog, Greta represents a doctor who explains a drug prescription to a patient (the User).

After an initial ‘greeting’, the Agent first of all informs the patient about his disease. Then, the patient can make any question concerning the topic in focus of the conversation; for instance, he may ask clarifications about the disease, its seriousness, the related therapy and so on; the system will answer in a way that is appropriate to the context. Obviously, different plan steps will be selected to respond to the User question. **Figure 4** shows a few annotated moves of the example dialogue. The tag names refer to a given communicative function whose values are specified between quotes.

The Agent’s dialog turn starts, as we mentioned before, with the turn-allocation function that indicates that the agent has the initiative. In the first dialog move (S0), Greta shows her empathy to the user: while informing the patient about his problem, she will show she is sorry for his illness. The affective communicative act is specified by the value ‘sorry-for’ of the corresponding attribute of the performative tag. This affective function is not displayed, for instance, in the case of dialog with another doctor: just information delivering is a more appropriate task in this case. In the same dialog move, we may notice that, since the doctor is certain of her diagnosis, she informs the patient in an assertive manner, verbally and not verbally. The other tags that are of relevance in this dialog are the <adjectival> one, related to quantitative attributes of an object or a concept, the <deictic> one, used to refer to a point in space, and the <belief-relation> one, related to rhetorical relations present in complex dialog acts.

U0: Selection of the overall dialog goal
“Explain(S,U, therapy(angina))”

S0:
```
<APML><turn-allocation type=”take”>
<performative type="inform" affect="sorry-
for" certainty=”certain”> I'm sorry to tell
you that you have been diagnosed as suffering
from what we call angina
pectoris,</performative><belief-relation
type=”eoa”> which <performative type="inform"
certainty=”uncertain”> appears to be
<adjectival type="small">mild. </adjectival>
</performative> </belief-relation></turn-
allocation></APML>
```

U1: ***What is angina pectoris?***

S1:
```
<APML><turn-allocation type=”take”>
<performative type="inform"
certainty=”certain”><belief-relation
type="gen-spec">This is </belief-relation> a
spasm of <deictic obj=”chest”>
chest</deictic></perfomative>
<belief-relation type="cause-effect">
resulting from <performative type="inform"
certainty=”certain”>overexertion when heart
is diseased.</performative></belief-
relation></turn-allocation></APML>
```

U2: ***Which is the therapy?***

S2:
```
<performative type="suggest"> <belief-
relation type="cause-effect"> To solve
<deictic obj=”patient”>your</deictic>
problem, </belief-relation> you should take
two drugs.</performative>
<performative type="inform"> <topic-comment
type="comment"> The first one </topic-
comment> is Aspirin <belief-relation
type="sequence">and</belief-relation>
<topic-comment type="comment"> the second one
</topic-comment> is Atenolol </performative>
```

… and so on.

Figure 4 . An example of annotated dialog.

Annotations are translated into facial expressions. **Figure 5** shows some examples of these expressions. As we said before our facial model is compatible with MPEG-4 compliances. The text of each dialog move with its tags is given as input to the Greta module and to Festival which provides the duration of phonemes and a wav file (an audio file). Phonemes are the smallest temporal units we are considering. Knowing the phoneme duration enables us to retrieve the exact duration of any expression as defined by the tags in the dialogue move, thus ensuring the synchrony between speech and other visual activities.

Figure 5. Some expressions of Greta's affective state: 'neutral', 'sorry-for' and 'surprise'.

5 Conclusions

In this paper, we have briefly described the architecture of a conversational agent embodied in a 3D face (Greta) that tries to achieve a believable behavior while interacting with the user. The type of conversation that Greta is able to undertake with the user is a query/answer dialog for information-giving applications. However, the architecture we presented can be applied in other domains and, with small changes in the DM engine, for other types of dialogs. In the future, we plan to look at how to link the dialogue manager to a high-level dialogue planner in order to generate plans that takes into account changes in the interaction context.

REFERENCES

1. J M Allbeck and N I Badler: Consistent communication with control. In C Pelachaud and I Poggi (Eds): Proceedings of the Workshop on "Representing, annotating and evaluating non-verbal and verbal communicative acts to achieve contextual embodied agents". Autonomous Agents 2001.
2. E. André, T. Rist, S. Van Mulken, M. Klesen and S. Baldes: The automated design of believable dialogues for animated presentation teams. In S. Prevost, J. Cassel, J.Sullivan and E.Churchill (eds), Embodied Conversational Characters. MIT Press, 2000.
3. J. Cassell, J. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjalmsson, and H. Yan. Embodiment in conversational interfaces: Rea. In Proceedings of CHI'99, pages 520--527, Pittsburgh, PA, 1999.
4. MT Cassotta, B. De Carolis, F. de Rosis, C Andreoli, ML De Cicco. User-Adapted Image Descriptions from Annotated Knowledge Sources. In AI*IA 2001: Advances in Artificial Intelligence, 7th Congress of the Italian Association of Artificial Intelligence, Bari , September 2001.
5. B. De Carolis, F de Rosis, C. Pelachaud and I. Poggi. Verbal and nonverbal discourse planning. In Proceedings of IJCAI 2001 Seattle, August 2001.
6. C. Elliott. An Affective Reasoner: A process model of emotions in a multiagent system, Technical Report No. 32 of The Institute for the Learning Sciences Northwestern University,1992.
7. FESTIVAL Home Page: http://www.cstr.ed.ac.uk/projects/festival/
8. Human Markup Language Home Page: http://humanmarkup.org
9. J.C. Lester, S.G. Stuart, C.B. Callaway, J.L. Voerman, and P.J.Fitzgerald. Deictic and emotive communication in animated pedagogical agents. In S. Prevost J. Cassell, J. Sullivan and E. Churchill, editors, Embodied Conversational Characters. MITpress, Cambridge, MA, 2000.
10. A. Ortony, G.L. Clore and A. Collins. The Cognitive Structure of Emotions. Cambridge University Press, 1988.
11. C. Pelachaud, E. Magno-Caldognetto, C. Zmarich, and P.cosi. An Approach to an italian talking head. Eurospeech '01, Aalborg, Denmark September 3-7 2001.
12. I. Poggi, C. Pelachaud, and F. de Rosis. Eye communication in a conversational 3D synthetic agent. Special Issue on Behavior Planning for Life-Like Characters and Avatars of AI Communications. 2000.
13. J. Rickel and W.L. Johnson. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. Applied Artificial Intelligence, 13:343--382, 1999.
14. TRINDI Home Page: http://www.ling.gu.se/projekt/trindi/