

humaine

D6g

Building an Affective Interactive ECA

Workpackage 6 Deliverable



Date: 31st August 2007

IST project contract no.	507422
Project title	HUMAINE Human-Machine Interaction Network on Emotions
Contractual date of delivery	<i>August 31, 2007</i>
Actual date of delivery	<i>August 31, 2007</i>
Deliverable number	D6g
Deliverable title	Building an Affective Interactive ECA
Type	Report
Number of pages	23
WP contributing to the deliverable	WP 6
Task leader	University of Paris 8
Author(s)	Catherine Pelachaud and WP6 Partners
EC Project Officer	Philippe Gelin

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.....	4
2	BRIEF OVERVIEW OF WORKPACKAGE 6 AND ITS EXEMPLAR	5
2.1	The field covered by Workpackage 6.....	5
2.2	The research objectives.....	6
2.3	Overview of main elements of the exemplar.....	6
3	ACHIEVEMENTS TOWARD MAIN ELEMENTS OF EXEMPLAR.....	8
3.1	Element 1: Cognitive influences on action.....	8
3.2	Element 2: Creating Affective Awareness	9
3.3	Element 3: Backchanneling.....	11
3.4	Element 4: Coordination of signs in multiple modalities	11
3.5	Element 5: Expressivity	13
4	PLANNED PROGRAM OF RESEARCH	16
4.1	Element 1: Cognitive influences on action.....	16
4.2	Element 2: Creating Affective Awareness	16
4.3	Element 3: Backchanneling.....	17
4.4	Element 4: Coordination of signs in multiple modalities	18
4.5	Element 5: Expressivity	19
5	REFERENCES.....	21

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 pre-specified: 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. (p 4)

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 (p.16)

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. (p. 21)

The consultation phase has now ended. Near-final plans were presented to the whole network at the Plenary in May 2005, and adjustments have been made accordingly.

This deliverable reports the plans that have now been set out for the remaining months of the project. They are necessarily provisional, because they are subject to two reviews (in 2006 and 2007) before they are completed.

Work has begun on several aspects of the planned programme. It is also reported. Ethical issues affect the whole of HUMAINE, but rather than repeating essentially similar points in multiple deliverables, they will be handled coherently in a single document, D0o (Science and Society).

2 Brief overview of Workpackage 6 and its exemplar

Within WP6, we are particularly interested in the role of emotion in interaction. This encompasses abilities in three domains, perception, interaction and generation, 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. Thus, 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. We refer the reader to deliverables D6b, D6c, D6d and D6e for the evolution of the exemplar and in-depth descriptions of the aspects highlighted here.

2.1 The field covered by Workpackage 6

An 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... WP6 tackles these issues over three main domains:

- 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.
- 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.
- 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 and/or other ECAs.

2.2 The research objectives

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:
 - Cognitive influences on action
- Interaction Domain:
 - Create affective awareness
 - Backchanneling
- Generation Domain:
 - Coordination of signs in multiple modalities
 - Produce dynamic expressive visual and auditory behaviours

2.3 Overview of main elements of the exemplar

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, gestures and linguistic expression. It consists of two key sub-topics: attention shifting related to emotion and the adaptation of politeness behaviours to a user's emotional state.

2. *Creating Affective Awareness*

In this capability, we seek to create connections between a user and an ECA with the aim to maintain user's engagement during an interaction. Sub-topics in this element are creating affective bonds between the user and ECA, imitation and adaptation to generate or alter the behaviour of the ECA.

3. *Backchanneling*

Backchannel utterances and signals can convey simple information of critical importance to the success of communication. Here we seek to investigate backchanneling undertaken by the listener in order to provide the speaker with immediate feedback about the state of the communication, in order to improve the interaction capabilities of an ECA.

4. *Coordination of signs in multiple modalities*

ECAs must show expressivity in a consistent and natural looking way across modalities in order to be believable. Sub-topics in this element consider coordination of multimodal behaviour from corpora, relationships between signs of emotion in

different modalities and gesture repositories for generating appropriate multimodal ECA behaviour.

5. Expressivity

In this capability we are considering expressive aspects of speech and behaviour by looking at sub-topics such as an intermediate level of behaviour parameterisation utilising dimensions of expressivity, expression through acoustic parameters of the voice based on speech synthesis techniques and emotional body gesture accounting for the situation context.

3 Achievements toward main elements of exemplar

3.1 Element 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, gestures and linguistic expression. 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 collaboration with WP7 has taken place. Here, we emphasise 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 interaction with the environment or social entities within that environment. These shifts may result in the production of visible phenomenon, such as gaze shifts, that are of significance to others in the environment.

Our work here has been developing as follows: We have enhanced the performance and functionality of our core attention model by adapting it to utilise graphics processing hardware, a necessary step in meeting the constraints of real-time interaction.

With WP7 partners we have defined a design for computational novelty relating to the early stages of appraisal theory. We have also designed and are starting to prototype elements of a general framework and scenario in which the novelty system can be embedded and evaluated. In addition, the framework has been designed so as to amalgamate aspects of emotion and attention into a single model, in order to perceive visually emotional and attentive stimuli from a virtual environment and to general emotional and attentive behaviours in a virtual agent towards such stimuli.

In collaboration with WP4 partners, we have started to introduce a number of image processing algorithms into our virtual environment for testing as part of our study into the notion of 'switchable sensing' – the idea of making agents capable of processing input from the real world as well as the virtual environment. We are investigating how these algorithms can fit into the design of our agent system.

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.

Our work here has been developing as follows:

The corpus of the Gamble study has to be analysed further regarding the reactions of the user towards emotional displays of the agent. This will give us further insights into which emotional displays of the agent were appropriate. At the moment, emotional reactions are triggered by cues derived from the state of the game at a given moment. The integration of a recognition system for emotional speech will allow for a more specific reaction of the agent. This system will first be tested in a single user scenario (see Section 4.2.d).

3.2 Element 2: Creating Affective Awareness

In this capability, we seek to create connections between a user and an ECA with the aim to maintain user's engagement during an interaction. Such ECAs should not only perceive the user's 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.

Our work here has been developing as follows:

During the Gamble study it was observed that, at least in this scenario, the agent became more believable when it started misbehaving, e.g. by insulting the user or by overtly expressing the negative emotion of rage about the current state of the game. An analysis of the user's reactions to the agent's emotional display will reveal if this is always the case in this scenario. Moreover it was observed that users started talking about the agent instead of talking with the agent, trying to explain the agent's behaviour and reactions to one another. Although this can be interpreted as very impolite behaviour towards the agent, it is a rich source of information to gain insights in the user's ideas about the agent and its performance and has to be analysed in detail.

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.

Our work here has been developing as follows:

The animation of the agent is done with Animation Score, a tool we have developed this past year. This tool allows for the control of the agent no more from high level specification (e.g. the communicative functions of the APML tags) but at the level of behaviours. One can specify the behaviours the agent should display through time. Behaviours for different modalities (facial expression, gaze, head movement, gesture) can be specified in Animation Score. The tool is interactive and allows one to visualise in real-time the

animation. Moreover one can specify values for all the expressivity parameters for each behaviour.

c. *Adaptation*

In order for an agent to be adaptive, information about the emotional state of the user(s) must be obtained. Here, the user's expressive gestures and language usage are analysed to infer his/her emotional state – the inferred state is then used to alter the ECA's reactions.

Our work here has been developing as follows:

We conducted perceptual tests towards the development of a computational model for emotion recognition by means of the analysis of the dynamics of movement and gestures. These tests have also been conducted on new sets of affective videos. Part of the work has been done in collaboration in collaboration with WP4.

We have developed a system to allow easier implementation of new features for our ECA Greta. The system facilitates the integration with other systems and will work in real-time. A first application of this architecture is an interface between the Greta and the EyesWeb systems. EyesWeb is used to continuously analyse video data coming from a video source, like a camera or AVI files and extract expressivity features (only the parameters of amplitude, speed and fluidity of movements are considered for the moment). Data are received by Greta which does not perform the same gesture but which replicates the expressivity (amplitude, speed, fluidity) of the original movement in real-time. The modularity of the implemented system allows us to switch between different modules with the same functions; that is, for example one could use expressivity data coming from pre-defined data files instead of EyesWeb for testing purposes.

d. *User's engagement*

Our work here has been developing as follows:

We have continued investigating how to act on user's engagement. We conducted experiments with subjects to test our audio-visual feedback system [Castellano et al. 06b]. We aimed to evaluate the effectiveness of the audio-visual feedback in reproducing movement expressivity of subjects and the impact on user's engagement. A triangulation between methods was used to evaluate the system: notes of the experimenter, video recordings and interviews of the subjects were compared. Results show that most part of the users perceived the changes in the audio-visual feedback generated by their movement and many of them understood the mapping we designed. Users' comments confirmed that our system provided strong sensations of participation, interaction and immersion.

3.3 Element 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 and signals 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.

Our work here has been developing as follows:

- **Generation of rules for backchannel behaviours:** We have worked on testing, evaluation and updating the back-channel framework. Work on the detection and interpretation of back-channel elicitation cues (head movements and prosodic features such as pitch detection) has been first tested on recorded data. Different versions of the rules generating the back-channel behaviours are being defined and tested.
- **Reactive Backchannel:** We have conducted studies on meaning of visual behaviours signals for backchannel and analysed the data. This study has been conducted across several cultures, namely France, Italy and the Netherlands. Later on we aim to pursue another evaluation study related to the level of mimicry and temporal synchrony to get for reactive behaviours.

3.4 Element 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 one 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 behaviours during non basic multimodal behaviours observed in non acted data, using multimodal corpora. Several studies about emotion in the social sciences are based on behaviours performed by actors during in-lab recording when being instructed to act a single emotion. Real life emotions are often complex and involve blends of emotions [Devillers et al. 06, Martin et al. 05].

Our work here has been developing as follows:

As the exploratory copy-synthesis approach that we have defined revealed to be successful for the study of levels of representation for blended emotions and expressive behaviours [Martin et al. 06b, Buisine et al 06], we aim at going further in this direction. We are working on making the copy-synthesis approach more automatic in order to ease the process of going from the

annotations to the replay by the Greta expressive agent. We are currently considering the representation languages under definition within HUMAINE for emotions (EARL) and behaviours (Gesticon).

Another strand of research deals with the automatic analysis of acted multimodal data from Geneva's GEMEP corpus [Baenziger et al 06]. This research aims for a better understanding and finally a model of the finer grained temporal synchronisation between expressive gestures and speech. For this purpose models for the automatic segmentation of speech at the phoneme-level have been developed, in order to provide exact information on the location of possible synchronisation points. These automatically derived phonetic segmentations provide a reasonable basis but still need to be manually validated and corrected if necessary. This task is currently performed for a significant subset of the GEMEP-corpus. On the other hand, computer vision algorithms are adapted for performing the automatic tracking of hands and faces on a per-frame level, and thus providing the base-level information necessary for all subsequent quantitative analysis of speech-gesture synchronisation.

b. *Gesticon*

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

Our work here has been developing as follows:

Currently Animation Score, our tool to control an agent from the specification of its behaviours through time (and not from communicative functions), is working with behaviours described within Gesticon (newly called Gestuary)/BML.

Links have been created between work conducting within Gesticon/Gestuary and SAIBA. SAIBA is an international efforts aiming at defining a unified representation languages for specifying and controlling emotions, communicative functions and behaviours.

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.

An example of a body gesture that depends on the context is the reaction movement. Reactions are unconscious behaviours that are not often implemented in virtual humans. The ability of reacting in virtual characters improves their realism. Based on an observation experiment of real people reacting we have identified different types of reactions: avoid, face and protect. These types of reactions are associated with personality traits [Garcia-Rojas et

al. 06]. To synthesise these reactions into virtual characters, we have created a semantic model that represents the animation synthesis, character's geometry, and a description of individual virtual humans through personality, emotion, age, gender, etc [Moccozet et al. 06]. Inside this model we have presented a reactive behaviour controller that describes the synthesis of a type of reaction according to individual parameters and properties of stimulus. This reaction has been implemented to the upper body and lower body of a 3D character using Inverse Kinematics [Garcia-Rojas et al. 07]. In this research, we have proved that using Inverse kinematics and well defined parameters we could create different reactive movements. The advantage of the semantic model is that allows scalability for implementation of concepts.

3.5 Element 5: Expressivity

Emotion is not simply expressed through a static facial expression or setting of vocal cords. Acoustic and visual expressions of emotion are 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.

Our work here has been developing as follows:

- **Analysis of emotional gestures:** We are currently working on computer-vision algorithms for extracting features about the spatial and temporal properties of gestures and postures displayed in a video corpus of acted emotions (the GEMEP corpus provided by the Geneva Emotion Research Group, [Baenziger et al 06]). A first version of a tool that performs frame-wise tracking of hand-positions has been developed. This tool supports an interactive tracking procedure, i.e. the user is asked for feedback whenever the tracking algorithm runs into decision problems and the user is asked for feedback on the obtained result. Thus manual validation of the tracking-results is an integral part of the tracking procedure, guaranteeing the soundness of the obtained results. For the tracking of face-positions a fully automatic tool has been developed. Currently the tracking tools are undergoing final refinements.
- **Synthesis of analysed emotional gestures:** We are working on developing a system that takes in input videos from the GEMEP corpus with real actors performing gestures in different emotional conditions and controls the Greta animation. The system analyses the actors' motor behaviour, extracts expressive indicators and controls the Greta animation so that the gesture performed by the agent communicates to users the same emotional state expressed by the actors. Perceptive tests with users are being conducted to allow us to evaluate both the analysis algorithms providing expressive cues and the synthesis approach for the generation of gesture expressivity in ECA.

- **Extension of behaviour expressivity to other domains:** We have conducted an analysis of 2D cartoons where we annotated gesture expressivity of the characters and we have investigated the modulations of this expressivity over time, named breaks. We have observed that these breaks can act as a rhetorical relation of similarity on the one hand, and as a rhetorical relation of contrast on the other hand [Ech Chafai et al., 2006].

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.

Our work here has been developing as follows:

The fruitful cooperation with the PAVOQUE project is ongoing. An **expressive database** for a limited domain unit selection voice has been designed and will be recorded. Signal analysis and modification algorithms are being pre-selected, evaluated and applied in database analysis and subsequent synthesis. One aspect of using an expressive target for selecting units is the use of copy synthesis: an expressive utterance, recorded with the synthesis voice but not included into the runtime data set, is analysed, yielding an **acoustic parametrisation** at a suitable degree of abstraction.

We have carried out first experiments with HMM-based speech synthesis (Krstulovic, Hunecke and Schröder, 2007), adapting a neutral German HMM-based synthesis voice to a small expressive data set of excited soccer comments. First results indicate that the out-of-the-box voice adaptation tool is not yet sufficient for adaptation of the relevant parameters: only mean pitch was adapted acceptably. A copy-synthesis test with the vocoding technique used (Mel-Log Spectral Approximation, MLSA) showed that the vocoder preserved the relevant features correctly; improvements should thus concentrate on the statistical models and the parameter generation from these models.

A **workshop on “Paralinguistic speech: between models and data”** has been organised at DFKI on 3rd August 2007, addressing the issue to build bridges between the two separate camps of “modellers” and “recognisers”, which traditionally have remained quite separate. Encouraging these camps to talk to each other is a first step towards a better mutual understanding, which may ultimately enable the two sub-communities to benefit from each others work.

For the next year the implementation of the issues mentioned in Section 3.5.b is planned. Furthermore we started work with the W3C incubator group on an **emotional mark-up language**. The support for this mark-up language with our tools will also be an issue of our work.

Forthcoming works will be directed towards the **refinement of the focus generation model**, according to the Valence-Intensity-Domain model [Aubergé and Rilliard 06]. We will focus on the two main points: accurate copy of prosodic cues on the focused part and neutralisation in the vicinity of the focused part, in order to avoid conflicting accents, one of the main

drawbacks inherent to the use of corpus-based synthesis. Signal processing techniques or/and unit selection will be used to reach the prosodic targets given by the model.

Listening tests will be carried out in order to evaluate the model on the three axes: valence (is the focus perceived or not?), intensity (if the focus is perceived, how strongly is it perceived?), and domain (is the focus perceived on the good linguistic domain?), as well as in terms of naturalness and acceptability. The perceptual exploration of the mapping of voice parameters to affect will continue and be further consolidated. To supplement and evaluate the output of perception experiments [e.g., Yanysheskaya et al 2006], production data will be analysed for a subset of the same affects.

A fundamental goal is to understand the **prosody** of the voice in expressive speech, i.e. not just pitch variation, but also the correlated source parameters. Towards this goal it will be important to contextualise the expressive dimension within what is currently understood about the dynamics of prosody. Analytic work is also planned here and there will be interaction with a TCD project on Irish Prosody.

The **cross language/cross cultural** investigations of the past three years will also be consolidated, and developed to yield initial principles for differentiating the universal from the culture-specific.

In order to serve the research community with an easy to use tool for emotional speech synthesis we developed Emofilt, the Mbrola-based multi-lingual emotional speech synthesiser. Interfaces to the widely used MARY speech synthesiser and the Festival framework have been implemented that enable researchers to integrate emotional expression with the respective synthesisers. The extensions with respect to graded and time-variant emotions were solved by parameter interpolation. The blending of emotions will be investigated by combinations of different parameter-categories denoting different emotion-related states, e.g. combining a sad intonation contour with a voice quality appropriate for anger.

4 Planned Program of Research

4.1 Element 1: Cognitive influences on action

a. *Emotion related attention shifts*

We will be conducting a further user evaluation study investigating user and agent conversation initiation with agents, based not only on gaze and locomotion direction, but also on the facial expression of the agent. The differences between user reporting of agent attention, emotion and intention when facial expression is and is not present will be of significant interest to our endeavours in creating perception-based behaviour models.

With WP7 partners we will be implementing, prototyping and elaborating a subset of elements that have been specified in our design of a computational novelty system. Furthermore, we will be improving the links between this system and the more general agent framework in which it is to be imbedded, so that novelty detection amalgamates into the framework as a single module in a generic perceptual / appraisal pipeline. The general framework will thus allow for the novelty module to be used and evaluated in a practical scenario, taking input from a virtual environment in order to influence the behaviour of an embodied virtual agent, primarily through appropriate facial expressions and gaze.

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

In our previous work, we focused on how to augment verbal politeness tactics by gestures. We will extend this approach by additional modalities, such as facial displays. The extension will be based on a corpus analysis conducted by Paris 8 from which rules linking politeness strategies and facial display types (masking, inhibiting, fake, felt) were derived.

4.2 Element 2: Creating Affective Awareness

a. *Creating Affective Bonds*

We will investigate how to create affective bonds between a user and an ECA by making the ECA respond to the user's emotional voice. Research so far has mostly dealt with offline evaluation of vocal emotions, and online processing has hardly been addressed. Online processing is, however, a necessary prerequisite for the realisation of human-computer interfaces that analyse and respond to the user's emotions while he or she is interacting with an application. While we used the Berlin emotional speech database for training a first classifier for our online processing component, we will record additional speech data from various speakers to improve performance. In addition, we will conduct an empirical study to investigate how people respond to an ECA that recognises the user's vocal emotions and shows empathy for him or her.

b. *Imitation*

We will continue using the Animation Score tool when replaying behaviours on an agent from the analysis of behaviours of a user from a video corpus. This will enable us to further study behaviour expressivity and which types are perceptually relevant.

c. *Adaptation*

We will work next on how to map user's movement expressivity to user's emotional state. Once the emotion has been recognised, the agent, when interacting with a user, can plan affective response accordingly to that. In fact, if the agent can understand the user's emotional state analysing the quality of his gesture, it can decide which facial expression to show. In this model the agent could interpret the user's state to improve the efficiency of communication.

d. *User's engagement*

A single user test scenario will be developed which will actively react to user's gaze behaviour to increase his/her awareness of the agent. This test scenario will be used to test user's reactions to emotional displays of the agent, e.g. an agent might smile if looked at and frown if the user gazes away. We will also employ a system to estimate the emotional state of the user from its speech input. The agent might e.g., give feedback to the user by mirroring his emotional state.

4.3 Element 3: Backchanneling

We will continue working towards developing a proof-of-concept backchannel and feedback ECA system. Following the framework for backchannel model presented in Deliverable 6d, two models should be considered: reactive and cognitive models. The former corresponds to simple mimicry emitted very often spontaneously and sincerely; on the other hand the latter reflect a conscious decision to provide backchannel to provoke a particular effect on the speaker or to reach a specific goal. The following main activities will occur:

- Extensive user tests are needed to find out how different back-channel strategies can lead to different impressions of the agent; the way the agent is engaged in the conversation (distracted, engaged, empathic, and so on) and to match different personality types. Furthermore, integration of the various components into a real-time system is on the agenda for the next phase of the project as well. This work will lead to new, improved versions of the architecture, taking up different development cycles.

Reactive Backchannel: The next steps planned concern the use of the adapted CDE architecture to implement a proof-of-concept backchannel and feedback ECA system. A subset of the architecture proposed in D6d for enabling an ECA system to show backchannel and feedback behaviour will be implemented. This will primarily concern the implementation of reactive components which do not require cognitive processing, such as an "audio feature extractor", a "reactive backchannel trigger", a "backchannel selector",

and a "backchannel scheduler". To the extent that speech recognition can be used or simulated, we intend to demonstrate the intended function of the "meaning analyser", the "cognitive backchannel trigger", and the "expressed backchannel meaning planner" by the use of mock-up modules.

4.4 Element 4: Coordination of signs in multiple modalities

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

BML. We are working on an implementation of the BML basic level 'realizer' for the Greta agent. BML is a new standard for the low-level description of virtual agents behaviour. The language also provides synchronisation mechanisms between the agent's modalities. Our realizer will allow us to use BML as a common platform for the specification of the agent's behaviour, both starting from high-level communication description (like for example APML) and from low level behaviour annotation. For example, it will be possible, starting from an emotional video with multiple levels of annotation (i.e., EmoTV), to automatically generate an APML file (from high-level annotation) and a BML file (from low-level annotation) and compute the corresponding agent's animations. In the future we will conduct perception tests to check if the emotional meaning conveyed by the synthetic agent is similar to the emotional content of the original videos.

Multimodal engine. We are developing a multimodal AMPL engine for the Greta virtual agent. Starting from an high-level description of the emotional/communicative intention of the agent (APML file) this module will compute the signals needed to convey the intended meaning on the following modalities: speech, face/head, gesture, torso. Signals on different modalities associated to the same meaning will be considered as a *composed* multimodal signal. For example, for a sad emotional state the engine may decide to produce a sad facial expression, stretching the arms down along the body and assuming a closed torso stance. This composed signal will be scheduled on the different modalities in a synchronised way.

Analysis of GEMEP-corpus. As described in Section 3.4, tools for deriving fine-grained temporal information from both speech and body-movements are under development. The work on the phonetic segmentation of speech, including manual correction has been finished for a reasonable sub-set of the GEMEP-corpus. This work will be continued, i.e. the proportion of manually validated data will still increase. The tools for hand- and face-tracking still need final refinements. All the tools and representation schemes will be properly documented in order to make them usable for other interested parties as well. The next step will be to analyse the uni-modal data from both the speech and the video channel obtained so far in order to investigate their properties in respect to temporal synchronisation.

- b. *Gesticon*

We aim to refine the Gesticon/SAIBA specification. At previous HUMAINE meetings (WP10 and plenary) a number of changes and additions to the current representation format were agreed on, which now need to be fleshed out in a

more detailed way.

The most important revision is concerned with the introduction of the concept of different levels of specification (LoS) in the Gesture Repository. We will add the definition of different LoS in the gesture repository. Two types of LoS can be distinguished. On the one hand, LoS may concern the level of precision of the data – stretching from mere animation names for motion-captured data, to fine-grained symbolic descriptions of behaviours. For this purpose for each type of channel (e.g. head, gaze, gesture...) and each level of detail a representational ontology will be defined.

On the other hand, LoS reflects the pertinence of different components of a gesture. E.g. studies have shown that not all dimensions of a gesture are relevant to convey its function or meaning. Thus we are looking for ways to represent this type of information in the Gesture Repository. by employing a typology of primary, secondary, and not relevant dimensions. Such descriptions would allow systems to automatically refine and adjust animations while pertaining their core meaning.

We aim to transform the behaviour representation scheme we have been using for the Greta agent technology to the Gesticon formalism. To ensure the portability of the language, i.e. to make sure behaviour descriptions can be interpreted identically by different agent geometries and animation players, we are collaborating with OFAI (in particular with the RASCALLI project) where a different agent technology is used.

c. *Context dependent emotional body gesture*

Our use case of a body gesture in a context dependent is the reactive movement in synthetic autonomous characters. So far we have modelled a semantic representation for individual reaction movements. This model represents the elements involved in this simulation: personality and emotion descriptors, behaviour controllers, movement controllers and geometrical aspects of the character. The results achieved provide different reactions to the same or to different stimuli depending on the internal state of a character (emotion and personality) and in the stimulus characteristics. However, as animations are procedural, specifically Inverse Kinematics, the movements are very robotic. To improve animations we aim to integrate emotional gestures to character's body parts that are not affected by the Inverse Kinematics. We are planning to integrate key framed movements such as facial expressions or hand movements that corresponds to the emotional state of the character. As a consequence, the emotional state needs to be updated after the reaction. For this, we are exploring some models already created for emotion update based on previous emotional state, induced emotion, and dimensions of personality.

4.5 Element 5: Expressivity

a. *Behaviour expressivity*

Analysis of emotional gestures: The tools for frame-wise tracking of hands and faces still need final refinements. The goal is to make them easily applicable for

partners in WP3 for actually performing the tracking tasks on the entire GEMEP corpus, in order to derive detailed and validated data on movement characteristics. In the future this data is to be used for building improved models on the influence of emotions on the size, shape and velocity of body movements. The obtained measurements also can be used directly for performing re-synthesis experiments with ECAs that allow for a comparison of 1:1 copying of movements with model-based synthesis of expressive behaviour.

Synthesis of analysed emotional gestures: As explained in Section 3.2-c, we have performed perceptive tests on emotional expressive videos. We have selected a set of 6 emotional videos from the GEMEP database in which 2 actors performed gestures in 3 emotional states: sadness, anger, happiness. Then we extracted 4 expressivity features using EyesWeb: speed, amplitude, fluidity and energy of movement. We have mapped these features on the expressivity parameters of the Greta virtual agent. We obtained 6 videos of Greta performing gestures with expressivity variations corresponding to the actors' movements. In the first experiment we asked the users to evaluate the actors and Greta's videos in a random order. In the second experiment we have mapped only some subsets of the 4 expressivity features. That is, we excluded one of the 4 features per video to see if the user recognition rate was improved or became worse. Results will help us (i) to understand if our expressivity mapping is correct and (ii) to discover which expressivity features are more/less important in conveying specific emotional states.

b. Speech Expressivity

The next step in the collaboration with PAVOQUE is the actual recording of the expressive speech synthesis corpus. Various expressive speaking styles produced by the same speaker will be recorded. The expressive styles envisaged for recording are inspired by the four SAL characters, coarsely representing the four quadrants in activation-evaluation space. Signal processing tools for changing the expressivity in unit selection synthesis (Schröder, forthcoming) will continue to be developed. We will also attempt to improve the quality of the adaptation in HMM-based speech synthesis.

We will continue working on the research to model blending, dynamics and mixed emotions under the limitation of different speech synthesis methods. One approach is the work in the Emotion Incubator W3C Group which aims at defining a language to describe emotional expression.

5 References

[**Abrilian et al 05**] Abrilian, S., Devillers, L., Buisine, S. and Martin, J.-C. EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. 11th Int. Conf. Human-Computer Interaction (HCII'2005), Las Vegas, Nevada, USA, 22 - 27 July Electronic proceedings, LEA, 2005.

[**Abrilian et al 06**] Abrilian, S., Martin, J.-C., Buisine, S. and Devillers, L. Perception of Movement Expressivity in Emotional TV Interviews. Humaine Summer School, Genova, Italy, September 22-28, 2006.

[**Aubergé and Rilliard 06**] Aubergé, V. and Rilliard, A., Le focus prosodique n'est pas que déictique, le modèle VID (Valence-Intensité-Domaine), JEP'2006, Dinard, France, June 2006.

[**Aubergé et al 06a**] Audibert, N., Vincent, D., Aubergé, V. and Rosec, O. Expressive speech synthesis: evaluation of a voice quality centered coder on the different acoustic dimensions, Speech Prosody 2006, Dresden, Germany, May 2006.

[**Aubergé et al 06b**] Audibert, N., Vincent, D., Aubergé, V. and Rosec, O. Evaluation of expressive speech synthesis, LREC'2006 Workshop on Corpora for research on emotion and affect, Genoa, Italy, May 2006.

[**Baenziger et al 06**] Baenziger T., Pirker H., Scherer K.: GEMEP - GENEVA Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions, in Devillers L., et al.(eds.), *Proceedings of LREC'06 Workshop on Corpora for Research on Emotion and Affect*, May 23, Genoa, Italy, pp.15-19, 2006.

[**Bevacqua et al 06**] Bevacqua, E., Raouzaïou, A., Peters, C., Cardakis, G., Karpouzis, K., Pelachaud, C. and Mancini, M. Multimodal Sensing, Interpretation and Copying of Movements by a Virtual Agent, Perception and Interaction Technologies, Kloster Irsee, Germany, June 19-21, 2006.

[**Buisine et al 2006**] Buisine, S., Abrilian, S., Niewiadomski, R., Martin, J.-C., Devillers, L., Pelachaud, C. Perception of Blended Emotions: from Video Corpus to Expressive Agent. 6th International Conference on Intelligent Virtual Agents (IVA'2006), 21-23 August 2006, Marina del Rey, USA. LNAI 4133, Springer, pp. 93-106. BEST PAPER AWARD.

[**Burkhardt et al 06**] Burkhardt, F., Audibert, N., Malatesta, L., Türk, O., Arslan, L. and Auberge, V. Emotional Prosody - Does Culture Make A Difference?, Proc. Speech Prosody 2006.

[**Castellano et al 06a**] Castellano, G., Camurri, A., and Scherer, K. Expressive gesture and music: analysis of emotional behaviour in music performances (in preparation).

[**Castellano et al 06b**] Castellano, G., Bresin, R., Camurri, A., Volpe, G. User-centered control of audio and visual feedback by full-body movements. Proceedings of Second International Conference on Affective Computing and Intelligent Interaction (ACII2007), Lisbon, September 2007..

[**Devillers et al 06**] Devillers, L., Cowie, R., Martin, J.-C., Douglas-Cowie, E., Abrilian, S. and McRorie, M. Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. 5th International

Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, 24-27 May 2006.

[**Garcia-Rojas et al 06**] Garcia-Rojas, A., Vexo, F. and Thalmann, D. Individualised Reaction Movements for Virtual Humans, In *4th International Conference on Computer Graphics and Interactive techniques in Australasia and Southeast Asia*, pages 79-85, 2006.

[**Garcia-Rojas et al 07**] A. Garcia-Rojas, F. Vexo, and D. Thalmann. Semantic Representation of Individualized Reaction Movements for Virtual Humans. *International Journal of Virtual Reality*, 6(1):25-32, 2007.

[**Gobl et al 06**] Gobl, C. Yanushevskaya, I., and Ní Chasaide, A. (2006). Cross-language perception of voice and affect. *Journal of the Acoustical Society of America*, 120, 3290.

[**Mancini et al 06**] Mancini, M., Castellano, G., Bevacqua, E. and Peters, C., Copying behaviour of expressive motion, (submitted, Mirage 2007).

[**Martin et al 05**] Martin, J.-C., Abrilian, S. and Devillers, L. (2005). Annotating Multimodal Behaviours Occurring during Non Basic Emotions. 1st Int. Conf. Affective Computing and Intelligent Interaction (ACII'2005), Beijing, China, October 22-24 Springer-Verlag Berlin.550-557.

[**Martin et al 06**] Martin, J.-C., Caridakis, G., Devillers, L., Karpouzis, K. and Abrilian, S. Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviours: Validating the Annotation of TV Interviews. Fifth international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, 24-26 May.

[**Martin et al 06b**] Martin, J.-C., Niewiadomski, R., Devillers, L., Buisine, S., Pelachaud, C. Multimodal Complex Emotions: Gesture Expressivity And Blended Facial Expressions. Special issue of the Journal of Humanoid Robotics. Eds: C. Pelachaud, L. Canamero. Vol. 3, No. 3, September 2006, 269-291.

[**Moccozet et al 06**] Moccozet, L., Garcia-Rojas, A., Vexo, F., Thalmann, D. and Magnenat-Thalmann, N., In Search for your own Virtual Individual. In *Semantics And Digital Media Technology Conference 2006* (to appear).

[**Ní Chasaide et al 06**] Ní Chasaide, A., Wogan, J., Ó Raghallaigh, B., Ní Bhriain, Á., Zoerner, E., Berthelsen, H. and Gobl, C. (2006). Speech Technology for Minority Languages: the Case of Irish (Gaelic). *Proceedings of the 9th International Conference on Spoken Language Processing, INTERSPEECH 2006*, Pittsburgh.

[**Peters 06a**] Peters, C. A Perceptually-Based Theory of Mind Model for Agent Interaction Initiation, Special Issue of the Journal of Humanoid Robotics, Special Edition "Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids", Eds: C. Pelachaud and L. Canamero, 3(3), September 2006.

[**Peters 06b**] Peters, C. Evaluating Perception of Interaction Initiation in Virtual Environments using Humanoid Agents (2006). In *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI)*, pp. 46–50, Riva Del Garda, Italy August 2006.

[**Peters 06c**] Peters, C. Designing Synthetic Memory Systems for Supporting Autonomous Embodied Agent Behaviour, *Proceedings of the 15th International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 14–19, September 2006.

[**Peters et al 06**] Peters, C., Pelachaud, C., Bevacqua, E., Ochs, M., Chafai, N. and Mancini M. Social Capabilities for Autonomous Virtual Characters, M. Proceedings of iDiG, International Digital Games Conference, Games Congress 2006, pp. 37-48, Portalegre, Portugal, September, 2006.

[**Pfleger and Löckelt 06**] Pfleger, N. and Löckelt, M. A Comprehensive Context Model for Multi-party Interactions with Virtual Characters. Proceedings of the 6th International Conference on Intelligent Virtual Agents (IVA 2006), pp. 157 – 168, 2006.

[**Schröder 06**] Schröder, M. Expressing degree of activation in synthetic speech. *IEEE Transactions on Audio, Speech and Language Processing*, 14 (4), pp. 1128-1136, 2006.

[**Schröder et al 06a**] Schröder, M., Hunecke, A. and Krstulovic, S. OpenMary - open source unit selection as the basis for research on expressive synthesis, *Blizzard Challenge Workshop 2006*, Pittsburgh, PA, USA, 2006.

[**Schröder et al 06b**] Schröder, M., Heylen, D. and Poggi, I. Perception of non-verbal emotional listener feedback. *Proc. Speech Prosody 2006*, Dresden, Germany, 2006.

[**Vincent et al 05**] Vincent, D., Rosec, O. and Chonavel, T. Estimation of LF Glottal Source Parameters Based on an ARX Model, *Interspeech'2005*, pp. 333-336, Lisbon, Portugal, September 2005.

[**Yanushevskaya et al 06**] Yanushevskaya, I., Gobl, C. and Ní Chasaide, A. (2006). Mapping Voice to Affect: Japanese listeners. *Proceedings of the 3rd International Conference on Speech Prosody*, Dresden, Germany.