Description of potential exemplars: Interaction

Catherine Pelachaud and WP6 members

Date: 30th November 2004
<table>
<thead>
<tr>
<th>IST project contract no.</th>
<th>507422</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project title</td>
<td>HUMAINE Human-Machine Interaction Network on Emotions</td>
</tr>
<tr>
<td>Contractual date of delivery</td>
<td>November 30, 2004</td>
</tr>
<tr>
<td>Actual date of delivery</td>
<td>November 30, 2004</td>
</tr>
<tr>
<td>Deliverable number</td>
<td>D6c</td>
</tr>
<tr>
<td>Deliverable title</td>
<td>Description of potential exemplars: Interaction</td>
</tr>
<tr>
<td>Type</td>
<td>Report</td>
</tr>
<tr>
<td>Number of pages</td>
<td>49</td>
</tr>
<tr>
<td>WP contributing to the deliverable</td>
<td>WP6</td>
</tr>
<tr>
<td>Task responsible</td>
<td>Catherine Pelachaud</td>
</tr>
<tr>
<td>Author(s)</td>
<td>See page 5</td>
</tr>
<tr>
<td>EC Project Officer</td>
<td>Philippe Gelin</td>
</tr>
</tbody>
</table>

Address of lead author:
Catherine Pelachaud
IUT de Montreuil
University of Paris 8
140 rue de la Nouvelle France
93100 Montreuil
France
Table of Contents

1 THE PLACE OF THIS REPORT WITHIN HUMAINE ............................................. 5

2 BRIEF OVERVIEW OF THE WORKPACKAGE, THE EXEMPLAR PROPOSAL, AND RELEVANT RESOURCES .............................................................. 7

2.1 The field covered by the workpackage .......................................................... 7
  2.1.1 Conception of the area before HUMAINE began ....................................... 7
  2.1.2 The current conception .............................................................................. 12

2.2 The exemplar proposal .................................................................................... 12
  2.2.1 The elements of the exemplar ................................................................. 13
  2.2.2 The ways in which the exemplar depends on other Workpackages .......... 16
  2.2.3 Proposed final output .............................................................................. 16

2.3 Related resources .......................................................................................... 16

3 RATIONALE FOR THE EXEMPLAR PROPOSAL .......................................... 17

3.1 Distinctive features of the approach proposed ............................................. 17

3.2 Rationale for emphasising this approach ..................................................... 17
  3.2.1 Perception Domain .................................................................................. 18
  3.2.2 Interaction Domain .................................................................................. 18
  3.2.3 Generation domain .................................................................................. 19

3.3 Rationale for subdividing the task ................................................................. 19

3.4 Measures taken to ensure coherence across subtasks ................................... 20

3.5 Measures taken to ensure coherence with other exemplars ......................... 20

4 TECHNICAL ASPECTS OF THE PROGRAM OF RESEARCH ..................... 21

4.1 Methodology overview .................................................................................. 21

4.2 Evaluation methodology and tools .................................................................. 26

5 DESCRIPTION OF CAPABILITIES ............................................................... 28

5.1 Capability 1: Cognitive influences on action .............................................. 28
  5.1.1 Emotion related attention shifts .............................................................. 28
5.1.2 Adapt politeness behaviours to the user’s emotional state................................. 29

5.2 Capability 2: Creating Affective Awareness.......................................................... 31
  5.2.1 Creating Affective Bonds.................................................................................. 31
  5.2.2 Imitation......................................................................................................... 33
  5.2.3 Adaptation....................................................................................................... 34

5.3 Capability 3: Backchanneling ................................................................................ 35
  5.3.1 Description..................................................................................................... 35
  5.3.2 Details........................................................................................................... 35

5.4 Capability 4: Coordination of signs in multiple modalities.................................... 35
  5.4.1 From multimodal emotional corpora to models of coordination between modalities ......................................................................................................................... 36
  5.4.2 Multimodal behaviour..................................................................................... 37
  5.4.3 Gesture repositories....................................................................................... 38

5.5 Capability 5: Expressivity .................................................................................... 39
  5.5.1 Behaviour expressivity.................................................................................. 39
  5.5.2 Speech Expressivity..................................................................................... 40
  5.5.3 Context dependent emotional body gesture ................................................. 41

6 PROPOSED ALLOCATION OF TASKS................................................................. 43
  6.1 Task 1: Representation Languages and Theoretical Framework ....................... 43
  6.2 Task 2: Mock-up of each capability.................................................................. 43
  6.3 Task 3: Proof-of-concept of each capability..................................................... 43

7 PROVISIONAL TIMELINE .................................................................................... 44
  7.1 Task 1: Representation Languages and Theoretical Framework ....................... 44
  7.2 Task 2: Mock-up of each capability.................................................................. 44
  7.3 Task 3: Proof-of-concept of each capability..................................................... 44

8 REFERENCES........................................................................................................ 45
1 The place of this report within HUMAINE

The HUMAINE Technical Annex identifies a common pattern that is followed by most of the project’s workpackages:

The measure of success will be the ability to generate a piece of work in each of the areas which exemplifies how a key problem in the area can be solved in a principled way; and which also demonstrates how work focused on that area can integrate with work focused on the other areas. We call these pieces of work exemplars. The exact form of an exemplar is not prespecified: it may be a working system, but it might also be a well-developed design, or a representational system, or a method for user-centred design.

To that end, each thematic group will work out a proposal for common action, embodied in one or more exemplars to be built during the second half of the funding period.

The process will begin with production by each thematic group of a review of key concepts achievements and problems in its thematic area; and drawn from the review, an assessment of the key development goals in the area. This review and assessment will be circulated to the whole network for discussion and comment, aimed both at building understanding of basic issues across areas, and at identifying the choices of goal that would be most likely let the different groups achieve complementary developments. That consultation phase will provide the basis for deliverables in month 11, which describe in some detail a few alternatives that might realistically be chosen as exemplars in each area, and their linkages to issues in other thematic areas. A decision and planning period will follow, involving consultation within and between thematic areas, leading to presentations at the second plenary conference, which will describe a single exemplar that has been chosen for development in each area, and the way work on the exemplar will be divided across institutions. The remainder of the project will be absorbed in developing the chosen exemplar.

The review and assessment documents were delivered in May, and the consultation phase has been ongoing since, using several channels, notably e-mail exchanges moderated by the co-ordinator; meetings of workpackage leaders by teleconferencing; meetings between workpackage representatives attending the WP4 workshop and the Summer School; and a consultation meeting of WP leaders in Paris on October 29th & 30th.

This deliverable is one of the group arising from the consultation phase, whose function is defined as to ‘describe in some detail a few alternatives that might realistically be chosen as exemplars in each area’. In general, we believe that we have progressed more quickly than we have expected, and that the alternatives described here are close to the ones that should be pursued. What remains to be completed is largely detailed planning. Given the intricacy of the network, that is not a trivial task.
The following persons have contributed to the deliverable:

Elisabeth André, Arnaud Blanchard, Gaspard Breton, Felix Burkhardt, Stephan Busemann, Antonio Camurri, Lola Cañamero, Ailbhe Ní Chasaide, Winand Dittrich, Christer Gobl, Stefanos Kollias, Kostas Karpouzis, Brigitte Krenn, Valérie Maffio, Maurizio Mancini, Jean-Claude Martin, Catherine Pelachaud, Danielle Pelé, Christopher Peters, Matthias Rehm, Olivier Rosec, Marc Schröder, Daniel Thalmann, Frederic Vexo, Gulatiero Volpe, and WP6 members

The institutions that have contributed are:

DFKI
DIST - University of Genova
EPFL
France-Telecom
ICCS
Limsi CNRS
OFAI
T-systems
Trinity College of Dublin
University of Ausburg
University of Geneva
University of Hertfordshire
University of Paris 8
2 Brief overview of the workpackage, the exemplar proposal, and relevant resources

2.1 The field covered by the workpackage

The area covered by this workpackage is described in the Technical Annex, particularly in Section 6.2, and in more depth in the review and assessment document for the workpackage. We summarise the area here partly so that the deliverable can be read as a stand-alone document, and partly to draw attention to changes of emphasis that have taken place during the first period of HUMAINE.

The core WP6 exemplar is to propose the definition of an Affective Interactive Embodied Conversational Agent (ECA) system with capabilities beyond those of present day ECAs.

As a long term goal the ECA needs to perceive the user’s emotional state and the context; she has to react to the user’s action and emotional states as well as to the events happening in the context; she also needs to show emotions, to adapt her behaviour to the user behaviour. She needs to catch the user’s attention, to maintain it while she is conversing with the user… At a high level the ECA should be active in the following domains:

- Perception Domain:
  - Agent perception of the users / other agents / world.
- Interaction Domain:
  - Interactants model of speaker(s) and listener(s)
  - Create affective awareness
- Generation Domain:
  - Produce dynamic expressive visual and auditory behaviours
  - Achieve coordination of signs of emotion in multiple modalities

2.1.1 Conception of the area before HUMAINE began

2.1.1.1 Perception Domain

In the perception domain, we are primarily concerned with work in two fundamental areas: perception by the user of the emotional displays of the system (see Section 2.1.1.1.1) and perception by the agent of stimuli in virtual and mixed reality environments (see Section 2.1.1.1.2).

2.1.1.1 Recognizing the emotional display of a system

To have an impact on the interaction with the user, the emotional display of a system must be recognizable by the user as such. A distinction must be made between the aim to provide life-
like or stylised emotional actions. Oudeyer (2003) presents a robot that is able to express its emotions by modulating the intonation of its voice. Lester et al (2000) present the pedagogical agent Cosmo that exhibits stylised emotional displays for specific situations like congratulatory acts. Paiva et al (2003) show in their SenToy environment that users are able to interpret and reason about the emotions of a computational system. The SenToy allows the user to influence the emotional state of game characters by the means of a tangible interface.

2.1.1.2 Perceptual Attention

Embodied agent perception and important related mechanisms, in particular attention, represent increasingly important areas of research for the development of plausible agents. The basis of many visual attention models is the feature integration theory of Treisman et al. (1986) that was derived from visual search experiments. According to this theory, features are registered early, automatically and in parallel along a number of separable dimensions (e.g. intensity, colour, orientation, size, shape etc.). Koch & Ullman (1985) have suggested a model based on this theory that leads to the generation of a master saliency map that encodes the saliency of image regions. Meaningful objects (conjunction of features) are identified at a second stage, which requires focused attention. Consequently, at the interface between the first and second stages there should be a bottleneck functioning as a gate allowing only part of the visual information to proceed to the second stage. Itti & Koch (1998, 2000) proposed the first computational model of saliency-based visual attention.

If the outputs of an embodied agent are desired to be humanlike, then the start of the process, i.e. the process of providing inputs to the system, should also be modelled in a way analogous to human processes. A number of authors have adopted the approach of attempting to create real-time systems that model human perception, even if in a crude way, in order to provide agent with a means to perceive their environment. Renault et al. (1990) and Noser et al. (1995) used synthetic perception and memory for solving an agent navigation task. Renault et al. (1990) introduce a synthetic vision system for the high level animation of agents. Noser et al. extend this model by adding memory and learning mechanisms. Kuffner and Latombe (1999) also present a perception-based navigation system for animated characters based on agents’ visual perception and memory. Hill (1999) provides a model of perceptual attention in order to create virtual human pilots for military simulations where perceived objects are grouped according to various criteria, such as type and proximity. A visual system provides a feedback loop to the overall navigation strategy. Musse and Evers (2002) have presented a system for the storage of an agent’s perception of a city environment. The memory model keeps a record of past experiences by using dynamic and hierarchical finite state machines to store a list of behaviours that were previously executed by the virtual human. Peters and O’Sullivan (2002) use a synthetic vision system linked with a computational model of memory to filter and store scene objects that are perceived by an embodied agent. Finally, Rapantzikos et al. (2004) introduce salient volumes as the means to locate and track attention-related shifts over time. Of particular note regarding contemporary work is that the systems involving perception, attention and memory for embodied agents have generally used them in a disparate fashion for specific applications, usually for the purposes of solving navigation problems.

2.1.1.2 Interaction domain

In the interaction domain, systems capable of exhibiting non-verbal as well as multimodal behaviours, including speech (Section 2.1.1.2.1), the ability for agents to attend (Section 2.1.1.2.2), react to user emotion and account for social context (Section 2.1.1.2.3) are all of relevance to our work.
2.1.1.2.1 Non-verbal and multimodal systems

Cassell and Vilhjálmsson (1999) present the BodyChat system, which automates the animation of communicative behaviours based on context analysis and discourse theory. Cassell et al. (2001) have also presented BEAT, a toolkit for the animation of expressive behaviours. The toolkit automatically selects appropriate gestures, facial expressions, pauses for an input text and provides synchronisation information that is necessary to animate the behaviours in conjunction with the character’s speech. Poggi et al. (2000) generate coordinated linguistic and nonverbal communication for the simulation of a dialog between two agents. The Olga project (Beskow, 1997) integrates conversational spoken dialogue (i.e., speech recognition and natural language understanding), 3-D animated facial expressions, gestures, lip-synchronized audiovisual speech synthesis, and a direct manipulation interface. In the human-human communicative process, verbal and nonverbal signals are active. Takeuchi and Naito (1995) introduced the notion of "situated facial displays." Situatedness means that the system not only follows its internal logic but is also affected by external events. External events include reactions of users, arrival of a new user interacting with the system, actions done by one of the users, and so forth. The NECA project (Krenn et al., 2002) developed web-enabled interactions between ECAs. Dialogues are generated by a sequential modular architecture consisting of scene generation (including an affective reasoning engine), multimodal natural language generation (generating verbal content and meaning-carrying gestures), speech synthesis, gesture alignment, and player-specific rendering.

The DIST – InfoMus Lab (Camurri et al., 2000, 2005) developed the EyesWeb open platform with a special focus on the multimodal analysis and processing of non-verbal expressive gesture in human movement and music signals. EyesWeb supports the user in experimenting computational models of non-verbal expressive communication and in mapping, at different levels, gestures from different modalities (e.g., human full-body movement, music) onto real-time generation of multimedia output (e.g., sound, music, visual media, mobile scenery). EyesWeb has been successfully used for a number of applications mainly in the performing arts. Nevertheless, it provides both a platform and a collection of software modules for real-time analysis of expressive gesture that can also be employed in the design and development of ECAs.

At UH, Arnaud Blanchard and Lola Cañamero are investigating the role of affect in low-level imitation using robots. By low-level imitation we understand rather automatic processes of synchronisation and coordination of behaviours that do not involve an explicit “intention” to imitate nor reasoning about the behaviour of the other agent, its mental, or affective state. An example of low-level imitation related to affect is the phenomenon of emotional contagion (Hatfield, 1994), in which an agent mimics automatically the expressive behaviour of another and, by doing so, enters in a similar (internal) affective state. Other roles that affect play in low-level imitation include influencing the choice of what, when, why and to whom to imitate or not to imitate based on affective bonds developed during the interaction.

Multimodal speech systems and animated agents have been created for the personalization of user interfaces (Cassell et al. 1994; Chopra-Khullar and Badler 1999; Pelachaud and Prevost 1994; Rist, André, and Müller 1997) and for pedagogical tasks (Badler et al 2000; Lester et al. 2000; Rickel and Johnson 1999), exhibiting nonverbal behaviours such as pointing gestures, gaze, and communicative facial expressions. The links between facial expression and intonation (Pelachaud and Prevost 1994) and between facial expression and dialogue situation (Cassell et al. 1994; Cassell et al. 1999; Thórisson 1997, Krahmer et al 2003) have been studied, and a method to compute automatically some of the facial expressions and head movements performing syntactic and dialogic functions has been proposed.
2.1.1.2.2 Environmental and agent awareness

Attention models have been used several times as a basis for providing automatic agent behaviours in specific situations, ranging from virtual helicopter pilots in combat environments (Hill, 1999) to conversational agents in social settings (Cassell & Vilhjálmsson, 1999; Poggi et al., 2000) to the modelling of more general attention behaviours during free viewing conditions in an environment (Chopra-Kullar, 1999; Gillies, 2001; Peters and O'Sullivan, 2003). Free viewing conditions may entail an agent looking at certain objects or locations of interest while standing in a natural environment when it is not considered to be engaged in any particular task. Emotional agents and social context

Several ‘talking heads’ capable of showing emotions to varying degrees have been developed. In particular, Kshirsagar and her colleagues (2001) have developed an agent that is able to react to the user’s emotion detected through his facial expressions. This reaction is based on a computational model of emotion behaviour that integrates a personality model. Carmen’s Bright IDEAS (Marsella et al., 2001) is an interactive drama where characters exhibit gestures based on their emotional states and personality traits. Through a feedback mechanism a gesture made by a character may modulate her affective state. A model of coping behaviours has been developed by Marsella and Gratch (2003). The authors propose a model that embeds information such as the personality of the agent and his social role.

In relation to emotion agents, another important consideration is the role of social context in an agent’s behaviour. Poggi et al. (2001) propose a model that decides whether or not an agent will display her emotion depending on several contextual and personality factors. Prendinger et al (2002) integrate contextual variables, such as social distance, social power and threat, in their computation of the verbal and nonverbal behaviour of an agent. They propose a statistical model to compute the intensity of each behaviour. Rist and Schmitt (2003) modelled how social relationship and attitudes toward others affect the dynamism of an interaction between several agents.

Camurri and colleagues (1998, 1999) proposed an architecture modelling the emotional, rational, and reactive components of an emotional agent and the interactions between such components. The architecture was successfully employed for developing emotional agents in the context of interactive installations for children (“Città dei Bambini”, Genova, Italy).

2.1.1.3 Generation Domain

Relevant work in this area has focused on generating expressive behaviours (Section 2.1.1.3.1) and expressive audio and visual speech synthesis (Section 2.1.1.3.2).

2.1.1.3.1 Expressive Behaviour

Chi et al. (2000) have implemented Laban principles of Effort and Shape to render gesture performances more expressive by varying kinematic and spatial spline properties of existing keyframed actions. An extension (Byun & Badler, 2002) has been proposed to use Effort parameters to modify facial animation data streams in MPEG4 FAP format. Additionally it is attempted to tie the EMOTE system to OCC and OCEAN models of emotion and personality (Allbeck & Badler, 2002). Another recent approach using OCC and OCEAN models to drive agent facial animation is presented in (Egges et al., 2002). Barrientos (2002) proposes a mechanism for extracting expressive or emotional parameters from an analysis of handwriting and then using these parameters to control the gesturing of an agent through multi-step-interpolation between predefined animation clips. In the framework of dimensional emotion models, Raouzaïou et al. (2002) propose a way of modelling primary expressions using FAPs
and describe a rule-based technique for the synthesis of intermediate expressions. A relation between FAPs and the activation parameter proposed in classical psychological studies is established, extending the archetypal expression studies that the computer society has concentrated on. The overall scheme leads to a parameterised approach to facial expression analysis and synthesis that is compatible with the MPEG-4 standard. Fiume and Neff (2002) have presented a convincing simulation of muscle tension. Earlier notable examples of generating expressive movement include Perlin and Goldberg’s use of noise functions (Perlin & Goldberg, 1996), and Bruderlin et al.’s use of signal processing techniques to change frequency characteristics (Bruderlin et al., 1985).

2.1.1.3.2 Audio and Visual expressive speech synthesis

Previous work on emotional speech synthesis can be divided into two main strands. In the first strand, emotionality is recorded in concatenative synthesis databases (Iida et al., 2003; Johnson et al., 2002), which allows for fairly natural expressivity without any explicit modelling of the acoustic properties of emotions. As emotionality is only “played back” by concatenating units from a database containing speech material recorded with the targeted expression, this approach completely lacks flexibility. The second strand, which historically appeared first, aims to explicitly model the acoustic properties of emotional states, either in terms of emotion categories (Cahn, 1990; Burkhardt & Sendlmeier, 2000) or emotion dimensions (Schröder, 2004). While the parametrisable speech synthesis methods used in this second strand of research (formant or diphone synthesis) sound less natural than the corpus-based methods used in the first strand, this approach is scientifically more promising in the long term. Its success depends on the development of parametrisable synthesis methods that are at the same time natural-sounding.

Common to all existing approaches to emotional speech synthesis is their lack of dynamicity: Emotionality is considered a constant over an utterance. In most research (e.g., Murray & Arnott, 1995), the verbal content is even explicitly rendered unemotional, in order to isolate the perceptual effect of the prosodic settings. More recently, we have proposed to conceive of the speech prosody as a modality that should “fit with” the emotional message expressed through other modalities such as the verbal content of the spoken utterance (Schröder, 2004). This has been taken up by Hamza et al. (2004), who modelled speaking styles motivated by applications (“good news” vs. “bad news” speaking styles). What has not yet been addressed, though, is the dynamic change of emotional tone as a function of the semantics of the text.

Audio-visual synthesis of emotional speech has been addressed as the combination of a body or facial model with a speech synthesiser (e.g., De Carolis et al., 2004; Krenn et al., 2002). The notion of mutual influence and the possibility of compensation or contradiction between modalities has not yet been addressed.

Summary

Up to now the ECA research community has been focusing on ECAs with more or less sophisticated presentation facilities. ECAs are to a certain extent able:

- to compute what to say based on discourse plans
- to generate synchronized behaviours.
- to talk with emotional intonation
• to show some emotional expressions (the so-called ‘universal’ facial expressions of emotions);

• to perform gesture and body movement.

2.1.2 The current conception

Within WP6, we are particularly interested in the role of emotion in interaction. This encompasses abilities in all three domains, as an ECA must also be able to perceive and interpret emotion as well as show emotion in order to interact in a desirable emotional manner.

2.1.2.1 Perception Domain

In order to interact in proper manner, an ECA must first be able to pay attention to, perceive and interpret her conversational partners (e.g. the user) and the context she is placed in; only then can she hope to adapt her (verbal and nonverbal) behaviour depending on the user, her role, and the social and cultural context she is placed in. At a fundamental level, the ECA should have some notion of conversational partner, emotion, gesture and environment.

2.1.2.2 Interaction Domain

Interaction is the central theme of WP6. Thus, we are particularly interested in the role of emotion (including emotion proper, mood, interpersonal stance and attitude) in interaction, i.e., how to create an interactive ECA able to not only show and perceive emotion, but also adapt her behaviour to convey some emotion or to induce some emotional state in the human user. Interaction is therefore the bridge between perception and generation.

2.1.2.3 Generation Domain

Even though the presentation side of ECAs is the most developed in ECA research, further improvement of this area is required, especially as regards a) multimodal integration and the display of emotional behaviours to improve the naturalness of ECAs, and b) the believability of ECAs which is influenced by the consistency of an ECA’s behaviour be it in terms of personality, cultural context and situation. Even though decisions on the form of generation (e.g. a particular facial expression or hand gesture) are taken on the basis of the aforementioned domains, it is the generation part that provides the capability for a proof-of-concept implementation and testing of those concepts and essentially initiates interaction with the user or/and other ECAs.

2.2 The exemplar proposal

Following the consultation period, the exemplar proposed for WP6 is Definition of Affective Interactive Embodied Conversational Agents. Such ECAs should have several capabilities within the three considered domains:
• Perception domain:
  o Agent perception of the users / agents/ events
• Interaction domain:
  o Interactants model:
    • Listener model
    • Speaker model
  o Create affective awareness
• Generation domain:
  o produce dynamic expressive visual and auditory behaviours
  o achieve coordination of signs of emotion in multiple modalities

The core task of WP6 exemplar is to work out a set of capabilities for an affective ECA to have in order to be able to sustain an interaction with a user or another ECA as well as to be able to perceive, adapt and generate emotions. These capabilities require broad and complex constructs such as social and communicative context, models of memory, of personality, of appraisal and the like. Figure 1 illustrates this framework.

![General Framework Diagram](image_url)

**Figure 1 General Framework**

### 2.2.1 The elements of the exemplar

Capabilities are the key elements of the WP6 exemplar. Several capabilities have been envisaged as being essential to the definition of an Affective Interactive Embodied Conversational Agent.

1. **Cognitive influences on action**

   This capability is concerned with certain aspects related to cognition that may influence the actions of an agent, such as gaze behaviours or gestures. In particular, it investigates attention shifts accounting for emotional stimuli and also the adaptation of
an agent’s politeness behaviours to an emotional user. These issues are related to WP7 themes and collaborate with WP7 will take place. Here, we are interested more in the social aspects.

a. *Emotion related attention shifts*

Emotion related attention shifts are concerned with the guidance and allocation of attention by an agent to emotional or potentially important stimuli within a scene for the purposes of conversation control. These shifts may result in the production of visible phenomenon that are of significance to others in the environment.

b. *Adapt politeness behaviours to the user’s emotional state*

Since emotions have an important impact on perceived face threat, here we seek to create an ECA that can adapt to the user by trying to mitigate such threats through use of gesture, mimics and speech and will also account for the causes of such emotions, rather than just their strengths.

2. *Creating Affective Awareness*

In this capability, we seek to create connections between a user and an ECA that will maintain user engagement in an interaction. Such ECAs should not only perceive the users emotion, but also adapt to it, react to it and imitate it.

a. *Creating Affective Bonds*

The creation of a bond between the user and ECA is investigated through monitoring the users level of engagement with an ECA and also by providing an explicit means for the ECA to try influence the users level of engagement in the interaction.

b. *Imitation*

Imitation allows the creation of an affective “awareness” and coordination, and to alter affective bonds by allowing agents to be capable of mimicking certain aspects of another, so that synchrony of behaviour may in some circumstances also lead to a synchronisation of emotions.

c. *Adaptation*

In order for an agent to be adaptive, information about the emotional state of the user(s) must be obtained. Here, expressive gestures are analysed to infer a users emotional state – the inferred state is then be used to alter an ECA’s reactions.

3. *Backchanneling*

ECAs need to become interaction partners. A key feature of spoken dialogue is backchanneling by the listener, providing the speaker with immediate feedback about the state of the communication. Backchannel utterances can convey simple information of critical importance to the success of the communication attempt, e.g., whether the addressee is still listening, understands, or agrees with the spoken content.
4. **Coordination of signs in multiple modalities**

In order to be believable, ECAs must show expressivity in a consistent and natural looking way across modalities. The ECA has to be able to display coordinated signs of emotion during realistic emotional behaviour. Such a capability requires to study and represent emotions and coordination of modalities during non basic realistic human behaviour, to define languages for representing such behaviours to be displayed by the ECA, to have access to mono-modal representations such as gesture repositories.

   a. *From multimodal emotional corpora to models of coordination between modalities*

   We aim at building a model of the coordination of multimodal behaviour during non basic multimodal behaviour observed in non acted data, using multimodal corpora.

   b. *Multimodal behaviour*

   We need to represent the relationships between the signs of emotion in different modalities, i.e., by designing a representation language.

   c. *Gesture repositories*

   An important precondition for generating appropriate multimodal behaviour in ECAs is the availability of gesture repositories.

5. **Expressivity**

Emotion is not simply expressed through a static facial expression or setting of vocal cords. Acoustic and visual expressions of emotion is dynamic; they evolve through time. The manner a movement is done provides relevant information on the affective state of the emitter. In this capability we aimed to consider the quality aspect of behaviour.

   a. *Behaviour expressivity*

   We aim to investigate the generation of phenomenologically accurate behaviours as we focus on developing an intermediate level of behaviour parameterisation, a set of dimensions of expressivity.

   b. *Speech Expressivity*

   We want to better understand how emotions and related states are expressed through the voice, and how the expression through the corresponding acoustic parameters can be achieved using speech synthesis techniques.

   c. *Context dependent emotional body gesture*

   We aim to provide ECA’s with the capability of modulating its body gestures (in particular upper-body: head, shoulders, arms, hands) according to the emotional state induced by the context –situation– in a virtual environment.
2.2.2 The ways in which the exemplar depends on other Workpackages

- WP3: appraisal model; defining the proper set of parameters to characterize expressive behaviour; which behaviours are relevant to convey emotion.
- WP4: analysis of audio and visual signals; analysis and recognition of user’s emotional state; analysis of the dynamic properties of user’s behaviours.
- WP5: collecting and labelling data; defining what to code and how to code it.
- WP7: models of planning, attention and decision making; architectures underlying low-level imitation and the influence of affect in it.
- WP8: models of engagement.
- WP9: for designing a given application, we will use user-centred design method in direct link with WP9.
- WP10: question of user’s manipulation, of invading user’s space....

2.2.3 Proposed final output

WP6 will conceive several proofs of concept of the different capabilities presented in this document. Moreover WP6 will work out an E-book. The outcome of this e-book is to serve the ECA community to further improve their system. This book will gather chapters describing the works done toward the creation of Interactive Affective ECAs and will also integrate multimedia materials that illustrate the presented works: movie clips, figures, audio clips but also pieces of software, detailed worked out examples. A section of the book will be dedicated to the representation language and the library of non-verbal behaviour, the gesticon.

2.3 Related resources

We will be using extensively concepts as well as we will refer to different techniques presented in other deliverables. In particular, we will refer to:

- WP3: concepts linked to the appraisal model; description of response patterning and of signalling of quality and intensity over all modalities
- WP4: visual and audio feature sequences based on acoustic and visual signals and signs analysis; means to perceive expressed emotions and salient events based on signs
- WP5: issues related to the elaboration of corpus and emotion labelling; corpus with automatic and manual annotation of signals in different modalities
- WP7: issues related to cognition and how emotions affect planning and decision making
- WP8: issues related to inducing emotions to users; relation of emotional activation and emotional decision-making with persuasion
- WP9: issues related to user-centred evaluation techniques
3 Rationale for the exemplar proposal

The exemplar proposal represents a choice to follow a particular line of development rather than others that are possible (or might seem possible to the outsider). The key reasons for making this particular choice are as follows.

3.1 Distinctive features of the approach proposed

Up to now the ECA research community has been focusing on ECAs with more or less sophisticated presentation facilities. ECAs are to a certain extent able:

- to compute what to say based on discourse plans
- to generate synchronized behaviours
- to talk with emotional appropriate intonation
- to show some emotional expressions (the so-called ‘universal’ facial expressions of emotions)
- to perform gesture and body movement

Even though the presentation side of ECAs is the most developed in ECA research, further improvement of this area is required, especially as regards a) multimodal integration and the display of emotional behaviours to improve the naturalness of ECAs, and b) the believability of ECAs which is influenced by the consistency of an ECA’s behaviour be it in terms of personality, cultural context and situation.

3.2 Rationale for emphasising this approach

The WP6 exemplar is not aiming at developing an integrated ECA system fully equipped with perception, interaction and generation capabilities. Rather it aims at defining relevant capabilities an ECA should have to fulfil the need to embed “emotions in interaction”. To provide with a clear vision on how these capabilities fit in an ECA system, we will first work out what is the theoretical framework of such an ECA system. The framework will reflect the desired modular architecture needed to support emotions in interaction. It is not intended to specify this architecture at a great level of technical details as far as system integration is concerned; rather it is aimed at establishing precisely which modules are required and how they are linked with each other.

Even if we are not going to develop an integrated ECA, results from the joint research can potentially produce an advance both in the state of the art of ECA research and, in general, to research on multimodal (affective) interactive systems. ECAs are thus selected as a paradigmatic example of multimodal interactive system: they encompass fundamental research issues that are common to many kinds of multimodal interactive systems (e.g., systems for edutainment, entertainment, performing arts, therapy and rehabilitation, …). Therefore, in the network, building a shared knowledge on ECAs and carrying out research and experiments on them is likely to produce progresses not only with respect to ECA research but also in many other applications scenarios the HUMAINE partners are involved in.
3.2.1 Perception Domain

3.2.1.1 Perception by the system of human emotional behaviour

A key issue is the investigation of expressive gesture as a first-class conveyor of expressive, emotional content. Analysis of expressive gesture can provide ECAs with the capability of extracting information about the emotional behaviour of the users. Advances in expressive gesture processing can be obtained:

- By defining and carrying out experiments focusing on specific aspects of non-verbal communication and interaction through expressive gesture: such experiments can contribute in creating the theoretical basis and in providing some first experimental evidence on the role of expressive gesture processing in interaction.

- By defining models and algorithms aiming at extracting expressive features and at interpreting expressive content, even if restricted to the test-bed scenarios on which the experiments above focused.

In this perspective, one main issue is to find how to use a corpus of human multimodal emotional behaviour to carry out experiments, identify and analyse expressive features, testing and evaluating the developed algorithms.

3.2.2 Interaction Domain

3.2.2.1 Agent Interpretation of discourse signals

In order to participate in discourse in a plausible manner, agents must be able to interpret signals made by others; only then should such signals be processed and stored internally. This involves more than a reversal of the problem of mapping emotions into animation parameters, as interpretation is dependant not only on the external signals being conducted, but also modulation by internal factors particular to the agent, such as the agents relationship with the signaller and personal traits, for example, optimism or pessimism. The active interpretation of signals should also imply an enhanced role for the listener during discourse, possibly providing listeners with more active roles during discourse planning (DeCarolis et al., 2002).

3.2.2.2 Attention to emotion

Social attention for speakers and listeners based on facial expressions during social interaction is an important aspect in modelling conversational behaviour. During conversation, the speaker may pay attention to the faces of listeners in order to obtain feedback of listener agreement / empathy. The role of emotional competence in expressions of others should be accounted for, not only in speaker to listener attention, but also in listener to speaker, and listener to listener attention. Therefore, we view an emotion-based conversational attention construct as an important prerequisite to the interpretation of signals for discourse. Furthermore, facial reactions in the listeners to the speaker (e.g. mimicry to express empathy) may also be important for accounting for plausible viewer perception of interacting ECAs. When people are exposed to emotional facial expressions, they spontaneously react with emotion-relevant facial muscles that tend to mimic the facial stimuli.
3.2.3 Generation domain

3.2.3.1 Behaviour expressivity

We seek to provide a computational model to capture expressivity in human behaviour. Expressive behaviour is defined both by the signal chosen to convey the meaning of the communicative act and the quality or form of execution. We want to address the question of how visible behaviours vary in relation to emotional intensity. How are signals affected for a particular modality? And how are signals synchronized across modalities? In essence, we are looking to establish a mapping between the intensity of motion on the one side and the qualitative signal composition of an expression on the other hand.

3.2.3.2 Emotional Speech Modelling

Existing research has proven that extreme emotions can be expressed in speech synthesis sufficiently well to allow for their distinction by listeners based solely on the vocal parameters. This, however, is not what is required in applications, where the tone of voice primarily needs to be consistent with the situational context, the verbal content, and visual channels through which emotions are being expressed.

At least two choices exist for modelling nuances of emotion:

- Modelling of general emotional properties: the quality of such modelling needs to be improved along two axes: 1) models of the acoustic correlates of low-intensity emotions (including the ability to capture the acoustic correlates of smiled speech); 2) signal manipulation and corpus-based technologies for realising these models during speech generation.

- Modelling of low-intensity emotion-related states specific to a given application domain: For a given application domain, it may be more beneficial to model as closely as possible the types of emotion-related states relevant for that domain.

Voice quality is undoubtedly a major carrier of emotion. Listeners are highly sensitive to the speaker’s ‘tone of voice’. Being able to recognise those voice quality modulations that constitute tone of voice, being able to interpret their communicative functions and being able to regenerate them through synthesis are capabilities we would eventually wish to incorporate in emotion sensitive man-machine systems.

These goals are currently a good distance from potential realisation. Voice quality is the most elusive aspect of the speech signal to capture and the least well understood. In comparison to f0 and amplitude dynamics or to temporal features which have been fairly extensively studied for emotive speech (e.g., by Scherer and colleagues) there is little quantitative information available. Most of the analytic research has been directed at the pathological voice, and deals with very limited aspects of voice quality, concentrating either on measures of jitter and shimmer or on measures taken from sustained phonation of a single vowel.

We will focus on this dimension of the speech signal and on how it combines with other dimensions of the speech signal (f0 and temporal features) in signalling emotion.

3.3 Rationale for subdividing the task

The core task of WP6 exemplar is to work out a set of capabilities for an affective ECA to have in order to be able to sustain an interaction with a user or another ECA as well as to be
able to perceive, adapt and generate emotions. Each of the capabilities has a narrower scope than building a full ECA system. Their purpose is to gain practical feedback on theoretical considerations.

Working on individual capabilities rather than integrating all efforts into building one affective ECA allows us to go into more depth in each theme. Moreover using analytical approach as well as modularity allows reducing complexity by assuming independence of subsystems. Section 4 will provide more details on this aspect.

3.4 Measures taken to ensure coherence across subtasks

There is a strong need of dialog reinforcement between groups working on the different capabilities. This need arises from common research topics or common practicability questions across working groups. Indeed, we have seen that the specification of representation languages is not only a key issue for mock-up implementation, but these languages in the first place represent a means for encoding communicative information and using it to drive and control the animation of ECAs. The encoded information may cross over multiple modalities e.g. correspond to gesture, gaze and face signs. It may also represent semantic or syntactic elements. Thus representation languages are a critical issue regarding any work involving expressive agents. A particular task will be devoted to the specification of these representation languages.

As working methodology, we plan to coordinate the work of WP6 partners by continuing using the already available tools in the HUMAINE portal (mailing lists, forum, shared repository, …). Moreover several meetings are planned (e.g., the WP6 workshop to be held in Paris in early March). Exchange of personnel for short periods among partners are foreseen. Some are already scheduled to happen in the spring and summer.

3.5 Measures taken to ensure coherence with other exemplars

The representation language functions not only as integrative means for the capabilities at the core of the WP6 exemplar, but has also links to other exemplars, for instance:

Defining the representation language we will also address the requirements for the representation language to function as interface to wrappers for various player technologies such as the Greta player developed at Paris 8, Cantoche’s LivingActor technology, or the Haptek player employed in the Wizard of Oz Tool utilized in the WP9 exemplar.

We will also define interfaces to the annotation schemes developed for the annotation of emotion data in WP5. This way we will be able on the one hand to ensure coherence in the description of emotions, and on the other hand do not overload the representation language with information irrelevant at the interfaces of system components.
4 Technical Aspects of the Program of Research

4.1 Methodology overview

The WP6 exemplar will seek to carry out the following tasks:

- **Theoretical Framework:** Design of overall architecture for ECA system, that can perceive, interact and generate, based on:
  - comparison of already existing systems
  - requirement analysis for
    - real-time activity
    - embedment of ambient environment
    - elaboration of the clear conceptual links between the perception, interaction and generation domains

- **Elaborate a set of capabilities for an Affective Interactive ECA:** each instantiation will represent the mock-up development of a sub-system enabling a capability an ECA should have in order to be able to conduct emotional interaction

- **Proof-of-concept:** instantiation of sub-systems of the theoretical framework

At the highest level, the theoretical framework covers all three domains: perception, interaction and generation (see Figure 2 a). The capabilities described above fit into this framework in a theoretical manner by crossing over one or more of the domains; some capabilities will be active in all three domains, others in only one domain. Capabilities also represent the link between theory and practice; we are not aiming at a fully integrated system and no integration of systems is foreseen within WP6. Rather, a mock-up version of each capability will be described, where certain subparts of a capability may be modelled as ‘black-boxes’ or elaborated to different extents, as required. Representation languages are the ‘glue’ with which capabilities are fitted into the framework across the domains.

Figure 2 visualizes the set up of the WP6 exemplar. The theoretical framework will cover all three domains (perception, interaction, generation) (see Figure 2 a). We will elaborate a set of modules that constitute it (Figure 2 b). The corresponding architecture will not be linked to any particular capability. Rather each capability will be defined as a set of modules of the main architecture. Capabilities could share identical modules (see Figure 2 c).

We are not aiming at an integrated system. Rather we will work out a mock-up version of each capability. Modules defining a capability will be linked with each other through representation languages (see Figure 2 d). So modules of a sub-system enabling a capability will interact with the other modules using a file-based interface (see Figure 2 d). This requires the definition of a set of languages that specifies the format of input and output required by the modules. Using such an interface between the modules allows an easy and fast integration of modules.
The capabilities listed above will give rise to the definition of sub-systems that will be defined as subsets of the overall architecture (see Figure 3 c). This work will be carried out by groups gathering several partners of WP6. As it has been stressed several times, no integration of systems is foreseen within WP6. Therefore the development of the systems corresponding to the capabilities will be done through a mock-up implementation. Finally if a capability has one or more proof(s) of concept, they will contribute to the evaluation of the sub-system of this capability. Not all the modules will necessarily be elaborated; mock-up modules will also be possible (see Figure 3 c).

**Theoretical Framework**

![Diagram](image)

Figure 2a-c: The theoretical framework is (a) distributed across the three domains; (b) each domain is conceptually composed of a set of modules; (c) Capabilities are defined within this framework.
Figure 2d-e: The means of communication between the modules is (d) a set of formally specified representation languages; (e) file-based interfaces will be used as a low-effort integration of modules.
Figure 3: We distinguish (a) mock-up and elaborated modules; within (b) the overall theoretical architecture, we identify (c) a subsystem realising each capability, which will be implemented as a mock-up subsystem and as several proof-of-concept subsystems.
Key Terms

A number of terms are used in this deliverable in a very specific sense. The following glossary is to explain the intended meaning.

- **Theoretical specification of overall architecture**: A detailed description of the modules realising the ECA, and of the information flow between them. We envisage this architecture to be modular, i.e. to consist of identifiable modules with clear interfaces between them. The specification will be a detailed theoretical description (which is sufficiently detailed to allow for a mock-up implementation).

- **Modules**: Identifiable parts of the overall architecture, with defined input and output interfaces, including the form and the meaning of data read in and written out. A module has a clear functional role, e.g. Multimodal data fusion, or speech synthesis.

- **Sub-system**: A number of modules linked together as specified by the overall architecture. A sub-system is typically defined by the capability that it is able to instantiate.

- **Representation language**: A formal definition of a data representation format, e.g. Using XML, which serves as the format in which data is encoded when passed between the modules in the architecture. A full specification of a suitable representation language is planned.

- **Mock-up implementation**: A very simple implementation of (a sub-system of) the specified architecture. The mock-up implementation consists of mock-up modules communicating via file-based interfaces. The data passed between the mock-up modules is in the form as defined by the representation language, and has the meaning as defined by the architecture specification.

- **File-based interfaces**: A very low-tech level of integration of the modules. For example, one module writes its output into a file, and the next module reads its input from that file. In particular, the two modules do not need to run on the same system, nor does their integration need to be done in real time.

- **Capability**: Abstract specification of new kind of behaviour of the ECA system, which it was not able to perform up to now. Such a capability is desirable for increased believability of the ECA in the context of emotional interaction. The capability is realised as at least a mock-up implementation of a sub-system of the overall architecture. This implies the identification of such a suitable sub-system and the modules that it is composed of.

- **Proof-of-concept implementation of a sub-system providing a capability**: The level of implementation planned for the different capabilities. This goes beyond a mere mock-up implementation, but will also not provide a fully functional sub-system. Some key features of important modules for the respective capability will be implemented, while other less central modules in the sub-system remain in a mock-up state.
4.2 Evaluation methodology and tools

Using ECAs or virtual environments are widely believed to induce a unique experience in users. However, standard human–computer interaction (HCI) usability evaluation methods (e.g., usability inspection or empirical user studies) do not address the vicarious nature of activities performed through either a cognitive (point of view: e.g. cognitive requirements and efforts by the user) or social perspective (point of view: e.g. users' emotional and social requirements and constraints) within a VE or ECA space. Human-Computer interactions can be seen as a challenge to human cognitive abilities maximizing gains of valuable information per unit cost associated to their efforts in engaging with this new technology. On the other hand, looking at ECAs as a social phenomenon changes the focus from information gathering to social processes such as communication and forms of interactions. From this perspective, and this is the main focus of WP6, ECAs will create a new psychological space or experience, yet to be explored and evaluated.

Although no consensus definition of this new experience yet exists, one may describe it as “being there” or “interacting with”, almost illusionary, experiences created by virtual space or ECAs. Referred to as presence, perception of embodiment or virtual interactivity, these are seen as a primary driver for VE design and evaluation, and consequently have yet to prompt much research in an attempt to elucidate its underlying determinants and find measures for their assessment and user satisfaction in terms of cognitive or social requirements. It remains an open question, for example, which of the experiences mentioned would be at the forefront of the human-computer interaction patterns related to virtual space or agents. Obviously, this is an example of the urgency of evaluation as continuing process feeding back into the development process in contrast to evaluation as a usability problem at a late stage.

In using professional evaluation, at least four parties or stakeholders seem involved:

(1) The professional producer: the exemplar developer and technology specialist,

(2) The testing professional (e.g. psychologist): who administer the testing procedure,

(3) The test taker,

(4) The user: the individual or organization that will eventually use the information to make certain decisions.

Although existing models do provide some guidance on what factors evaluators should attend to, for example based on cognitive architectures and models, and how they may go about doing so, no single model or framework is likely going to satisfactorily capture the complexity of technical, organizational, and psychological issues inherent with ECAs or virtual environments. These factors seem to interact with traditional implementation concerns in an mostly unknown way. Traditionally, what is referred to as four parties is seen as four stages in a serial usability model. Grounded in this analysis, a theoretical model of exemplar implementation and evaluation should be adopted and articulated and contrasted with other views of implementation based on usability and design questions. In this sense, it seems essential to regard evaluation as a continuing process relevant at all stages of the developing process and not just a final addendum to development. Therefore, evaluation and testing in respect to psychological parameters or models might be considered at an early stage. Not the least, as it seems widely accepted that formal models of human-computer interactions related to virtual or ECAs environments seem to suffer from the inherent ambiguity and unpredictability of behaviours in very open tasks such as envisaged in the present project. Such technical and human factors problems add a layer to the already overwhelming set of...
social constructs to understand when trying to simulate social interactions (e.g. ECAs) more or less. One of the most challenging aspects of social interactions is the diffusion of mental representations among agents. Specifically this cognitive aspect needs continuous checking and evaluation in real life, even more so in virtual environments. These cognitive aspects will most certainly form severe constraints for the development of ECAs, for example. Again, early evaluation of such aspects when implementing interactive abilities in ECAs might be at the heart of a resource-economic strategy. It is anticipated that the various evaluation and testing phases are based on empirical and experimental techniques widely used in psychological testing. Detailed specifications seem premature at this stage but it is suggested that the testing model will follow the clinical testing model.

Although these may be satisfactory criteria for the evaluation of the psychological, e.g. cognitive and social, factors, an evaluation framework has to appreciate the value of mixing quantitative and qualitative methodologies when evaluating exemplars such as ECA prototypes. Therefore, different mixes of quantitative and qualitative methodologies are suggested. It is anticipated that by mixing methods there is greater potential for capturing and understanding the richness and complexity of virtual or ECAs environments than if either approach is used solely. While some evaluators may argue against mixing research paradigms, here, we support Tashakkori and Teddlie’s (1998) pragmatic stance. They stress the importance and predominance of the research question over the paradigm. This approach frees the evaluator to choose whatever methods are most appropriate to test the exemplar once it is articulated or specified sufficiently.

The approach described here illustrate the wide range of evaluation tools and methodologies available to the evaluator and the need to adopt the methodology to meet the requirements of stakeholders. A final more detailed evaluation programme will be based on the specifications of the exemplar by adopting the current framework of evaluation and testing.

Finally, can the costs be contained seems always an important question. On the basis of a science-based evaluative study, one will be able to provide the stakeholders with sufficient data to demonstrate the feasibility of a prototypical ECA or exemplar and the users' psychological requirements. Only through continual experimentation with various evaluation strategies and new tools, evaluators will be able to contribute significantly to our understanding of the role of emotions, including cognitive and social factors, in this unique new experience when interacting to various degrees with virtual agents or environments.
5 Description of Capabilities

5.1 Capability 1: Cognitive influences on action

This capability is concerned with certain aspects related to cognition that may influence the actions of an agent and investigates attention shifts accounting for emotional stimuli and the adaptation of an agent’s politeness behaviours to an emotional user. Although these issues are related to WP7 themes and will involve collaboration with that work package, here we are oriented more towards the social aspects of cognition applicable to ECAs.

5.1.1 Emotion related attention shifts

5.1.1.1 Description

The guidance and allocation of attention by an agent to a scene is of great importance, as it may lead to visible behaviours such as saccadic eye movements and gaze reorientation, which may further alter perception or trigger higher-level actions. This part of the capability is concerned with the guidance of an agent’s attention to emotional or potentially important stimuli in the environment, in particular other agents, for the purposes of conversation control.

Of particular interest is the balancing of attention to events within the conversation and to events external to the conversation, and the handling of endogenous (goal-directed, voluntary) and exogenous (outward, grabbed) attention. Exogenous attention, in particular, can be based on simple low-level image related events, such as a light flashing, or motion related, through the use of saliency volumes.

5.1.1.2 Details

The consideration and integration of attention to the environment and attention to a task is highlighted as a significant feature of this work. Here, we define the task as a conversation and are therefore interested in the guidance of attention during conversation scenarios. We emphasise the notion of an agent which is not only involved in conversation, but is also capable of paying attention to other potentially important events, possibly outside of the conversation, rather than acquiring a ‘tunnel vision’ for the conversation. Emotion is viewed as being indicative of a more concrete basis for defining notions such as importance in the context of attention. Therefore, the balancing and resolution of internal goals with unexpected external stimuli is of great consequence, implying some capacity for interruption handling.

For clarity, this part of the capability can be split into two conceptual components, both of which must interact:

Intra-conversational attention providing a description of, for example:

- Who in the conversation the listener(s) attends to and how much
- Who in the conversation the speaker attends to and how much

Extra-conversational attention:

Where or to whom a conversant attends external to the conversation and how much e.g. social contacts or stimuli in the environment.
Emotion is regarded as being an important factor in the allocation and resolution of both intra- and extra-conversational attention guidance. In order to achieve this, we intend to take the approach of endowing agents with a form of synthetic social perception that allows particular attention to be paid to other agents in the environment, including their bodies, gaze, faces and gestures. Such a mechanism is important for allowing agents to be more socially focused and acts as a basis for allowing the perception of emotions from other agents as well as aiding conversation control. This social attention mechanism may also allow the consideration of additional behaviours, such as ‘attention paid to attention’; i.e. certain behaviours that direct attention are often of interest to other entities capable of perceiving them as indicators of goals and intention.

The extra-conversational attention component must also account for stimuli in the environment; for example, stimuli eliciting or indicating fear or threat may cause changes to the focus of attention away from the conversation. A key point is that extra-conversational attention should take place within the framework of the conversation; this includes notions of where speaker and listeners are looking during conversation when mutual gaze is not taking place as well as under what conditions attention may be ‘grabbed away’ when mutual gaze is taking place.

5.1.2 Adapt politeness behaviours to the user’s emotional state

5.1.2.1 Description

Brown and Levinson’s (1987) theory on politeness describes traffic rules of social interaction focusing on verbal communication. The role of emotion in choosing appropriate communication strategies and the importance of non-verbal behaviour in mitigating face threats has not been investigated. But undeniably, emotions have an important impact on the perceived face threat, e.g., a very angry addressee will alter the speaker's perception of a potential face threat. Consequently, adapting politeness behaviours to the user’s (or other agents’) emotional state, an ECA should (1) try to mitigate face threats by an appropriate combination of gestures, mimics and speech and should (2) take into account not just the strength of emotions but also their causes.

This capability requires information from the perception domain regarding the user

- what kind of emotion (and with what intensity) does the user display
- who is the addressee of this emotional display
- what is the relation between the user and her addressees

and regarding the environment and the situational context, e.g., the kind of the interaction and the current state of the interaction.

This capability will employ a theory of politeness to contribute to the decision making process by deciding which verbal and non-verbal acts the agent should perform and whether emotional states should be suppressed or not. To further this process, information from the perception and interaction domain will be taken into account:

- the user’s emotional state
- the agent's emotional state
- context dependent variables like
The capability will have to decide on appropriate politeness behaviours that mitigate face threats by means of gestures, mimics, and speech. This includes means to decide for “false” emotional displays in order to appear natural, e.g., telling a joke involves corresponding non-verbal behaviours and emotional displays like smile or laughs.

The verbal and non-verbal politeness strategies will be represented by sequences of dialogue acts annotated in an APML-like language to be passed on to the generation component.

5.1.2.2 Details

This capability will employ a theory of politeness to contribute to the decision making process by deciding which verbal and non-verbal acts the agent should perform and whether emotional states should be suppressed or not taking into account not only the strength of emotions but also their causes.

To achieve these goals, we will investigate:

- **The cause of the user’s emotion**: We will need a situational representation that allows us to derive the potential cause of a user’s emotion. This may include information on social variables like distance and power as well as the kind of the interaction and the current state of the interaction. The potential cause of an emotion will inform the process of choosing a specific communication tactics. If a student is frustrated by continuing system problems, it will not suffice for the agent to sympathize with the student by saying “Let’s try it again. Together we will manage it.” On the other hand, if the student is frustrated because she was not able to solve the problem by herself this might be just the right move of the agent.

- **The when and how of displaying emotions**: Being polite often means holding back one’s own emotions or faking them for the sake of the interaction. Thus, taking the emotional state of speaker and/or hearer into account while choosing appropriate politeness behaviour has to be treated separately from the question what kind of emotional display will accompany the realization of the chosen strategy. This decision will be constrained not only by the emotional states of speaker and hearer but also by the situational setting like competition or cooperation.

- **Corpus Collection**: To facilitate the aforementioned research, we will collect a corpus of interactions between users and an agent in the test-bed GAMBLE which allows us (i) to create highly emotional situations and (ii) a high degree of control over situational factors.
5.2 Capability 2: Creating Affective Awareness

Getting the users to be engaged and maintaining the user engaged in an interaction is a key capability an ECA should have. In particular, ECAs should be encompassed with the capability to perceive user’s emotion, to adapt to it, to react to it and to imitate it.

5.2.1 Creating Affective Bonds

5.2.1.1 Description

A necessary requirement to create affective bonds between user and agent is to make sure that there is a perceived connection between them. Sidner and colleagues (2004) use the term engagement to refer to the process of initiating, maintaining or finishing such a perceived connection. Consequently, the capability includes methods to (1) monitor the user’s level of engagement and (2) means to explicitly influence that process.

This capability requires information from the perception domain, namely information

- regarding the user:
  - a. what is the user looking at (the agent, other users, the environment)?
  - b. what kind of emotions does he or she display to whom (the agent, other users, the environment) and with what intensity?

- regarding the interaction environment (e.g. the state of the game)

Furthermore, the capability requires information on the discourse history from the interaction domain, such as the turn taking behaviours of the user.

To assess the user’s level of engagement, information from the perception and interaction domain will be taken into account including:

- the user’s attentive behaviours (distribution of attention between the agent, other users, the application and the environment)

- the user’s dialogue behaviours
  - How often does the user take the initiative?
  - What is the duration of his or her turns?
  - Whom is he or she addressing (other users, the agent, off talk)?
  - What kind of verbal and nonverbal acts does he or she employ (agreement, disagreement, …)?

- the user’s emotive state derived from
  - from his or her expressions
  - from the discourse history
  - from the situational context
To increase the user’s level of engagement, the capability will include verbal and nonverbal means to:

- increase the user’s awareness of the agent
- encourage the user’s interaction with the agent
- deliberately influence the user’s emotive state

In addition, the capability includes the determination of appropriate affective responses. In order to influence the user’s level of engagement, we need to decide on (1) the type and intensity of the emotions to be displayed and (2) the channels of expressions to be used. For instance, if the user pays hardly any attention to the agent’s face, acoustic means of expression might be given priority.

The module for enabling this capability will deliver annotated listener and speaker behaviours to the generation domain.

5.2.1.2 Details

This capability will include verbal and non-verbal means that may help to increase the level of emotional awareness between user and agent. On the one hand, the agent should be sensitive to the user’s emotional state and act on it. On the other hand, the agent’s emotional signals should stimulate (unsolicited) responses from the user. To achieve these goals, we will investigate:

- **Emotional displays to increase awareness**

  A basic requirement to create affective bonds between individuals is to make sure that they are aware of being looked by each other. Establishing eye contact is an important means to establish such a connection. But how can the agent signal the user that it has noticed that the user is paying attention to it as well? We will investigate in how far emotional displays can be exploited to signal awareness. For instance, human speakers usually smile to show the addressee that they have noticed that he or she is looking at them as well.

- **Emotional displays to improve user engagement**

  According to Sidner and colleagues (2004), engagement is “the process by which individuals in an interaction start, maintain and end their perceived connection to one another.” We will explore various kinds of emotional response an agent might provide and investigate how they support the engagement process. For instance, the agent might show anger if the user shows anger (mirroring), the agent might show empathy if the user is upset (sharing) or the agent might feel ashamed if the user blames him or her (complementary). In addition, we will investigate how the intensity of emotions will influence the engagement process.

- **Measurements for affective bonds**

  To measure the degree of user engagement, we will set up a list of criteria including the user’s gaze behaviours, the user’s emotional responses, and the duration of the interaction. In addition, we will compare the user’s behaviour towards the agent with
that towards other humans and try to derive guidelines for the improvement of the agent.

We will create a highly emotional situation that allows us to investigate an ECA’s capability of adapting its reactions to the user’s emotional state as well as converging or diverging with the user’s behaviour appropriately. As an evaluation platform, we intend to use Gamble (Greta’s Affective MoBiLe dicing intEraction, see also WP8). This is a small game of dice where one of the game partners is substituted by the Greta Agent (developed by Catherine Pelachaud and colleagues). This game can only be won by deceiving the other players and by detecting such attempts from the other players. Thus, a highly emotional situation is created when the agent blames the user for deceit or when the user detects such an attempt and the agent has to react to it. This showcase will help us to investigate how subtle emotional signals employed by a conversational embodied agent are perceived by the human user. Currently, there are no specific attempts to provide the user with the impression that the agent is aware of her. As a consequence, the users don't pay as much attention to the emotional response of the agent as we would have wished.

The new version of the agent will be enhanced by a module for supporting the engagement process. The module will include an empirically validated repertoire of strategies to initiate, maintain and end a perceived connection. For instance, the agent needs to be able to deal with situations in which engagement behaviours diverge. The strategies will be based on attentive behaviours as well as emotional displays.

5.2.2 Imitation and affective links

5.2.2.1 Description

We are investigating the role of affect in low-level imitation using robots (see Section 2.1.1.2.1). Based on this work done with robots, developed under the exemplar of WP7, for this WP we aim to provide cues and guidelines to help the designers of ECAs to endow these agents, which have a very different embodiment and architecture, with similar capabilities.

5.2.2.2 Details

By low-level imitation we understand rather automatic processes of synchronisation and coordination of behaviours that do not involve an explicit "intention" to imitate nor reasoning about the behaviour of the other agent, its mental, or affective state. An example of low-level imitation related to affect is the phenomenon of emotional contagion (Hatfield, 1994), in which an agent mimics automatically the expressive behaviour of another and, by doing so, enters in a similar (internal) affective state. Other roles that affect play in low-level imitation include influencing the choice of what, when, why and to whom to imitate or not to imitate based on affective bonds developed during the interaction. Based on our work done with robots, developed under the exemplar of WP7, for this WP we aim to provide cues and guidelines to help the designers of ECAs to endow these agents, which have a very different embodiment and architecture, with similar capabilities.

Examples of scenarios relevant to this end include tasks in which two agents need to coordinate their behaviour and “get along” or “share a positive affective state” in order to accomplish the task, e.g. dancing or following another robot to obtain a needed resource. Features such as speed, beat, or amplitude in the execution of motor behaviours can for example be used to:
1) Mutually influence their affective states by automatically “matching” or “mimicking” those external features

2) Modify their affective links over the course of prolonged interaction, depending in the average quality of the synchronisation achieved

3) Learn the correlation between the internal affective state (e.g. a “mood” or a stronger emotion) and its behavioural effect on itself or on the other agent

4) Try to influence the behaviour and affective state of the other: e.g. trying to transform a behaviour that is inefficient to achieve a task into an efficient behaviour (with the right amplitude, speed, etc) by getting the “inefficient” robot to imitate one’s own behaviour, i.e. by “teaching” the robot the appropriate realisation of the task.

5.2.3 Adaptation

5.2.3.1 Description

We wish to obtain information about the emotional state of the users by analysing their expressive gestures and to adapt the virtual subject’s reactions depending on the inferred emotional state of the users. This capability requires the ECA to be endowed with sensors (e.g., video cameras, microphones) in order to capture data from users. Information can then be extracted at several levels, e.g., from low-level motion cues (energy, contraction/expansion), to mid-level expressive features (motion fluency, impulsiveness), toward high-level expressive, emotional descriptors (e.g., in terms of the models worked out in WP3). Simple reactive strategies (e.g., condition-action rules) can be employed for quickly adapting the subject’s actions.

5.2.3.2 Details

This task aims at obtaining information about the emotional state of the users by analysing their expressive gestures. The goal is to adapt the ECA’s reactions and behaviour depending on the inferred emotional state of the users. Information can be extracted at several levels, e.g., from low-level motion cues (energy, contraction/expansion), to mid-level expressive features (motion fluency, impulsiveness), toward high-level expressive, emotional descriptors (e.g., in terms of the models worked out in WP3). Simple reactive strategies (e.g., condition-action rules) can be employed to quickly adapt the ECA’s behaviour.

Domain: The task mainly concerns the perception domain. Interaction is concerned as for the definition of simple reactive rules.

Input: data from sensor systems (e.g., video-cameras, microphone). The ECA should be endowed with sensors in order to observe the user’s behaviour.

Output: a collection of features describing the behaviour of the user and her possible inferred emotional state (e.g., either in terms of basic emotions, or as stated in the component process model by K. Scherer, or in terms of the models worked out in WP3).
5.3 Capability 3: Backchanneling

5.3.1 Description
The ECA should give appropriate feedback while the user is speaking. This “backchanneling” has a communicative spoken component, even if it is not a fully verbal one. Examples include: "mmmmh" with a rising pitch and a frown might express doubt or disagreement; "mmmmh" with a falling pitch and a nod might express agreement or understanding; "lalala" with a gaze moving away from the speaker might indicate a lack of interest, etc.

These paralinguistic spoken backchannels could be considered “auditive gestures”. Interesting compensation phenomena are to be expected, in the sense that auditive or bodily gestures could be used with similar expressive meanings, either in combination or individually.

5.3.2 Details
In the Perception domain, this capability requires some incremental processing of the user input, so that appropriate moments in time for realising backchannel utterances can be identified. Ideally, it should be combined with some understanding of the intended user message, to give meaningful backchanneling information related to the user utterance.

In the interaction domain, the incremental processing result from the perception domain must be analysed in terms of the appropriateness to react using a backchannel utterance. This reaction can be conceptualised as belonging to two distinct levels:

i. a merely reactive kind of backchanneling feedback indicating that one is still listening;

ii. and a more cognitive, evaluative backchanneling communicating an evaluation of the user utterance.

In the generation domain, backchannels can be treated similarly to short, multimodal expressive acts, in which visual and/or auditive modalities may play a role, and that may or may not have a verbal component.

5.4 Capability 4: Coordination of signs in multiple modalities
This capability aims at enabling the ECA to show multimodal behaviour integrating text, tone of voice, facial expression and gesture enhanced with signs of emotions in these multiple modalities. This capability is uniquely relevant for the Generation domain.

The foremost task with respect to Capability 4 is to define levels of description and to design representations to describe the relationships between the signs of emotion in different modalities.

Therefore one of the main endeavours related to Capability 4 will be defining a rich representation language that allows us to describe affective multi-modal behaviour at various levels, such as which information is relevant at a level of utterance planning and from which (hypothetical) component does it come from, which information needs to be contributed by the speech synthesis component, how and at which levels of detail do information from speech synthesis and facial expression modelling interact, how comes gesture into play? Accordingly we will need to define, e.g. a “speech synthesis output language” specifying those aspects of emotion related to the tone of voice, a “facial expression related output
language” specifying the emotion related information in the face, a “gesture related output language” specifying the emotion related information conveyed by gestures, etc., and even more importantly we need to identify those aspects of the language where information conveyed through the different modalities can be integrated. Moreover, given an emotion (non basic or not) we will need to specify how to coordinate the signals across modalities.

We propose to tackle this overall goal in a number of different ways:

In one strand of our research we will focus on facial animation in dialogue context. In order to make believable facial expressions, we need to study precisely how to mix the emotions coming from all the expression channels. They are many expression channels (phonemic movements, conversational markers such as emphasis, punctuators, emotions at different levels...) that have different time scales and need to be synchronized and blended together.

In another strand of our research we will concentrate on the display of non-basic emotions by investigating the following capabilities:

- to represent coordination of signs of non basic emotions in multiple modalities
- to specify the multimodal behaviour to be displayed by the ECA
- to manage audio-visual speech

Again in another area we will work on the development of high level representations for gestures that allow for player independent descriptions for discourse accompanying gesture generation where also the fact is accounted for that in multimodal (affective) human expression we rather find a constrained mix of the modalities than an additive display of expressions conveyed via individual modalities.

5.4.1 From multimodal emotional corpora to models of coordination between modalities

5.4.1.1 Description

Here, we aim at building a model of the coordination of multimodal behaviour during non basic multimodal behaviour observed in non acted data. It will use multimodal corpora built in WP5; these corpora will be annotated, where possible, using the tools provided by WP4. This annotation is essential in providing insight on the synchronization of the different means by which expressions are conveyed (e.g. MPEG-4 FAPs for facial animation and BAPs for body gestures), as well as on the actual evolution of these features over time. Particular care will be given by working with other WP6 partners to ensure that the coding scheme used for annotation is adequate for the specification of ECA.

5.4.1.2 Details

The following questions (which address different levels) of specification (for ECA use) will be tackled:

- How to model / specify non basic emotional behaviour and the emotional context (e.g. the input expected from the interaction domain)?
How to model / specify the coordination between the modalities in emotional behaviour regarding temporal behaviour, semantic relations (redundancy, complementarity…), constraints and conflicts?

Proposed solution will include the definition of one or several representation language(s): either a single rich representation language that allows to describe affective multimodal behaviour at various levels, or several languages, one for each level. We will start by a selection and definition of representation languages for individual modalities.

Illustrative examples of models/rules will be described on how to map or integrate representations between different levels:

- emotions and multi-modal coordination
- multi-modal coordination and mono-modal specification of expressive behaviour

From multi-modal emotional corpora to model of coordination between modalities: a corpus of non acted behaviours involving non basic emotional patterns will be used as it will illustrate how to go beyond the specification of basic emotions; representation languages will be defined at different levels and with the dual goals of both enabling annotation of the emotional behaviour observed in the corpus and specifying emotional multimodal behaviour of the ECA.

5.4.2 Multimodal behaviour

5.4.2.1 Description

Humans are endowed with complex communication means. Messages can be transmitted through the voice, an arm gesture, a facial expression. Verbal and nonverbal behaviours may be considered as different modalities working toward a same goal. They may not provide the same exact information but the transmitted information contribute to the overall communicative process. They can not be interpreted separately. Their significations add to each other to produce the final communication: gaze, body posture, hand gesture, facial expression, intonation, all these signs are interleaved and participate to the communicative process. The creation of an ECA should include these different communicative modalities.

Thus, an ECA has to decide which modalities to use to express an emotion and a communicative act. This includes the specification of the speech-related interrelationships, including:

- lip-synch (the synthesis component needs to inform the visual component of the phoneme durations);
- audio-visual time alignment on higher-level information such as accented words, phrase breaks;
- audio-visual expression aligned on the semantics of the text (e.g., "I will NOT do it!" with an emphatic accent on NOT accompanied with a fist-in-the-air gesture)

5.4.2.2 Details

We need to represent the relationships between the signs of emotion and of communicative acts in different modalities. This capability is uniquely relevant for the Generation domain.
The work considered here embeds several steps:

- **Selection:** Given an information to transmit, select which modalities should display it; the information may come from the interaction component, a specification of the emotional state to express, and the communicative intention (if it is only to convey information, the text / propositional content to express would suffice).

- **Scheduler:** when assigning a modality to display an information, this modality might be already busy (when you are carrying an heavy object you can not use your hands to indicate a point in space). The scheduler will plan the attribution of modalities to express an emotion and a communicative act over time.

- **Synchronize the modalities:** each modality has its own synchronisation scheme (a blink differs from an arm gesture). One needs to model the temporal course for each of them.

In particular, we will focus on facial animation in dialogue context. In order to make believable facial expressions, we need to study precisely how to mix the emotions coming from all the expression channels. They are many expression channels (phonemic movements, conversational markers such as emphasis, punctuators, emotions at different levels...) that have different time scales and need to be synchronized and blended together.

### 5.4.3 Gesture repositories

#### 5.4.3.1 Description

An important precondition for generating appropriate multimodal behaviour in ECAs is the availability of gesture repositories, i.e. archives of gesture representations/descriptions that are used to guide the selection or creation of gestural movements as part of verbal and nonverbal communicative behaviours.

These representations need to be realized at a high level of abstraction in order to be reusable with different types of player technologies. The representations also need to be of such a kind that it is possible to account for the fact that facial expression, posture and gesture may constrain each other, and to allow for fine-tuning of the temporal alignment of verbal expression with the duration and dynamics of the accompanying gesture(s).

#### 5.4.3.2 Details

We propose the investigation of the construction of a gesture repository (gesticon) for behaviour generation of ECAs. This involves answering several challenging questions, including:

1. What should a Gesticon contain?
2. What should be the structure of a Gesticon?
3. Do we need a Gesture Grammar in addition to a Gesticon?
4. How should the Gesticon cooperate with other modalities in multi-modal generation?

For example, if gestures are to be based on some kind of repository, this repository must contain reference to form as well as semantics of meaning of gestures. It may also contain reference to the communicative functions of the gestures. Open issues include the question of...
whether a set of base gestures or gesture primitives can be used to construct all needed gesture types and variants. For individualised gestures, what are the determining factors and how should they be mapped onto form parameters. For gesture selection, should the gesture specifications in a gesticon include weights for some parameters so that prioritisation between multiple gestures can be made. In addition, what characteristics depend on context and which are independent. How should iconic gestures be dealt with? These are all open questions that must be tackled in order to proceed towards the construction of expressive agents with capabilities beyond those of contemporary attempts.

5.5 Capability 5: Expressivity

Human individuals differ not only in their reasoning, their set of beliefs, goals, and their emotive states, but also in their way of expressing such information through the execution of specific behaviours. Moreover, emotion is not simply expressed through a static facial expression or setting of vocal cords. Acoustic and visual expressions of emotions are dynamic; they evolve through time.

5.5.1 Behaviour expressivity

5.5.1.1 Description

The manner a movement is done provides relevant information on the affective state of the emitter. In this capability we consider the quality aspect of behaviour. We aim to define the expressivity attributes for generating the behaviours, and the mapping from emotion models onto suitable values for these parameters. To achieve this, current emotional model implementations are helpful, especially when they have to means to provide expressivity features for realistic emotions and not necessarily acted or extreme ones. To this end, we seek to move away from a generic action model and instead simulate individualized agents that portray idiosyncratic behaviours. Achieving expressivity in the agents’ facial animation and gesturing will be accomplished by modifying behaviour generation to reflect internal and external influences and a personal communicative style.

A key issue will be the definition of suitable strategies for mapping the information about users’ behaviour obtained from the adaptation modules onto real-time generation by the ECA of expressive behaviour. This objective can be achieved by developing mapping strategies integrating both fast adaptive and reactive behaviour (like in the adaptation module) and more high-level decision-making processes.

5.5.1.2 Details

We aim to investigate the qualitative aspects of co-verbal behaviour executed by Embodied Conversational Agents. Our attention is restricted to generating phenomenologically accurate behaviours as we focus on developing an intermediate level of behaviour parameterisation, a set of dimensions of expressivity. We regard expressivity parameters as useful enabling tools to mediate between high-level, qualitative goals such as the communication of emotion and personality, and low-level animation parameters. Our search for suitable techniques is guided by the simultaneous goals of allowing for the realization of a substantial subset of high-level functions, while at the same time providing an efficient conversion into animation parameters in a near real-time environment.

It is our goal to preserve the semantic value of each gesture during the computation of expressivity. Effective strategies have to adjust behaviour on multiple levels - from abstract
planning (whether to search for a gesture for a given text at all), to gesture phase-level modifications (whether or not to repeat a stroke), down to adjusting velocity profiles of key pose transitions.

Expressive motions for an ECA is generated through the use of expressivity attributes (partly detected via the analysis phase) acting as an intermediate reparameterisation that links high level concepts, such as emotion, to low-level animation parameters. We base our model from the analysis of the expressive gestures of the user e.g., the way he/she moves (with or without energy, in a fluent or bounded way etc.).

A key issue is therefore to investigate the strategies for mapping the high-level information on the user’s behaviour onto real-time generation of a suitable expressive behaviour for the ECA. Such strategies may be situated at different level of complexities, from simple static rules and functions to high-level decision-making processes.

5.5.2 Speech Expressivity

5.5.2.1 Description

For language and speech, the capability to show expressive behaviour consists of two main aspects: (1) elaborating our knowledge about emotional expression in language and speech, and (2) improving the methods for expression emotions in synthetic speech.

5.5.2.2 Details

5.5.2.2.1 Knowledge about emotional expression in language and speech

We will concentrate on the interaction of voice source parameters (including f0 and intensity) and temporal features in affect signalling. This task will encompass the following subtasks:

- establishing “baseline” behaviour of the prosodic and segmental voice variation in affectively neutral utterances,
- looking at how the neutral prosody (for these dimensions of the speech signal) is transformed in affectively coloured repetitions of the same utterances,
- examining through synthesis the role of voice source parameters (including f0 and intensity) and temporal features in affect signalling,
- cross-cultural differences in affective attribution of voice quality difference.

These analyses will involve close inspection of relatively limited numbers of utterances and affective samples. The goal is to try and establish some of the kinds of variation that large corpus based approaches will eventually need to focus on, and which technology providers will need, to incorporate affective “tone of voice” into man-machine interaction systems. This task will require very specific recording types (provided within WP 5).

5.5.2.2.2 Methods for expressing emotions in synthetic speech

The methods to express emotions naturally depend heavily on the synthesis technology used. We will experiment with several techniques including HMN, PSOLA, unit-selection and diphone-synthesis.
We will develop strategies for different kinds of modifications taken separately or combined (F0, duration, energy and timbre) and do proof-of-concept experiments utilizing copy synthesis.

Further we will investigate how possibilities to simulate a mixture and graded emotions with a rule-based synthesizer based on a discrete emotion concept as well as dimensional models and compare different approaches.

5.5.2.2.3 Evaluation of the existing speech synthesis capabilities

Subjective tests will be led in order to evaluate the different speech synthesis capabilities available (eg but not restricted to TD-PSOLA, HNM and formant synthesisers). The speech sequences will be evaluated by a panel of listeners in laboratory according to four criteria:

- the similarity between each synthetic (neutral modified) and target sequences;
- the acceptability of the synthesised speech;
- the naturalness of each synthetic speech sequence;
- the preference between the speech synthesis sequences.

5.5.3 Context dependent emotional body gesture

5.5.3.1 Description

People display different gestures with their arms and hands depending on their emotional state and the context: nervous tics, hands in pockets, body balancing, avoiding gaze, trembling … Moreover, under certain circumstances (in presence of a particular person, under stress) such gestures get accentuated and/or some others may appear.

We aim to provide ECA’s with the capability of modulating its body gestures (in particular upper-body: head, shoulders, arms, hands) according to the emotional state induced by the context –situation- in a virtual environment.

5.5.3.2 Details

We will consider an emotional model in which high emotional states are displayed for a few seconds while emotional gesturing remains at a less intense level for longer periods. Different temperaments will be defined which will shape the role of an ECA and its “personal view of life”. Some agents will be naturally cheerful or optimistic, while others can be negative or pessimistic.

We will define sets of body emotional gestures for each of the 6 basic emotions identified by Ekman: anger, disgust, sadness, enjoyment, and fear/surprise. Emotional gestures will be modulated to display emotions corresponding to different contexts. Gesture modulation will consist on modifying amplitude of movements and selecting the kind of gesture to display (animation sequences involving one or more body regions: arms, hands, trunk, etc.). Agent temperaments will be defined as a set of parameters and rules describing the “tolerance” and “sensitivity” of the ECA when facing different situations.

A library of temperaments (parameters and rules for modulating emotional gestures) will allow us to generate expressive animation which can be interpreted as the manifestation of an
emotional state in a given context. The temperament of each ECA will help it decide how to behave according to a given situation. Advanced animation techniques (real-time body animation, coding and rendering) will be used to modulate context-dependent emotional gestures in interactive applications. This research will extend our work on providing virtual characters with the capability of reacting to their environment (Conde & Thalmann, 2004; Gutierrez et al., 2004) and a Virtual Human Animation Engine (Gutierrez et al., 2002) designed to synthesize body emotional gestures as a function of the cultural profile of a Virtual Human.
6 Proposed allocation of tasks

6.1 Task 1: Representation Languages and Theoretical Framework

Deliverable: report describing the representation languages and the theoretical Framework of the overall architecture; due Month 23

6.2 Task 2: Mock-up of each capability

Deliverable: report describing the mock-up of each capability; Month 35

6.3 Task 3: Proof-of-concept of each capability

Deliverable: report describing the proof of concept of each capability; Month 46
7 Provisional timeline

7.1 Task 1: Representation Languages and Theoretical Framework

<table>
<thead>
<tr>
<th>Key actions</th>
<th>Estimated completion</th>
<th>Type of associated deliverable</th>
<th>Estimated date of delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representation languages; specification of the main architecture</td>
<td>M 12-23</td>
<td>report</td>
<td>Month 23</td>
</tr>
</tbody>
</table>

7.2 Task 2: Mock-up of each capability

<table>
<thead>
<tr>
<th>Key actions</th>
<th>Estimated completion</th>
<th>Type of associated deliverable</th>
<th>Estimated date of delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mock-up of each capability</td>
<td>M 18-35</td>
<td>report</td>
<td>Month 35</td>
</tr>
</tbody>
</table>

7.3 Task 3: Proof-of-concept of each capability

<table>
<thead>
<tr>
<th>Key actions</th>
<th>Estimated completion</th>
<th>Type of associated deliverable</th>
<th>Estimated date of delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proof of concept</td>
<td>M 20-46</td>
<td>report</td>
<td>Month 46</td>
</tr>
</tbody>
</table>
8 References


