Multimodal Intelligent Information Presentation

Text, Speech and Language Technology

VOLUME 27

The titles published in this series are listed at the end of this volume

Multimodal Intelligent Information Presentation

Edited by

Oliviero Stock
ITC-irst, Trento, Italy

and

Massimo Zancanaro
ITC-irst, Trento, Italy

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 1-4020-3050-9 (PB)
ISBN 1-4020-3049-5 (HB)
ISBN 1-4020-3051-7 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Sold and distributed in North, Central and South America
by Springer,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Springer,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

Printed in the Netherlands

TABLE OF CONTENTS

INTRODUCTION

Information Technology often moves slowly and the potential of innovative concepts requires time to be fully understood. We know what happened with the history of the computer and its directions of development in general. Though the achievements of hardware technology were amazing, software development methodology lagged behind and did not consent full exploitation of the hardware. Even more strikingly, it took over three decades, until in the Seventies for the computer to be recognized as an interactive tool aimed not only at automatic computation, but also for assisting human beings in any of their activities. Especially with the appearance of the personal computer, and of graphic, intuitive interfaces including the concept of direct manipulation, the computer grew closer to naïve users. At the same time, it became evident that much further work was needed in human-computer interaction to make computers friendly and usable by a larger public.

In the following decades, aside investing on the traditional themes - computing power, infrastructure, system development methodology, etc. - Information Technology put effort into improving the intelligence of systems and in their capability of understanding ways of communicating that are natural for humans. In the Nineties the web era opened a new phase, based substantially on two concepts: connectivity (one single information space, irrespective of physical location of resources) and a simple navigational modality of interaction (inspired by the concept of hypertext, itself about as old as the computer). Mobile computing has added one additional dimension to Information Technology and pervasiveness of computing is coming into our lives. Today everyone realizes that we share many aspects of our routine activities with computers. It is understood that something must be done so to make access to information feasible for all (e.g. interpretation of requests more complex than could be possible with a menu-based interface, selection of existing information so that only the most relevant is displayed, etc.). Researches in the field, however, are realizing that computers have the capability of doing more than just mimicking human-human communication. Presenting information flexibly, integrating various means of communication, providing intelligent feedback are just some examples. Intelligent Information Presentation is a field that is assuming an increasing strategic importance. Current intelligent presentation systems are task specific and refer mainly to some specific technologies, but they have something important in common. They all posit that presentations have to take into account the user's characteristics, the user's specific needs and the context in which the interaction takes place. The process of elaborating a presentation starts from some level of internal representation of the content and from some knowledge or experience of previous presentations and exploits the available means of communication for conveying the information in the most appropriate way. In some cases, the system may even have its own communicative goals that may decide which one to pursue as well as when and how.

O. Stock and M. Zancanaro (eds.), Multimodal Intelligent Information Presentation, VII–XII

At the root of the theme of Intelligent Information Presentation we can consider several scientific areas, but at least three are fundamental. Probably the first to be mentioned is Natural Language Generation, the branch of natural language processing that deals with the automatic production of texts. The field normally is described as investigating communicative goals, the dynamic choice of what to say, the planning of the overall rhetorical structure of the text (called sometime strategic planning), the actual realization of sentences on the basis of grammar and lexicon (sometimes called tactical planning), and so on. With a similar objective but with different means, the field of Adaptive Hypermedia combines hypertext (hypermedia) and user modeling. In contrast to NLG, text or multimedia existing documents are manipulated. Adaptive Hypermedia systems build a model of the goals, preferences and knowledge of the individual user and use this throughout the interaction for adaptation of the hypertext to the needs of the user. In other words, AH takes into account the fact that users vary in knowledge, cognitive skills and reasons for searching information. By keeping a model of some aspects of the user's characteristics, the system can adapt to and aid the user in navigating and filtering information that best suits his or her goals. A third important field is computer graphics; it has experienced a fundamental passage toward the end of the Eighties, when it was understood that graphics production should start from internal representations and communicative goals in a way similar to language production. This passage has led to the possibility of developing multimodal systems, that in output would consider the available modalities, possibly the context and the user characteristics, and operate so that the message is allocated and realized in a coordinated way on several media.
Intelligent Interactive Information Presentation has gone further along that line: it relates to the ability of a computer system to automatically produce multimodal information presentations, taking into account the specifics about the user, such as needs, interests and knowledge, and engaging in a collaborative interaction that helps the retrieval of relevant information and its understanding on the part of the user. It may include dimensions such as entertainment and education, opening important connections to areas that were not related to the world of human-computer interaction, such as broadcasting or cinematography. This vision has led to novel concrete aggregations. This is evident in a number of projects, in Europe as well as in America and in Japan, where the teams have included very diverse expertise. The vision has even caused the establishment of novel institutions: a good example is the Institute for Creative Technologies, created at the University of Southern California with the goal of bringing together researchers in simulation technology to collaborate with people from the entertainment industry. The idea was that "more compelling simulations could be developed if researchers who understood state-of-the-art simulation technology worked together with writers and directors who knew how to create compelling stories and characters." (see the chapter by Swartout et al. in this volume).

Looking at the market prospects, many areas of IT will benefit from results in Intelligent Interactive Information Presentation. Future developments in the field are expected to have a significant impact on business output by increasing

productivity and improving quality of business documents' content through rich user feedback. Training of employees at a company is another area that i3p technology can significantly enhance by improving the learning process. Systems able to provide assistance to customers that face, for instance, technical difficulties with an electronic product, can translate into significant savings for businesses, while also providing high-quality and tailored assistance to their customers. They can substantially simplify the presentation of information to nonexpert computer users as well as provide services for individual with various impairments. The tourism industry (about to become the most important business on the web) is an excellent example of the trend toward increasing demand from users for a personalized service. A privileged general area of application is education: Intelligent Information Presentation technology will bring a whole new experience for the user providing personalized information to allow individuals to learn at their own pace while gaining access to relevant information that efficiently complements their knowledge base. Obviously, interactivity in the entertainment industry can also be significantly enhanced by Intelligent Information Presentation technology by creating a personalized experience for the user. And many more areas could be indentified.

With this book we have put together a collection of highly representative works that show the advancements and the trends in the field. The book is organized along five clusters.

The first cluster encompasses three contributions on information presentation using lifelike characters.

The contribution by Poggi, Pelachaud, de Rosis, De Carolis and Carofiglio illustrates the architecture of Greta, a believable embodied conversational agent built in the context of the EU Project MagiCster. Greta engages the users in natural conversations. It is provided with a mind, a dialogue manager and a body. The mind triggers an emotion on the basis of the agent's personality, the occurring events and the current dialog move. The dialog manager selects the appropriate dialog move to perform. The body, which includes a 3D face model is capable of expressing a set of nonverbal communicative functions.

The second chapter, by Paggio and Jongejan, describes the multimodal communication components of a virtual world application developed at the Centre for Language Technology within the Danish research project "Staging of 3D Virtual Worlds". The application is a virtual farm in which the user can interact with an autonomous farmer agent to participate in simple conversations and actions concerning work on the farm. Communication is multimodal in that the agent can respond to verbal input combined with a limited number of deictic, iconic, and turn-taking or giving hand gestures. Speech and gesture inputs are synchronised and, if the relevant syntactic and semantic constraints are satisfied, a unified interpretation of both is generated by a feature-based multimodal parser.

Finally the work of Theune, Heylen and Nijholt discusses the output modalities available for information presentation by embodied, human-like agents which include both language and various nonverbal cues such as pointing and gesturing. Nonverbal modalities can be used to emphasize, extend or even replace the language

output produced by the agent. In this chapter, the authors discuss the issues involved in extending a natural language generation system with the generation of nonverbal signals.

The second cluster comprises three contributions on the topic of information presentation on mobile devices. Mobile navigation systems are considered one of the possible breakthrough technologies for broad band wireless internet access (i.e. for UMTS). Although quite sophisticated systems already exist for different transportation means, e.g. car navigation systems and more recently also pedestrian navigation systems, these systems only function in a single well defined context. They are designed either to be operated by a driver in a car, or a by pedestrian in an outdoor scenario. In contrast to that personal navigation systems are supposed to work in all of these contexts, regardless of the chosen mean of transportation or other external factors.

The first chapter of this cluster, by Baus, Krüger and Stahl, focuses on the conceptual and technical requirements of such a personal navigation system. They compare their approach with classical mobile navigation systems and argue that resource-adaptiveness has to be achieved on three different levels of the navigation task: the wayfinding level, the presentation planning level and the presentation rendering level.

Stock, Zancanaro and Not discuss the opportunities and challenges offered by the area of cultural heritage appreciation for innovative, natural–language centered applications. They introduce the PEACH project, aimed at exploring various technologies for enhancing the visitors' experience during their actual visit to a museum.

Finally, in his contribution, Rist discusses the problem of the overwhelming variety of new portable computing and communication devices. This adds a new level of complexity to the design of usable information services since different users may not be equipped equally in terms of output and input capabilities. With a special focus on graphical representations, this contribution sketches several application scenarios taken from projects conducted at DFKI and discusses the technical approaches to accomplishing adaptation tasks that result from device restrictions, such as limited screen real estate, lack of high resolution and color.

The cluster on Natural Language Generation presents four contributions.

First, the chapter by Andrè and colleagues describes a system called AutoBriefer which automatically generates multimedia briefings from high-level outlines. The author of the briefing interacts with the system to specify some of the briefing content. In order to be scalable across domains, AutoBriefer provides a graphical user interface that allows the user to create or edit domain knowledge for use in the selection of some of the briefing content. The system then uses declarative presentation planning strategies to synthesize a narrated multimedia briefing in various presentation formats. The narration employs synthesized audio as well as, optionally, an agent embodying the narrator.

Calder and colleagues address a similar issue in their chapter. They introduce The M-PIRO project which targets the concept of personalized information objects. M-PIRO's technology allows textual and spoken descriptions of exhibits to be generated automatically from an underlying language-neutral database and existing

free-text descriptions. The resulting descriptions, produced in three different languages (English, Greek, and Italian), are tailored according to the user's interests, background knowledge, and language skills.
Carenini and Conati in the last chapter of this cluster describe a framework that helps students learn from examples by generating example problem solutions whose level of detail is tailored to the students' domain knowledge. When presenting a new example to a student, the framework uses natural language generation techniques and a probabilistic student model to selectively introduce gaps in the example solution, so that the student can practice applying rules learned from previous examples in problem solving episodes of difficulty adequate to her knowledge. Filling in solution gaps is part of the meta-cognitive skill known as self-explanation (generate explanations to oneself to clarify an example solution), which is crucial to effectively learn from examples.
Finally, Piwek, Power, Scott and van Deemter explore NLG applications in which extra media are brought into play from the point of view of planning the layout of the document. The main point of the chapter is that the extra media are not simply added to plain text, but integrated with it: thus the use of formatting, or pictures, or dialogue, may require radical rewording of the text itself.
The fourth cluster comprises two contributions on the topic of Virtual and Augmented Reality.
Bell, Feiner and Höllerer describe a view-management mechanism that addresses dynamically changing visibility relationships among moving objects, and show how to apply it to 3D user interfaces. The focus is on augmented reality systems, in which users wearing head-tracked, head-worn displays can view graphics overlaid directly onto the real world. Examples from an implemented view-management testbed are used to demonstrate some of the different ways in which the system and its users can exert control over what is seen in augmented reality. These examples include automatically laid out annotations that provide desired information about the surrounding environment, and a hand-held tracked physical display and head-tracked virtual situation-awareness aid that respond to user control and interact with the annotated environment.
The work of Swartout and colleagues presents the Mission Rehearsal Exercise Project, which confronts a soldier trainee with the kinds of dilemmas he might reasonably encounter in a peacekeeping operation. The trainee is immersed in a synthetic world and interacts with virtual humans: artificially intelligent and graphically embodied conversational agents that understand and generate natural language, reason about world events and respond appropriately to the trainee's actions or commands.
The final cluster of the book is about the future. It consists of a contribution by Bunt, Kipp, Maybury and Wahlster that introduces a recommended action plan and a roadmap for the topic of multimodal interactive information presentation.

The occasion that gave impulse to this volume was the European CLASS Project (IST-1999-12611). CLASS (http://www.class-tech.org) was launched at the request of the European Commission with the purpose of stimulating cross-project

collaboration among EC-sponsored Human Language Technology projects as well as between those projects and relevant projects world-wide.
Among the initiatives of CLASS, two workshops were held recently in Verona, and Copenhagen with submissions from Europe, the US, Canada, Australia and Japan. Subsequent to the workshops, the authors of the best selected extended abstracts in the proceedings have been invited to write a long paper for the present volume that was reviewed also by authors of other papers in the collection.
Finally, we would like to thank Ivana Alfaro for her help in this enterprise and Erika Belli for her work on the manuscript.

Oliviero Stock
Massimo Zancanaro

LIFE LIKE CHARACTERS

The research field on lifelike characters is a promising area for investigating challenging issues in Artificial Intelligence and Human-Computer Interfaces. Lifelike characters are a kind of multimodal interfaces where the modalities are those natural in human conversations: speech, gestures as well as facial expressions and body postures.

The hypothesis is that intelligent user interfaces can take advantage of embodied intelligence to facilitate human-machine interaction. From this point of view lifelike characters constitute another argument in the long-term debate on anthropomorphism in the interface. Indeed, human-like embodied interfaces have become popular as the front end to many web sites and as part of many computer applications. However, these interfaces have just the graphical appearance of a body without any internal representation of the communicative functions of a body and they often produce far from satisfactory results. The contributions in this cluster take the opposite approach, they all try to investigate how to properly model the communicative functions of a body in order to produce realistic behavior.

The contribution by Poggi, Pelachaud, de Rosis, De Carolis and Carofiglio illustrates the architecture of Greta, a believable embodied conversational agent built in the context of the EU Project MagiCster. Greta engages the users in natural conversations. It is provided with a mind, a dialogue manager and a body. The mind triggers an emotion on the basis of the agent's personality, the occurring events and the current dialog move. The dialog manager selects the appropriate dialog move to perform. The body, which includes a 3D face model is capable of expressing a set of nonverbal communicative functions.

The second chapter, by Paggio and JongeJan, describes the multimodal communication components of a virtual world application developed at the Centre for Language Technology within the Danish research project "Staging of 3D Virtual Worlds". The application is a virtual farm in which the user can interact with an autonomous farmer agent to participate in simple conversations and actions concerning work on the farm. Communication is multimodal in that the agent can respond to verbal input combined with a limited number of deictic, iconic, and turn-taking or giving hand gestures. Speech and gesture inputs are synchronised and, if the relevant syntactic and semantic constraints are satisfied, a unified interpretation of both is generated by a feature-based multimodal parser. Finally the work of Theune, Heylen and Nijholt discusses the output modalities available for information presentation by embodied, human-like agents which include both language and various nonverbal cues such as pointing and gesturing. Nonverbal modalities can be used to emphasize, extend or even replace the language output produced by the agent. In this chapter, the authors discuss the issues involved in extending a natural language generation system with the generation of nonverbal signals.

I. POGGI, C. PELACHAUD, F. DE ROSIS, V. CAROFIGLIO AND
B. DE CAROLIS

GRETA. A BELIEVABLE EMBODIED CONVERSATIONAL AGENT

1. INTELLIGENT BELIEVABLE EMBODIED CONVERSATIONAL AGENTS

A wide area of research on Autonomous Agents is presently devoted to the construction of ECAs, Embodied Conversational Agents (Cassell et al. 2000; Pelachaud & Poggi, 2001). An ECA is a virtual Agent that interacts with a User or another Agent through multimodal communicative behavior. It has a realistic or cartoon-like body and it can produce spoken discourse and dialogue, use voice with appropriate prosody and intonation, exhibit the visemes corresponding to the words uttered, make gestures, assume postures, produce facial expression and communicative gaze behavior.

An ECA is generally a Believable Agent, that is, one able to express emotion (Bates, 1994) and to exhibit a given personality (Loyall & Bates, 1997). But, according to recent literature (Trappl & Payr, in press; de Rosis et al., in press a), an Agent is even more believable if it can behave in ways typical of given cultures, and if it has a personal communicative style (Canamero & Aylett, in press; Ruttkay et al., in press). This is, in fact, what makes a human a human. More, an ECA must be interactive, that is, take User and context into account, so as to tailor interaction onto the particular User and context at hand.

In an ECA that fulfils these constraints the communicative output, that is, the particular combination of multimodal communicative signals displayed (words, prosody, gesture, face, gaze, body posture and movements) is determined by different aspects: *a.* contents to communicate, *b.* emotions, *c.* personality, *d.* culture, *e.* style, *f.* context and User sensitivity. At each moment of a communicative interaction, all of these aspects combine with each other to determine what the Agent will say, and how.

In this paper we show how these aspects of an ECA can be modeled in terms of a belief and goal view of human communicative behavior. We then illustrate Greta, an ECA following these principles which is being implemented in the context of the EU project MagiCster1.

2. WHAT AND HOW WE COMMUNICATE

Before focusing on the Agent's dialogue, let us see how a single move can be represented and simulated. An Agent S (Sender) generates a set of beliefs - about

O. Stock and M. Zancanaro (eds.), Multimodal Intelligent Information Presentation, 3–25

itself or about external objects and events - and has the goal to make Agent A (the Addressee) believe them. To achieve this, S produces a set of communicative signals (for example words, gestures, gaze, head, face and body behavior), taken from its communicative repertoire and temporally ordered in a particular way.

Now, what are the beliefs an Agent may conceive to communicate? What is the structure of its communicative repertoire? What are the rules for the temporal ordering and synchronization of different signals? But also: how does the system go from the input – a set of beliefs – to the output – that particular arrangement of words, voice, hands, body, face signals? In fact, our communicative repertoire is very complex, in order for us to modulate our communication in a very sophisticated way. For example, words have formal and informal variants, connotations, positive or negative, tender or insulting nuances; and not only is the verbal lexicon so rich, but we can also communicate by subtly different intonations, gestures, facial expression, gaze, posture, spatial behavior.

In each move of a dialog, how do we choose the best way to communicate, the combination of verbal and non verbal signals that are most fit to express our communicative goal? How do we activate the goal of using that given word, replacing it with a gesture or using both to communicate our meaning?

According to a goal and belief model of social action (Conte & Castelfranchi, 1995), choice, that is, the decision to pursue a goal instead of another is determined by the relative values attached to the alternative goals. But the value of a goal in its turn stems from the value of its superordinate goals, or from the algebraic sum of the values of two or more of them. So, whoever discovers his car was stolen might shout. But if this happens to someone who is just starting a work where he needs his car, his shout will be sharper or longer and his utterance more aggressive. In other words, resources we specifically use in a given communicative situation (a gesture in the place of a word, a very colloquial term instead of a more formal one) are determined by a number of permanent and contingent factors.

3. PERMANENT AND CONTINGENT FACTORS

As an Agent enters a communicative interaction, having the goal of communicating some meanings, two kinds of factors affect the final aspect of his/her communication: permanent and contingent ones (Table 1). The former are the goals and resources coming from the Senders' biological and cultural endowment, that are always active in them; the latter are the goals activated and the resources provided by the contingent situation in which the Senders communicate.

Long-lasting internal resources are (1) personality, (2) social identity (age, gender, cultural roots) and (3) cognitive traits. Among them, we may distinguish (4) innate and (5) culturally learned features: innate ones may be (6) a higher or lower capacity of making inferences, different reasoning styles (more abstract, intuitive, or imaginative); or (7) the different aptitudes, partly depending on neurological dispositions, towards musical, mathematical, visual or linguistic skills.

Table 1. Factors affecting the choice of communication resources

contents to communicate communicative choice

<table>
<tr><td colspan="3">PERMANENT</td><td colspan="3">CONTINGENT</td></tr>
<tr><td colspan="3">1 S's personality</td><td rowspan="2">11
Self</td><td colspan="2">14 Physical resources
(motor capacity and energy)</td></tr>
<tr><td colspan="3">2 S's social identity</td><td colspan="2">15 Cognitive resources
(drunk, concentrated)</td></tr>
<tr><td rowspan="10">3
S's
Cognitive
Traits</td><td rowspan="3">4
Innate</td><td>6
Inference capacity</td><td rowspan="10">12
Other</td><td rowspan="3">16
Physical resources</td><td>18
Sensory capacity</td></tr>
<tr><td>7
Aptitudes</td><td>19
Media</td></tr>
<tr><td rowspan="7">5
Learn-ed</td><td>8
Knowledge Base</td><td rowspan="4">17
Cognitive resources</td><td>20
Knowledge Base</td></tr>
<tr><td>9
Cultural techniques, norms, values</td><td>21
Inference capacity</td></tr>
<tr><td rowspan="2">10
Communica-tive repertoire</td><td>22
Communicative repertoire</td></tr>
<tr><td colspan="2">23
Personality</td></tr>
<tr><td rowspan="5">13
Situation</td><td rowspan="2">24
Physical setting</td><td>25
Available modalities</td></tr>
<tr><td>26
Presence of referents</td></tr>
<tr><td>27
Social setting</td><td>28
S-A relation
(status, role, affect)</td></tr>
<tr><td colspan="2">29
Type of encounter
(service, affective)</td></tr>
<tr><td colspan="2">30
Relations to others
(in public, in private)</td></tr>
</table>

Other cognitive capacities are culturally dependent: culture entails beliefs about the environment (8) and about strategies of behavior typical of a population (see Section 8.), but also norms on how to do things, how to behave, what and how to

communicate (9), and finally the communication repertoires (verbal and nonverbal) one comes to learn since infancy (10).

The combination of personality traits, attitudes and culturally learned behavior habits gives rise to style, an idiosyncratic tendency to behave in some peculiar way that allows others to recognize the Agent's identity. The contingent resources taken into account for the choice of the output communicative behavior are provided, instead, by the intrinsic features of context (Poggi & Pelachaud, 2000).

The presence or absence of these resources in oneself (11), the Other (12) or the Situation (13) activate specific goals in Senders as they plan their discourse and then the communicative signals to exploit. If I am very tired (physical resources, 14) I'll tend to speak low and not to make conspicuous gestures; if my interlocutor is a bit deaf (sensory capacity of other, 18), I'll speak loud. If he is a tourist (Other's communicative repertoire, 22) I'll talk slowly; with a student (Knowledge Base, 20) I'll explain things at length; with a not so smart person I'll explain even obvious causal links (inference capacity, 21). And if the interlocutor is a touchy person (personality, 23), I'll be more indirect in my criticism.

Beside the features of oneself and the interlocutor, resources present or absent in the environment trigger different goals about what and how to communicate: in a noisy discotheque (physical setting, 24), I'll use gestures instead of words (available modality, 25), while I will not do this on the phone (absence of referents, 26). But more than the physical, the social setting (27) is relevant: we use more polite words with a high status Addressee (S-A relation, 28) or in formal situations (type of encounter, 29); less colloquial words in public than in private (relation to others, 30).

How do we choose, then, how to communicate in a given situation? Presence or lack of individual resources lead us to choose one signal or combination of signals instead of another. Our hypothesis is that every combination of an Agent's multimodal communicative behavior (that particular word, uttered by a particular intonation, while making that gesture, that gaze, facial expression, and posture) is the result of the final choice of the communicative goal to pursue. This goal is selected among the Agent's different goals, determined in their turn by contingent events (content to communicate, felt emotions, context and interlocutor) and by long-lasting features (the Agent's culture, personality and style).

In Sections 4 through 9 we overview how the above aspects (meanings to communicate, emotion, personality, culture and style) can be viewed according to a goal and belief model of human communicative interaction, and how they may be represented in an ECA. In Section 10 we describe the architecture of the ECA "Greta", while showing how some of the above principles can be applied to construct a Believable Embodied Conversational Agent.

4. MEANINGS TO CONVEY

Let us first overview the beliefs that may form the content of a communicative act. Three classes of meanings can be distinguished (Poggi, 2002 b): Information on the *World*, Information on the *Speaker's Identity* and Information on the *Speaker's Mind*.

Information on the World. As we communicate, we provide information on concrete or abstract events, their actors and objects and the time and space relations among them. Such information is provided mainly through words, but also by gestures or gaze. To mention the referents of our discourse, we may point at them by deictic gestures or gaze; to refer to some properties of objects we may use iconic or symbolic gestures and even gaze (as when we squeeze eyes to mean 'small' or 'difficult').

Information on the Speaker's Identity. Physiognomic traits of our face, eyes, lips, the acoustic features of our voice and often our posture provide information on our sex, age, socio-cultural roots, and personality. And, of course, our words can inform on how we want to present ourselves.

Information on the Speaker's Mind. While mentioning events of the external world, we also communicate why we want to talk of those events, what we think and feel about them, how we plan to talk of them. We provide information on the *beliefs* we are mentioning, our own *goals* concerning how to talk about them and the *emotions* we feel while talking (Poggi 2002 a).

About *beliefs*, we inform about:

1. *degree of certainty*: words like 'perhaps', 'certainly'; conditional or subjunctive verb modes; but also frowning, which means 'I am serious in stating this'; opening hands, which means 'This is self-evident';
2. *metacognitive* information: that is, the source of mentioned beliefs, whether they originate from memory, inference or communication (we look up when trying to make inferences, snap fingers while trying to remember,...)

We inform about the following *goals*:

1. *performative* of the sentence (by performative verbs, intonation, facial expression);
2. *topic-comment* distinction (by batons, eyebrow raising, voice intensity or pitch);
3. *rhetorical relations*: class-example (saying "first…second…third..."; counting on fingers) topic shift (expressed through posture shift);
4. *turn-taking and backchannel*: raise hand for asking turn; nod to tell the Interlocutor we are following what she says.

Finally, we inform on the *emotions* we feel while talking (by affective words, gestures, intonation, facial expression, gaze and posture).

5. COMMUNICATIVE REPERTOIRE

Let us now see what is the structure of the communicative repertoire (Table 1, n.10): the set of innate and learned features and behaviors that we can use as signals in order to convey the meanings specified above. The Human Agent is endowed with different "mode-specific communicative systems", that is, sets of rules to link meanings to signals that work in different modalities.

Several parts of the Human Agent's body produce communicative signals: so head, face (with eyes and mouth), hands, trunk and legs can be viewed each as the depository of a communicative system. Communicative systems may be of at least

two kinds: "codified" and "creative". In a "codified" system, or "lexicon", the same signal-meaning link is shared and coded in the Agents' memory. Not only words or symbolic gestures but also gaze, facial expression, posture form lexicons, where each signal represented in an Embodied Agent's mind corresponds to a specific meaning. In a "creative" system, instead (for instance the system of iconic gestures), what is coded in memory is only a small set of inference rules about how to create a new signal starting from a given meaning, or about how to retrieve a meaning from a given signal (Magno Caldognetto & Poggi, 1995).

In our hypothesis, in all communicative systems, meanings may be represented in terms of 'mental images' or logical propositions. For example:

Let `Ai` denote the Sender and `Aj` the Addressee, `a` an action and `g` a goal that may be achieved by means of `a`. The performative of a 'Peremptory order' may be represented as follows:

1. Goal Ai (Do Aj a)
2. (Goal Ai g) $\wedge$ Bel Ai (Achieve a g))
3. Goal Ai (Bel Aj (Power-on Ai Aj a))
4. If (Not (Do Aj a)) then (Feel Ai Angry)

while the performative of an 'Imploration' may be represented as follows:

1. Goal Ai (Do Aj a)
2. (Goal Ai g) $\wedge$ Bel Ai (Achieve a g))
3. Goal Ai (Bel Aj (Power-on Aj Ai a))
4. If (Not (Do Aj a)) then (Feel Ai Sad)

In both ordering and imploring, the Sender wants the Addressee to do some action. However, the two performatives differ for the power relationship between the two interlocutors (`Ai` has power over `Aj` in the former, and the reverse in the latter), and for the potential emotion (anger vs. sadness) in case `Aj` does not perform the requested action (Poggi & Pelachaud, 2000).

A compositional representation can be adopted not only on the side of the meaning, but also for the signals: a signal is represented as a combination of behavioral parameters, each with its number of values. Gestures are decomposed into hand-shape, arm and wrist position, type of movement; gaze into direction of eyes, eyelid aperture, movements of the eyebrows, and so on (Poggi, 2002 b). This compositional representation on the two sides enables achieving a high flexibility in the correspondence between signals and meanings. It opens, as well, the possibility that a combination is not always conveyed by the same signal, but may be expressed through different parameters of signals taken from different communication systems (Pelachaud & Poggi, 2002).

The overall communicative repertoire, in fact, manages the combination of signals taken from different communicative systems, the distribution of meanings across them, and their temporal ordering and synchronization.

6. EMOTIONS

Before showing why and how emotions are an important determinant of an Agent's communication, let us define them according to a goal and belief model of action and social interaction (Conte & Castelfranchi, 1995).

6.1. Emotions and goals

Actions in our life are often part of a plan aiming at some goal. Take for example an action of Oetzi, the 5000 B.C. pre-historic man of Similaun, who chooses a stone apt to sharpen well and makes a lance for chasing the wild-pig successfully. The actions of looking for a good stone and to sharpen it are just means for the complex action of chasing. But also chasing is aimed at feeding himself and his group, which in turn aims at the goal of survival. The goals of our everyday plans are not ends in themselves: they all aim at more general goals of biological import that are common to all humans, like the biological goals of survival and reproduction and some sub-goals of them, physical well-being, safety, loving and being loved, self-realization, image and self-image. These are *terminal* goals, that are ends in themselves and ones to which we assign the highest weights. So much that, if two of them are incompatible (as for instance freedom vs. life itself), giving up one of them is a heavy renunciation. With respect to terminal goals, the goals of our everyday life are *instrumental* goals, in that they directly or indirectly serve our terminal goals. For instance, chasing the wild-pig with a sharpened stone is instrumental to survival: if the lance is not sharp enough and does not hit the wild-pig to death, the wild-pig might aggress and kill Oetzi. Instrumental goals are more or less important to us, depending on the strength of their link with terminal goals: at the extent to which an Instrumental Goal is likely to be the only possible means to reach a Terminal Goal, that Instrumental Goal receives a high weight, just because it inherits its weight from the Terminal Goal it serves.

Emotions are a biological device aimed at monitoring the state of reaching or threatening of our most important goals, be they terminal or instrumental (see, for instance, Carbonell, 1980; Oatley & Johnson-Laird, 1987). Anytime something happens (or the Agent believes it happens) that is likely to produce the achieving or threatening of a highly weighted goal, the biological device of emotion is triggered: from the agent's interpretation of the situation, a complex subjective state originates, generally of a short duration and with different degrees of intensity. This state includes physiological, expressive and motivational aspects. If Oetzi throws his lance but sees it has not run into the wild pig's heart, fear is triggered since his goal of survival is challenged: physiological reactions are activated (blood flowing away from face to limbs) some of which may show in the perceivable state of his body (pale face, tremors); and the specific goal of escaping, that might serve the terminal goal of survival, is activated.

There is a strong relationship, then, between goals and emotions: goals both cause emotions and are caused by emotions. They cause emotions since, if an important goal is achieved or threatened, an emotion is triggered: emotions are therefore a feedback device that monitors the reaching or threatening of our high-

weighted goals. At the same time, emotions activate goals and plans that are functional to re-establishing or preserving the well-being of the individual, challenged by the events that produced the emotions. So, fear triggers flight, anger triggers aggression, guilt triggers the goal of helping the harmed person or of escaping sanction (Castelfranchi, 2000).

6.2. Emotion triggering vs. emotion display

Emotions may be implied in communication in at least two ways.

1. they may be the very reason that triggers communication: we activate the goal of communicating just because we want to express our emotion;
2. they may intervene during our communication, as a reaction to what our interlocutor is saying, or to some thought suddenly coming to our mind, either related to the ongoing dialogue or not.

In both cases, the triggering of emotion does not necessarily imply that the Agent displays it. There are many reasons why we may refrain from expressing our emotion, and the final (aware or non-aware) decision of displaying it may depend on a number of factors (Prendinger & Ishizuka, 2001; De Carolis et al., 2001). Some of them concern the very nature of the emotion felt (emotion nature), others the interaction of several contextual (scenario) factors.

1. Emotion nature
    a. *Intensity* (a more intense emotion might be more likely displayed);
    b. *Valence* (it is not the same to display negative or positive emotions);
    c. *Social evaluation* (some emotions, like envy or shame, are subject to social sanction: then it is more difficult to express them);
    d. *Addressee* (it is different to express an emotion to the one who caused it or to a third person).
2. Scenario Factors
    a. Agent's *Display motive* (displaying or not depends on whether you do it to be helped, consoled, or if you want to demonstrate or teach something);
    b. Agent's *personality* (an impulsive person is generally more keen to displaying than a reflexive or a shy one);
    c. *Interlocutor*'s features (displaying depends on the other's personality, empathy, intelligence...).
    d. Agent - Interlocutor *Role relationship* (whether he has power over you or you over him);
    e. Agent - Interlocutor *Personal Relationship* (you might not display your being worried to someone you love, if you want to protect him);
    f. Type of *social interaction* (being in public makes a difference for emotion display).

7. PERSONALITY

Personality is also linked to goals: it can be viewed in terms of weights people attribute to terminal goals (Carbonell, 1980; Poggi and Pelachaud, 2000). For example, a sociable person gives more importance than others to knowing and staying with other people. A selfish person, when the goals of physical safety and others' care are conflicting, chooses to pursue the former, while an altruist pursues the latter. A proud person attributes a high value to self image and autonomy, while a dependent person cares others' image more than self-image.

Since both emotions and personality have to do with the relative importance of goals, there is also some link between emotion and personality. Some personality traits may be viewed in terms of the general 'propensity to feel emotions' (Plutchick, 1980; Poggi and Pelachaud, 1998). Picard (1997) calls 'temperament' this subset of personality traits, while other authors relate them directly to one of the factors in the 'Big-Five' model: for instance, neuroticism (Mc Crae & John, 1992). These traits imply, in a sense, a lower threshold in emotion feeling (Ortony et al., 1988). For instance, a 'shy' person is keener to feel 'shame', especially in front of unknown people. A 'proud' person, who attributes a high weight to his goals of self-esteem and autonomy, will feel particularly proud of himself every time one of these goals is achieved. And, conversely, every time they are threatened (if, say, he is obliged to ask for help), the person will feel the opposite emotion, shame. Thus, a personality trait (proud) is related to attaching a higher weight to a particular goal (self-esteem, autonomy); and, since that goal is particularly important to that kind of person, the person will feel the corresponding emotions (pride or shame) with a higher frequency or intensity.

8. CULTURE

Culture may also be viewed in terms of different weights on goals. Both a Somali shepherd and an Italian housewife have the goal of feeding themselves and their family; but the sub-goal chosen to pursue this goal may be for the former to search the bush, for the latter to go shopping in a store.

8.1. Goals and culture

Humans pursue their goals by using external resources (presence of food, characteristics of the territory, climate conditions) and internal resources (physical strength, body agility, manual skill, beliefs and intelligence). Since different physical environments provide different resources, each population, given the environment in which it lives, comes to accumulate a set of beliefs on the instrumental goals that most easily and economically serve the biological terminal goals in that environment. At the extent to which an instrumental Goal is (or is likely to be) the only possible means to reach a terminal Goal, given the external conditions available, that instrumental Goal will receive a higher weight than other possible ones. Thus, the instrumental goal chosen becomes a strategy of survival typical of that culture; and culture, overall, may be defined as a set of beliefs on the

typical techniques to pursue goals. Of course it also entails beliefs about the external world. And since language is both produced by beliefs and a vehicle of them, culture typically shows up in language. Language is made of the beliefs of a population, and a way to organize them, a set of rules on how to conceptualize and categorize information. Consequently, it also implies a set of settled communicative techniques, that is of settled instrumental goals, stating how to convey information.

More, culture entails values and norms. Values are evaluative beliefs about what is good and then has to be pursued as a goal (Miceli & Castelfranchi, 1989). But since particular ways to behave may be good or bad according to the environment, again due to what are the most useful techniques of survival, different populations in different environments may hold different values. Where individualistic behavior proved convenient, individualistic values will develop, while in environments where collectivistic behavior is more fit, values centered on the family or the group will hold.

Norms are obligations that rule the relationships among people in a group (Conte & Castelfranchi, 1995). In a culture more centered on interdependency, a norm (then, a highly weighted goal), may prescribe to be cooperative, even when this implies intruding in the other's affairs; in a culture centered on the individual's autonomy, the goal of keeping one's privacy will be more weighted, and a norm will hold of not intruding in others' affairs and of contrasting others' intrusions.

Now, since values and norms generate goals in people (the goal to pursue that value or to respect that norm), if they are thwarted, they provoke emotions. Not living up to one's values may induce shame, while violating norms may cause guilt. Therefore, if two populations have different values and norms, they will also feel these emotions out of different events.

To sum up, culture is a set of beliefs shared by a population about the environment in which the population lives and the best techniques (the most highly weighted instrumental goals) to reach the biological terminal goals in that environment, given the means-end relations that hold in the given physical conditions and the set of beliefs accumulated. Culture also includes beliefs about how to gather and organize beliefs, and about the norms and values that are functional to techniques of goal achievement that best fit the surrounding environment. According to this definition, let us try to figure out how the way people communicate changes with cultural differences, by trying to distinguish what is universal (biological) and what is culturally determined in the different aspects of communication. These differences may then be simulated in a Believable Embodied Agent, in words or discourse planning, in gesture, gaze, facial expression, body posture and proxemic behavior.

8.2. Semantic rules versus norms of use in communication systems

A communication system includes two kinds of rules: semantic rules and norms of use. *Semantic rules* concern the correspondence between meanings and signals, be they words, signs or sentences in a verbal or sign language, or gesture or gaze items. These are rules of the following type:

if you want to communicate the meaning ”I greet you”, say “Hello”
if you want to communicate the meaning ”I greet you”, raise your eyebrows
if you want to communicate the meaning ”I greet you”, wave your hand

Norms of use, instead (those studied by Pragmatics and Sociolinguistics) do not state how some meaning has to be conveyed, but if some meaning can, should or should not be conveyed in a given situation. They are rules of this type:

if you meet a person you know, apply the rule for the meaning “I greet you”, or
if you meet an unknown person, do not apply the rule for the meaning “I greet you”.

Now, in some communication systems (for example spoken languages) both semantic rules and norms of use vary across cultures. But in others (typically, the facial expression of primary emotions) semantic rules might be everywhere the same (a grimace of anger is performed in the same way in all cultures), while cultural difference holds in the norms of use (in Japan expressing anger is much more sanctioned than in the USA).

8.3. What is universal and what is cultural in speech, gestures, and face communication?

- Words and sentences

Even if at a deep level the syntax of all languages might be universal (as argued by the Chomskian Universal Grammar approach), specific words and syntactic rules differ from one language to another. Some mechanisms like iconicity hold in all languages, but cultural variation across languages holds at various levels: in the strategies of discourse planning, the importance of politeness or rhetoric, in the rules defining what to speak about, how much to mention the self, whether to convey new information or just make reference to shared knowledge, and so on.

- Gestures

Some gestures are culturally codified (e.g. the gestures for ‘OK’ or ‘Victory’), while others are biologically codified (raising fists up to show elation). If we want to simulate the former in a culture-sensitive Agent, they will have to be varied from a culture to another. Non-codified iconic gestures, instead, might be generated all through the same set of inference rules (supposedly universal), whatever the culture the Agent comes from; but if the referent represented is typical of a culture or an action is performed in a way typical of it, then also a creative gesture may be culturally dependent.

- Gaze and Facial Expression

Facial expression and gaze are more likely to be universally shared than gesture. They can communicate information on the world (we point at things with chin or gaze, squeeze eyes to say that something is little or difficult), information on the Speaker’s beliefs, goals and emotions (we raise eyes while we remember or make

inferences; we frown to communicate anger, concentration, or an order) and information on the Speaker's identity (our face and gaze provide information on sex, age, ethnicity, personality, sometimes even social class).

Focusing on the expression of emotions, even if the so-called *basic emotions* (happiness, sadness, anger, fear, surprise and disgust) are felt in all cultures, and everywhere they trigger an innate universal neural program for facial expression (Ekman, 1982), this does not imply that people in different cultures always show their emotions in the same way in the same situation. First, an emotion is triggered by the cognitive categorization of a situation on the part of the subject, and a situation that in a culture, because of its beliefs, norms and values, is categorized as a cause of sadness, in a different culture might be seen as a cause of happiness (take the death of a martyr in the Islamic culture). Second, the emotion display is filtered not only by cognitive and personality traits of the Agent and of the Interlocutor, their relationship and the situation, but also by the cultural norms about the expression or non-expression of given emotions.

9. STYLE

Style is an internal feature of an Agent that affects its choice of communicative resources. We define style as the idiosyncratic stable tendency of a specific agent to exhibit specific communicative or non-communicative behaviors. Communicative style is then the tendency one has to choose some signals, or arrangements or aspects of signals, instead of others. Agent X uses formal words also in private informal situations; Y's discourse is always thoroughly explained; Z makes ample gestures; W avoids the Addressee's gaze. In its being a stable tendency to communicative choice, style is the ultimate result of the combination and interaction of the permanent goals that rule an Agent's behavior, namely the goals stemming from its personality, attitudes, and culture. A particular combination of these goals, at the moment of the communicative output, affects the choice of a particular signal or set of signals, or simply the value of some parameters in the production of signals. High level goals implied in a specific personality get instantiated into lower level goals. For example, in an introverted person, the general goal of not being too visible may result in a goal of avoiding too wide gestures. A tendency to visual thought may lead to use metaphors in discourse. The habit of explaining to students may induce to be always didactic.

Style can manifest itself both in the choice of whole signals (I may prefer a metaphorical to a literal word) and in the choice of a particular way to produce a signal. For example, people may differ in their style as to various aspects of gesture use (Ruttkay et al., in press). Differences may concern:

1. whether to use a gesture or not, which depends on at least three factors:
    a. *width of repertoire*: we cannot convey a meaning by a gesture if our culture or personal gestural competence does not include a gesture for that meaning;

   b. *threshold*: the degree of formality, as well as other factors like need for redundancy, may provide a threshold for the use of a gesture;
   c. *redundancy*: I may use both a word and the corresponding gesture if I want to stress some meaning or if I want to be very clear.
2. ways of performing gestures**.** While some parameters and sub-parameters of nonverbal signals like gestures are subject to lexical and sociolinguistic variation, some sub-parameters of movement: tense or relaxed (*tension*), wide or narrow (*amplitude*), smooth or angular (*manner*) might be more affected by style. That is, they are directly determined by the current goal set of the Sender and then by his/her emotional or physiological state.

In conclusion, each signal chosen and the combination of particular values in each parameter or sub-parameter of this signal is determined by the combination of goals that are activated at a certain moment in the Sender, all stemming possibly by the goal set described above.

10. GRETA'S EMOTIONS, PERSONALITY, AND COMMUNICATION

In this Section, we show how some of the principles introduced above have been applied in the construction of an Embodied Agent: Greta. Specifically, the device of emotion triggering and its influence on the Agent's behavior have been already implemented, along with some aspects of its personality, communicative planning, voice and face communication. Some aspects of gestural communication are still to be thoroughly implemented in Greta such as gesture expressiveness (Chi et al., 2000). But this architectural model may also support adaptation to some features of the user target population, such as culture and style (de Rosis et al., in press a; Ruttkay et al., in press; Hartmann et al., 2002).

In order to achieve a high degree of flexibility, necessary to this aim, we base the architecture of our Agent on the following principles:

- the *Mind* and the *Body* of the agent are separated. This separation, in our opinion, helps in achieving a higher flexibility in adaptation of the Agent's behavior and appearance. It offers the opportunity to vary its mental state and its reasoning capacity according to the context and to establish the forms of communication to adopt by considering the technical resources available. Considering interdependency between the body and the mind would mean to include context-dependent rules concerning the different aspects of manifesting behavioral differences during communication into the planning process, and thus to make this step very complicated since it has to solve constraints at both inner and outward levels. In our approach (independence of Body and Mind), at the Mind level, only constraints on the meanings associated to a communicative act are represented, leaving to the Body the task of deciding which signal to employ according to the context:

- at the *Mind* level, represented according to the BDI (belief, desire, intention) theory and architecture (Rao and Georgeff, 1991), the way in which its mental state is related to the context may be represented explicitly. This enables the Agent to vary the decision made (including the 'discourse' that achieves a given

'communicative goal') and the triggered and displayed emotions, according to its mental state's structure;

- at the *Body* level, that can use different physical representations, a context-dependent meaning-signal translation is performed in order to render the selected meaning consistently with the chosen physical body and context rules.

- *relations* between the various phases of the generation process have to be conceived in such a way as to facilitate a flexible adaptation to different features that influence communication. For this reasons, the I/O interface between Mind and Body is constituted by an XML specification that is easily wrapped according to available resources and rendering rules.

10.1. The System Architecture

These principles have been applied, in the context of the MagiCster project, to develop an Embodied Conversational Agent that appropriately combines verbal and nonverbal signals when delivering information, to establish a natural communication with the User. Our Agent achieves a quite rich expressiveness in conversation, by displaying various types of information typically transmitted in human-human dialogs (syntactic, dialogic, meta-cognitive, performative, deictic, adjectival, and rhetorical relation).

It is embodied in a 3D talking head whose name is 'Greta' (Pelachaud, 2002); it has a personality and a social role, as well as the capability of expressing emotions, consistently with the context in which the conversation takes place and with its own goals, and to adapt her behavior to some user features relevant for the selected application domain. At present, we simulate advice-giving dialogs, where the main function of Greta is to provide suggestion and useful information to the User, in the application domain. In our 'mixed-initiative' system, the User can ask Greta some questions; this opens a question-answering sub-dialog after which, if needed, Greta revises her discourse plan according to the User request. At the moment, Greta provides information and advice in the medical domain adapting her behavior to some user features such as the age (children, teens, adult), the role (patient, nurse, parent of a patient) and some cultural features that are relevant for providing more appropriate suggestions both from the content and the signal rendering points of view.

The architecture of our Conversational Agent includes the following main components (Figure 1): i) the Agent's MIND including i.1) the 'Emotional Mind' and i.2) the Dialog Manager, and ii) a Generator of the Agent's Body including ii.1) a Signal Generator, ii.2) a Signal Wrapper and iii.3) the chosen body (Greta in this case) that renders the specified behavior.

Even if we are working in the field of advisory dialogs, the system is domain-independent. A particular application domain is therefore selected as a first step of interaction: a conversational goal in that domain is passed to the Dialog Manager (DM) and the dialog may start. From this goal, an overall discourse plan is produced for the Agent, by retrieving an appropriate 'recipe' from a plan library formalized according to DPML (de Rosis et al., in press b). This plan represents how the Agent

will try to achieve the specified communicative goal during the conversation and is selected according to user features and context. The way the dialog goes on is a function not only of this plan but also of the User Moves and of the Social Context.

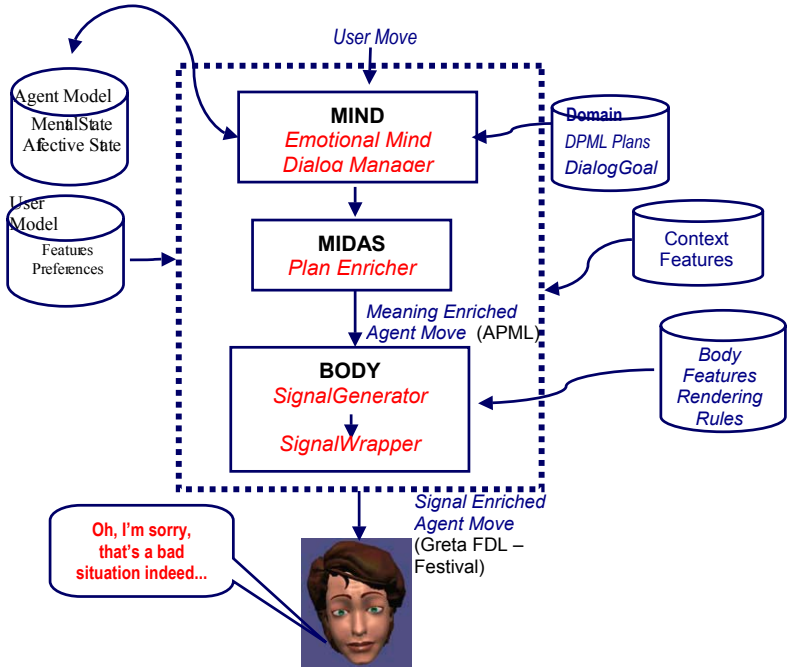

Figure 1. The Architecture of our Conversational Agent

This context includes variables concerning the topic of the dialog, the personality traits of the two interlocutors and their 'attitudes' towards the dialog itself. For instance, the Agent may be defined as more or less 'empathic' towards the User, more or less easily affected emotionally, more or less prone to feel specific emotions with a high intensity (for more details, see Carofiglio et al., in press).

Once an Agent's first move has been generated, the User replies with a move in his or her turn. The DM then asks the emotion simulation module (Emotional Mind) whether a particular affective state of the Agent is activated as a consequence of this move and with which intensity. Then, the DM selects the next move according to the goal activated by the user move, the Agent's new affective state, and the Social Context.

When the DM selects a particular Agent's move, this has to be enriched by the plan enrichment module (MIDAS) that adds tags indicating the meanings to be conveyed. At this point, a further adaptation can be performed by adding meaning

according to the user features, culture and context. This enriched move is then passed to the Body Generator, that interprets it by converting meaning into signals and by rendering them into an expressive behavior. The selection of signals, at this point, can be adapted to the available communication channel of the chosen body, user features and cultural features. We now describe in some more detail the above mentioned modules.

1. The *Emotional Mind* is responsible for updating the Agent's mental state by deciding whether a particular affective state should be activated and with which intensity, and whether the felt emotion should be displayed and how, according to the context variables. Mind is based on a dynamic belief network (DBN), that combines a belief network (BN) 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 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.
2. The *Dialogue manager* (DM) is built on the top of TRINDIKIT (http://www.ling.gu.se/projekt/trindi/). It controls the dialogue flow by iterating the following steps:
    a. after a 'dialog goal' has been specified, an appropriate discourse plan is selected from the library of plan recipes and the first move is generated according to the first step of this plan. The 'dialog goal' becomes the main topic of the conversation;
    b. at the end of this first move, the initiative is passed to the User, that may ask questions to the agent about any subject concerning the main topic under discussion;
    c. the User move is translated into a symbolic communicative act (through a simplified interpretation process) and is passed to the DM;
    d. the DM decides "what to say next" by selecting the sub-plans to execute. At the same time, the information state of Trindi is updated;
    e. the DM goes on cycling over steps b. to d., until the user leaves the conversation.
3. The *Plan Enrichment* module, called MIDAS, has the role of translating the symbolic representation of a dialog move into an Agent's behavior specification. A dialogue move may be very simple (i.e. greet, ask) or it may correspond to a small discourse plan (for instance: describe an object with its properties). In both cases MIDAS generates a tagged output expressed according to a particular XML specification which is interpretable by the Body of any Animated Agent (Affective

Presentation Markup Language: APML. See De Carolis et al., 2002 for details). APML is used as a middle-ware level, to overcome integration problems between the mind and body components and, at the same time, to allow independence and modularity between them. The move is then represented as an APML string, in which the verbal part of the dialog act is enriched with the tags that are needed by the speech and body generation components of the Agent to produce the required expressions: rhetorical relations, deictic or adjectival information, certainty values, metacognitive or turn-taking expressions.

4. The *Body Generator* module interprets the APML-tagged dialog move and decides which signal to convey on which channel for each communicative act. We defined a lexicon of facial expressions and gaze: that is, a set of (meaning, signal) pairs. Each meaning in the taxonomy specified in Section 4. (certainty, metacognitive, comment etc.) corresponds to a particular configuration of parameters of gaze or facial expressions that Greta is able to exhibit. For example, among the same class "certainty", the meaning "certain" corresponds to a "small frown" while the meaning "uncertain" to "raise eyebrow". In order to adapt this correspondence between meanings and signals, we started to introduce some translation conditions depending on the user and the culture of the target user population. The Body we use is a combination of a 3D face model compliant with the MPEG-4 standard (Pelachaud, 2002) and of the speech synthetizer Festival (Black et al., 2002). The facial model is capable of performing the face and gaze expressions foreseen for our conversational agent. Each signal may be expressed as a set of facial parameters. We have developed a language to describe facial expressions easily (De Carolis et al 2002). The text of each dialogue move with its tags is given as input to the animation module and to Festival, which provides the duration of the phonemes and a wav file (an audio file). Phonemes are the smallest temporal unit considered. 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 synchrony between speech and other visual activities. Tags get instantiated by their corresponding signals which are in turn transformed into facial parameters' values. When several tags occur in the same text span, if their corresponding signals act on the same facial regions with different values (e.g. on the eyebrow region the agent should perform a frown and a raised eyebrow, or with the head, a head nod and a head shake), a conflict may arise. In such cases a conflict solver module takes as input the co-occurring meanings and produces as output a complex expression made of a mix of signals from different meanings (Pelachaud & Poggi, 2002).
5. The various components (Mind, MIDAS, Greta and the DM itself) are connected through a *Graphical Interface* which controls activation, termination and information exchange for the various processes involved in the dialogue management.