humaine

D6d
Proposal for exemplars and work towards them:
Emotion in Interaction

Workpackage 6 Deliverable

ist
information
society
technologies

Date: 30th November 2005

IST project contract no.	507422
Project title	**HUMAINE** **Human-Machine Interaction Network on Emotions**
Contractual date of delivery	*November 30, 2005*
Actual date of delivery	*November 30, 2005*
Deliverable number	D6d
Deliverable title	Proposal for exemplars and work towards them: Emotion in Interaction
Type	Report
Number of pages	52
WP contributing to the deliverable	WP 6
Task leader	University of Paris 8
Author(s)	E. André, E. Bevacqua, F. Burkhardt, A. Camurri, G. Castellano, L. Devillers, A. Egges, A.Garcias-Rojas, D. Heylen, B. Krenn, N. Magnenat-Thalmann, J.C. Martin, R. Moore, C. Pelachaud, C. Peters, H. Pirker, I. Poggi, G. Raimundo, A. Raouzaiou, M. Rehm, D. Romano, O. Rosec, M. Schröder, G. Volpe, P. Wallis
EC Project Officer	Philippe Gelin

Address of lead author: Catherine Pelachaud
                IUT de Montreuil
                Université de Paris VIII
               140 rue de la Nouvelle France
               93100 Montreuil
               France

Table of Contents

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. (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 25 months of the project. They are necessarily provisional, because they will be 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).

The following persons have contributed to the work reported in the deliverable:

E. André, E. Bevacqua, R. Bresin, F. Burkhardt, A. Camurri, G. Castellano, N. DeCarolis, L. Devillers, A. Egges, A.Garcias-Rojas, S. Gupta, O. Gurnasson, D. Heylen, R. Kooijman, B. Krenn, M. Lamolle, M. Mancini, J.C. Martin, V. Maffiolo, N. Magnenat-Thalmann, R. Moore, J.A. Palacios, A.A. Palacios, C. Pelachaud, C. Peters, H. Pirker, I. Poggi, G. Raimundo, A. Raouzaiou, M. Rehm, D. Reidsma, D. Romano, O. Rosec, M. Schröder, A. Sharizan, G. Volpe, M. Walker, P. Wallis, Yorick Wilks.

The institutions that have contributed are:

DFKI, ICCS, UH, Paris 8, OFAI, KTH, UA, EPFL, DIST, INESC ID, CNRS-LIMSI , FT-RD, USFD, TS, Utwente, CNR, U-Bari, MIRALab-UNIGE, ISTC

2 Brief overview of Workpackage 6 and the exemplar proposal

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.

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:

    o Cognitive influences on action

- Interaction Domain:

    o Create affective awareness

    o Backchanneling

- Generation Domain:

    o Coordination of signs in multiple modalities

    o Produce dynamic expressive visual and auditory behaviours

2.2.1 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. 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 have taken 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 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.

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, the user's expressive gestures and language usage are analysed to infer his/her emotional state – the inferred state is then be used to alter the 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 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.

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 How the subtasks link to each other

Several means are being used to ensure links between subtasks:

- Representation languages: are 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 has been devoted to the specification of one of these representation languages (cf the description of the tasks Gesticon and EmoLang). These languages are also essential in the creation of mock-up systems.

- Collaboration: Several partners are working conjunctly on a same subtask. Exchange programs sponsored by Humaine have been used and are going to be used to allow partners to visit each other and work more closely.

- Subtasks: that have been designed within WP6 are complementary one from another. But they are also necessary to create an affective interactive ECA. Such an ECA can not have the capability to perceive the users to be able to interact expressively with him.

2.2.3 How the subtasks link to other aspects of HUMAINE

- WP3: appraisal model; defining the proper set of parameters to characterize expressive behaviour; which behaviours are relevant to convey emotion. WP3 has and will continue to provide information of fundamental issues (eg, property of facial expressions of emotions)

- WP4: analysis of audio and visual signals; analysis and recognition of user's emotional state; analysis of the dynamic properties of user's behaviours. Tight links have been created with WP4 and will continue as it is important for an interactive

system to perceive the user's emotional state. A new emphasis of this collaboration will concern the detection of emotional states differing from the classic 6 prototypes emotions. States linked to the level of engagement a user is in a conversation will be detected. Level of engagement is linked to the level of attention the user is paying to the ECA as well as to the level of interest the user is showing. These levels are defined by the gaze direction of the user, his amount of gaze toward the ECA, his head direction, etc. Technologies defined with WP4 allow one to extract these visual signals.

- WP5: collecting and labelling data; defining what to code and how to code it. Collaboration on the EmoTV corpus has already happened. Further collaboration, specially regarding corpus on backchannel, is foreseen.

- WP7: models of planning, attention and decision making; architectures underlying low-level imitation and the influence of affect in it. A lot of exchanges are happening with WP7. Cognition is an essential issue in the creation of an interactive ECA within an interactive system.

- WP8: models of engagement. During an interaction, one needs to check how the outputted message reaches one's interact. With WP8 we are building a model that uses backchannels as cues of success during persuasion task.

- WP9: While in WP6 we are interested in specifying capability an affective interactive ECA ought to endow, in collaboration with WP9, we aim to understand which of these capabilities are the most important for a given application. We will use user-centered method to reach our goal.

- WP10: Question of user's manipulation, of invading user's space, etc are discussed with WP10.

3 The planned program of research

3.1 Element 1: Cognitive Influences on Action

3.1.1 Leader

Christopher Peters, University of Paris 8

3.1.2 Main participants

University of Augsburg (UA): Elisabeth André, Matthias Rehm

University of Paris 8 (UP8): Christopher Peters

University Sheffield (USFD): Peter Wallis, Yorick Wilks, Swati Gupta, Daniela M. Romano, Marilyn Walker.

3.1.3 Main steps planned towards producing element 1

Subtask	Carried out by	Start / end dates
1) Emotion related attention shifts	UP8	36 (mockup) 48 (proof of concept)
2) Adapt politeness behaviours and language to the user's emotional state a) Corpus-based analysis of eye gaze behaviours and conversational language in input to measure engagement of human conversational partners b) BDI with Goal Tagged Activities	UA, USDF USFD	36 (mockup) 48 (proof of concept) 36 (mockup) 48 (proof of concept)

These subtasks are detailed below

3.1.3.1 Emotion related attention shifts

ECA's that can establish engagements with others through a social attention mechanism (UP8)

Planned Work:

Our aim is to develop an attention-based capability that allows ECA's or other agent's to establish interactions within a virtual environment using perception of direction of attention and facial expressions of others. In essence, this capability focuses on the computation of attention paid to emotion- and attention-related expressions. The ECA should be able to attend to its environment, using gaze and emotion cues of others to focus attention and infer

interaction details from potential interactants. The core work on this component will surround the construction of a social perception and interpretation module that, by utilising synthetic vision, memory and attention capabilities, will allow an agent to attend to, sense and interpret a number of behaviours relating to attention and emotion of other agents in order to plan autonomous interaction initiation more effectively.

Outputs:

Artificial regions of interest in the environment, where social/emotion-related visual stimuli are treated in a special manner. This output can be used e.g. to drive agent looking behaviour, to drive conversation initiation behaviour.

Attention and interest metrics based on social perception that may be used for planning more plausible agent conversation initiation strategies.

3.1.3.2   Adapt politeness behaviours to the user's emotional state

Corpus-based analysis of eye gaze behaviours to measure engagement of human conversational partners (UA, USFD)

Planned Work:

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. The user's emotional state is one crucial parameter for this process, knowledge of the current interaction context another. Based on this information the capability will not only decide on a specific politeness strategy, but it will decide on appropriate politeness behaviours that mitigate face threats by means of gestures, mimics, language and speech. Not accompanying a joke e.g. by corresponding non-verbal behaviours like smiles or laughs may result in the addressee taking the utterance literally. Based on the insights gained from the SEMMEL-corpus (Rehm and Andre, 2005), a gesture selection mechanism will be developed to enhance the verbal part of the mitigation attempt with accompanying gestures. The capability will also provide means to decide for false emotional displays in order to appear natural. According to DePaulo et al. (1996), emotions are the number one topic people lie about. Thus, the capability has to decide when to display the calculated emotion and when to suppress it for reasons of etiquette and believability.

Although Brown and Levinson (1987) claim their politeness strategies to be universal, the realization of each strategy depends on the available linguistic means of a language. Enhancing the verbal realization of a strategy with non-verbal behaviours like gestures, raises the question of universal applicability of these behaviours. Gesture usage seems to follow culture-specific patterns (Ting-Toomey, 1999), thus to clarify this point, a cross-cultural comparison of non-verbally enhanced politeness behaviour will be conducted in Germany (University of Augsburg) and France (University of Paris 8).

Output:

The outputs of the politeness maker are verbal and non-verbal politeness strategies that will be represented by sequences of dialogue acts annotated in an ECA mark-up language to be agreed on within Humaine.

BDI with Goal Tagged Activities (USFD)

Planned Work:

My interest is in using BDI over CTA (Cognitive Task Analysis) to implement interactive dialog systems. Behaviours such as being polite, gossip, and the expression of status recognition are all behaviours that we do every day, but that we do not need to think about. When dealing with such automatic and goal tagged activities, something other than introspection and CTA, which we have previously used (see Wallis et al), is needed, and the aim is to develop a methodology for collecting believable low-level behaviours in a format suitable for synthetic characters.

The challenge is to capture the automatic behaviour of humans in a way that provides coverage, and provides systematically believable behaviour. How do we express anger, and how does that expression change when talking to a child, a policeman, or the priest? What emotions do we humans use to explain the behaviour of others, and what factors change the way we relate emotion to perceived behaviour?

Output:

1) A methodology for collecting believable low-level behaviours
2) Second workshop at CHI'06 as a follow-up to "Abuse: the darker side of human computer interaction" (INTERACT'05)

3.2 Element 2: Creating Affective Awareness

3.2.1 Leader

University of Sheffield: Daniela M. Romano

3.2.2 Main participants

Università Degli Studi di Genova (DIST): Antonio Camurri, Ginevra Castellano, Gualtiero Volpe (DIST)

University of Augsburg (UA): Elisabeth André, Matthias Rehm

University of Paris 8 (UP8): Catherine Pelachaud, Christopher Peters

University of Sheffield (USFD): Daniela M. Romano, Marylyn Walker, Swati Gupta, Orn Gurnasson, Jorge Arroyo Palacios, Arturo Arroyo Palacios, Ahmad Sharizan

3.2.3 Main steps planned towards producing element 2

Subtask	Carried out by	Start / end dates
1) Creating Affective Awareness This is the ability to create an affective relationship with other humans or objects	UA, USFD	36 (mockup) 48 (proof of concept)

2) Detection and imitation This is the ability to replicate the emotional state gained and reproduce it	USFD	36 (mockup) 48 (proof of concept)
3) Adaptation This area of research look at creating systems that are able to sense the emotional state of the user and adapt his performances to it	DIST, UP8, USFD	36 (mockup) 48 (proof of concept)
4) User's engagement This area of research looks at manner to enhance the user's experience and keep him interested in the task he/she is performing.	DIST, UA, UP8, USFD	36 (mockup) 48 (proof of concept)

3.2.3.1   Creating Affective Awareness

Planned Work:

At UA, based on the insights gained from the SEMMEL-corpus, a gesture selection mechanism will be developed to enhance the verbal part of the mitigation attempt with accompanying gestures. The capability will also provide means to decide for false emotional displays in order to appear natural.

3.2.3.2   Detection and imitation

Planned Work:

At **USFD**, two models (a personality model for a believable adaptable intelligent ECA and a social interaction model with ECA) are going to be integrated to allow the human users to freely converse on selected topics with a synthetic character, exercising all the nuance of the English language, in particular on emotion triggering topics. Moreover, various experiments are planned to study and understand the effect of emotions on the voluntary and involuntary human responses to emotional stimuli.

3.2.3.3   Adaptation

Planned Work:

At **DIST**, we seek the development of a dynamic computational model for emotion recognition and interpretation that goes beyond dealing with basic emotions: the idea would be to look for a correspondence between movement parameters' dynamic and emotion categories or dimensions.

A validation of the model will be performed through the integration of the extracted dynamic in ECAs (by means of possible collaborations with other partners), with the aim to contribute to create systems able to sense the emotional state of the users.

At **USFD**, a personality model for a believable adaptable intelligent synthetic 3D character has been developed. A character driven by the model is able to engage in believable social affective interactions with the users and the other characters in the world. Planned work involves the personality model being installed on a robot able to express emotions with his facial expressions and postures.

3.2.3.4   User's engagement

Planned Work:

At **DIST**, We are going to keep working on the analysis of emotional engagement of subjects exposed to multimodal emotional stimuli.

Checking the emotional state of the users or the emotional content perceived by them in stimuli they have been exposed to, can help to find ways to evaluate user experience in emotion-oriented systems.

At **University of Augsburg**, we elaborate a number of key functional components of research and their relevant high-level inputs and outputs as follows:

Components:
- Methods to monitor the user's level of engagement and means to explicitly influence this process.
- Module for supporting the engagement process, including an empirically validated repertoire of strategies to initiate, maintain, and end connections between agent and user.

Input:
- Regarding the user: what is she looking at (agent, other user, environment), what kind of emotions does she display to whom and with what intensity regarding the environment, e.g., state of the game.

Output:
- Annotated listener and speaker behaviours for the generation domain.

Our planned work here is also very relevant to Creating Affective Awareness.

At the **University of Paris 8** we aim to develop an ECA able to establish, maintain and end interactions. The ECA should be able to perceive the attention and level of interest to communicate of its interlocutor to start and/or to maintain a conversation. Thus, communication is not worthwhile without the interlocutor's engagement. So far we have concentrated on two phases: the establish phase where the prospective speaker decides whether to engage a conversation with the prospective listener by assessing the level of attention of the latter; the maintain phase where the speaker checks whether the conversation is effective by considering the interlocutor's backchannels.

Our model embeds a computational model of attention and visual perception.
- The model of attention requires the computation of:
  - Bottom-up visual attention module: It takes as input visual information from the virtual environment. It outputs the saliency maps based on orientation, intensity and colour.

- Social attention module: It takes as input a list of agent percepts in the 3D environment and outputs which agent's face to look at.
- The model of perception computes for each single agent in the 3D world
  - Social perception module: It takes as input the gaze, body and facial expressions of the other agents in the 3D world. It outputs metrics such as level of interest and attention level
  - Theory of Mind module: It takes as input the interest level and outputs a theory of whether the other is interested in conversing or not.

This work is also relevant to Adaptation.

3.3 Element 3: Backchannel capability

3.3.1 Leader

DFKI (Marc Schröder)

3.3.2 Main participants

University of Paris8: Catherine Pelachaud, Elisabetta Bevacqua, Christopher Peters

University of Twente: Dirk Heylen, Ruben Kooijman

OFAI: Hannes Pirker, Brigitte Krenn

ISTC: Isabella Poggi

DFKI: Marc Schröder

3.3.3 Main steps planned towards producing element 3

Subtask	Carried out by	Start / end dates
Elaboration of a theoretical model	Utwente, UP8, OFAI, ISTC, DFKI	18-30
mock-up implementation	Utwente, UP8, OFAI, ISTC, DFKI	18-36
proof-of-concept implementations	Utwente, UP8, OFAI, ISTC, DFKI	30-48

3.3.4 Proposed ECA sub-system architecture

In the context of the general conceptual architecture for ECAs, we propose a modular architecture for an ECA sub-system capable of providing the functionality needed for generating back-channeling behaviour (Figure 1).
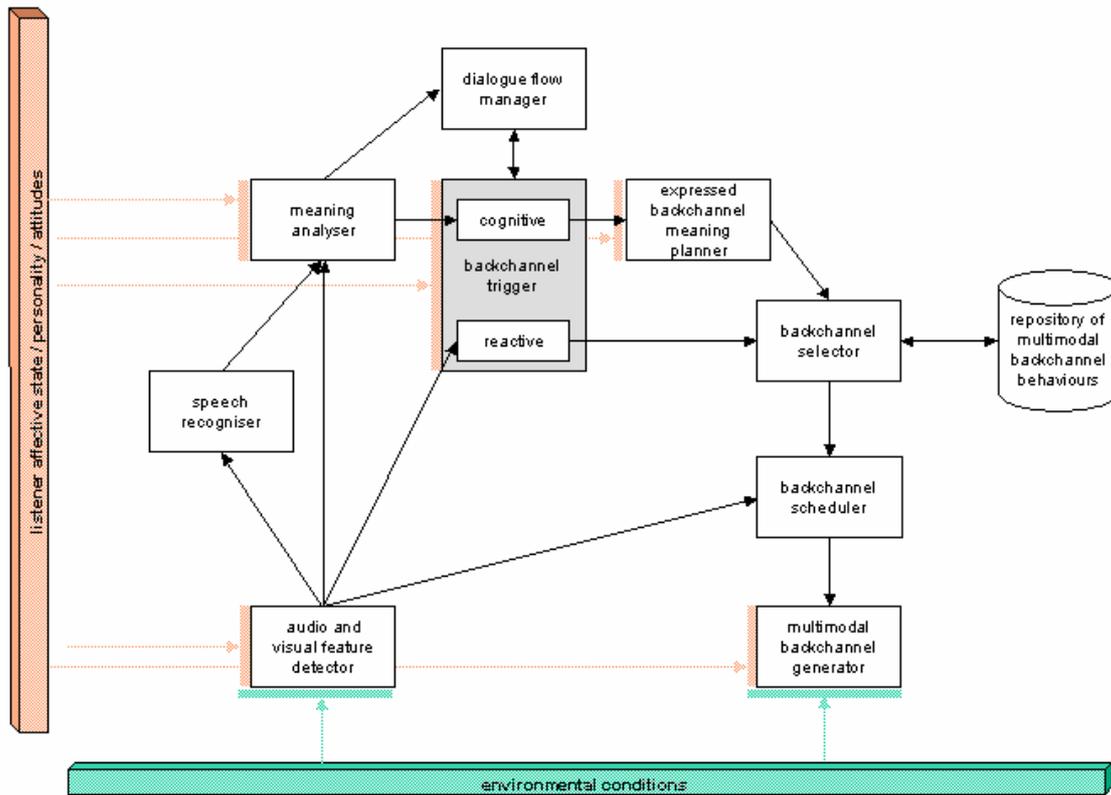
Figure 1: Architecture for backchannel capability

At the heart of the architecture (Figure 1) is the backchannel trigger module, which takes the decision that a backchannel should be produced. This decision relies, in particular, on input from a dialogue flow manager, because backchannels will only be appropriate when the listener does not intend to take the turn, but wants to leave the floor to the speaker. The backchannel trigger is embedded into a perception-interpretation-expression framework.

At the perception end, the audio and visual feature detector module performs basic signal processing tasks. For speech, it can detect pauses, low/mid/high boundary tones, loudness increase, disfluencies, etc. For face and body input, it can detect things like user gaze direction, posture, gestures, head movements, and facial expressions. In combination with the speech recogniser module, the textual content of spoken messages can be obtained. The feature detector may also construct a set of mid-level abstractions, such as syntactic or prosodic phrases.

At the expression end, a backchannel selector determines the form of the backchannel behaviour to generate, by drawing on two types of information. On the one hand, it draws on a repository of multimodal backchannel behaviours, selecting a suitable backchannel for the intended communicative function. On the other hand, the form of the backchannel may be dynamically influenced by the behaviour of the interlocutor, e.g. in imitation. A backchannel scheduler determines the right moments in time to realise these backchannel utterances through the different multimodal generator modules.

In the interaction part of the architecture, two types of perception-expression loop can be distinguished: a reactive and a cognitive loop [Peters.etal05]. Correspondingly, the backchannel trigger has a reactive and a cognitive component.

The reactive backchannel trigger draws its input directly from the low-level and mid-level output of the feature detector. In particular, the meaning of the input is not analysed. The types of backchannel feedback that can be triggered by this reactive component include self-revelation and imitation. Self-revelation feedback covers basic attention and interest signals, as well as the general mood of the listener and his interpersonal stance toward the speaker. However, as meaning analysis is not performed on the input, reactive backchanneling is not a reaction to the content of what the speaker is currently saying, but only reflects the global state of the listener. Imitation feedback is based on a low-level reproduction of speaker behaviour by the listener. The output of the feature detector, including vocal features such as pitch level/range, as well as gaze direction, posture, head movements, gestures, and facial expressions, are used in the backchannel selector in order to determine the form of the backchannel. Such low-level mimicry may lead to plausible displays even if no understanding of the meaning is involved, and might serve as an approximation of empathy displays.

The cognitive backchannel trigger draws on the output of a meaning analyser, which constructs meaning from the input, by interpreting the content of the linguistic message using natural language understanding techniques, as well as the non-verbal input. The meaning is evaluated with respect to the emotions, attitudes, and interpersonal stances contained in the message. Cognitive processing of the message is also performed, leading to a sense of understanding, interest and agreement with the content of the message. When the cognitive backchannel trigger decides that a backchannel reaction to the message is appropriate, a planner component is activated that decides on an appropriate backchannel expression as a feedback response to the input. This may encode the results of the cognitive processing, such as the degree of understanding, interest and agreement. It may also encode the listener's emotions, attitudes and interpersonal stances, which may or may not be the same as the speaker's, which were inferred by the meaning analyser. This mechanism may result in a more sophisticated empathy display than the simple imitation mechanism in the reactive system, in the special case that the listener encodes the same emotion as was decoded from the speaker's message. This more sophisticated type of empathy display may result in more credible behaviour, firstly because it reacts to the actual meaning of the message, and secondly because the surface form of the expressed behaviour may be very different from the one used by the speaker.

3.4 Element 4: Coordination of signs in multiple modalities

3.4.1 Leader

Jean-Claude MARTIN (CNRS-LIMSI)

3.4.2 Main participants

CNRS-LIMSI: Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian

University of Paris8: Catherine Pelachaud, Maurizio Mancini

University of Twente: Zsofia Ruttkay

OFAI: Hannes Pirker, Brigitte Krenn

ISTC: Isabella Poggi

EPFL: Alejandra Garcia Rojas

ICCS: Amaryllis Raouzaiou

3.4.3 Main steps planned towards producing element 4

Three sub capabilities have been identified. For each of them, this document mentions the contributions provided by each partner and the related parts of the schema.

Subtask	Carried out by	Start / end dates
1) From multimodal emotional corpora to models of coordination between modalities A copy-synthesis approach grounded on a corpus of TV interviews and an expressive ECA Video-based models of face and hand expressivity	CNRS-LIMSI, UP8, ICCS	1 - 36 (mockup) 36 - 48 (proof of concept) 1 - 36 (mockup) 36 - 48 (proof of concept)
2) Gesture repositories Representation language for physical properties of gestures	OFAI, UP8	1 - 24 (Gesticon - 1st specification draft) 24 - 36 (Gesticon - 2nd specification draft)
Repertoire of gestures	Univ. Twente	24 - 48 (Gesture repository)
Context dependent emotional body gestures	EPFL	1 - 36 (mockup) 36 - 48 (proof of concept)

These subtasks are detailed below

3.4.3.1 From multimodal emotional corpora to models of coordination between modalities

Planned work:

There has been a lot of psychological researches on emotion and nonverbal communication of facial and vocal expressions of emotions (Ekman 1999), and also on expressive body movements (DeMeijer 1989; Wallbott 1998). Yet, these psychological studies were based mostly on acted basic emotions or limited with respect to the number of modalities. Thus, real-life multimodal corpora are indeed very few despite the general agreement that it is necessary to collect audio-visual databases that highlight naturalistic expressions of emotions as suggested by (Douglas-Cowie et al. 2003). Moreover, results from the literature in Psychology are very useful for the specification of Embodied Conversational Agents, but yet provide few details, nor do they study variations about the contextual factors of multimodal emotional behaviour. Very few researchers have been using context specific multimodal corpora for the specification of an ECA (Kipp 2004).

The study of multimodal emotional corpora for the modeling of coordination between several modalities is studied over 3 projects:

1. *Corpus analysis (CNRS-LIMSI)*: This work is done in strict collaboration with WP5.
    a. Input (from WP5) = annotations of emotion and multimodal behaviours observed in a video (cf. the EmoTV coding scheme)
    b. Output = multi-level emotional expressive profile of a video
        i. emotional profile (soft vector combining verbal labels, valence, intensity),
        ii. multimodal profile (activation rates in each modality),
        iii. expressivity profile (% of each value for the expressivity parameters)
2. *Video-based models of face and hand expressivity (ICCS):* This work is done in strict collaboration with WP4.
    a. Input :
        i. analysis of videos of facial expressions and features extraction
        ii. image sequences of videos captured at an acted session including actors every one of them performing different gestures.
    b. Output :
        i. intermediate facial expressions
        ii. temporal evaluation of emotions (face-gestures)
        iii. definition of expressivity parameters for the gestures
        iv. synthesized gestures
3. *Computational model for expressive gestures and facial expressions (UP8 / KTH):* This work is also partially being developed as part of the Expressivity Capability
    a. Input: annotations of emotion and multimodal behaviours observed in a video
    b. Output:
        i. generation of ECA animation
        ii. computation of the expressive gesture animation
        iii. computation of the expressive facial expression animation

2) Gesture repositories

This task is worked out along 3 main lines:

1. Representation language for physical properties of gestures (OFAI and UP8): use the representation language which has been developed for building a Gesticon, where the

representation language is used in order to encode the mapping between "function" (i.e. a meaning, an emotional content, a pragmatic function etc.) and a certain "form" (i.e. a gesture, posture, facial expression). Refine the first draft of the Representation Language.

2. Repertoire of gestures (Univ. Twente): Twente works on a repertoire of gestures for agents giving a presentation and agents giving directions. This involves both hand and arm gestures. In the presentation case we also look at head movements and gaze.

Context dependent emotional body gestures (EPFL): In Vrlab-EPFL we aim to increase believability in ECA's animation by addressing the Reflex movements as unconscious behaviour. Reflex movements normally are repetitive and robotic sequences. To improve believability in this behaviour is useful the incorporation of descriptors of human characteristics as personality, gender, etc. We work on a Semantic based representation of reflex supported by the concept of ontology. With this ontology model we want to be able to generate a solution space to obtain a suitable reflex animation of individualized ECAs.

3.5 Element 5: Expressivity

3.5.1 Leader

OFAI: Hannes Pirker

3.5.2 Main participants

DFKI: Marc Schröder

EPFL: Alejandra Garcia Rojas

FT-RD: Valerie Maffiolo, Olivier Rosec

INESC: Guilherme Raimundo

CNRS-LIMSI : Laurence Devillers, Jean-Claude Martin

University of Genève MiraLab: Arjan Egges, Nadia Magnenat-Thalmann

OFAI: Hannes Pirker, Brigitte Krenn

University of Paris8: Catherine Pelachaud, Maurizio Mancini

T-S: Felix Burkhardt

Università Degli Studi di Genova (DIST): Antonio Camurri, Ginevra Castellano, Gualtiero Volpe

3.5.3 Main steps planned towards producing element 5

In order to tackle the vast area of modelling and generating expressive behaviour in ECAs, this capability within WP6 was sub-divided into different sub-capabilities. We keep up the distinction between Behaviour expressivity (comprising of facial expression, gesture and posture), which, according to Deliverable D6c "aim(s) the qualitative aspects of co-verbal behaviour executed by Embodied Conversational Agents" and Speech expressivity. In this

section we thus mainly deal with single modalities, but of course there is no strict separation from the work described in section 3.4, which is explicitly concerned with multi-modality

Subtask	Carried out by	Start / end dates
Expressive Behaviour - Expressive Gestures - Complex Facial expressions: Blending	 UP8, KTH UP8	1- 48 1-24 12-36
Expressive Speech – Blending of emotions – Control of voice-quality Copy synthesis	 T-S FT, DFKI FT, T-S	1-48

3.5.3.1   Expressive behaviour

Work on expressive behaviour constitutes the main body of research in the "Expressivity"-capability. It comprises the specification and animation of facial expressions, gestures and postures.

Gesture Expressivity (DIST, KTH, UP8)

UP8 and KTH aim at further developing the GretaMusic application. In particular, in collaboration with WP9, they will work out an evaluation study to test the impact of the GretaMusic application on user's emotional experience. As a way to measure it, user will select interactively which facial expressions and expressivity parameters that better characterized his emotional experience. A set of high level parameters to describe facial expressions and expressivity parameters will have to be defined.

We consider the analysis of expressive gesture of the user, e.g. the way he/she moves during an emotional phenomenon, as the starting point to model the relationship between high level semantic space such as emotions and low level physic space such as movement parameters to be integrated in the animation of an ECA. Our intention is to keep working on analysis of expressive qualities of movement and on a refinement of algorithms and techniques our layered approach is based on. Validation of our analysis algorithms and techniques could be performed through possible integration of the extracted dynamics in ECAs (by means of possible collaboration with other partners).

Facial Expressivity: Blending of Emotions

UP8 intends to extend their model on complex facial expressions to include the combination of the expressivity parameters of the blended emotions. This will enable them to deal with the masked behaviours observed in real data (LIMSI's EmoTV-corpus) and apply the copy-synthesis approach that they have defined for gestures. They also will extend their model to include dynamic facial expressions and other types of emotion blending (eg, sequence of expressions).

Graphical platforms for ECA animation (MIRALAB, INESC)

MIRALab has developed an integrated face and body animation engine specifically targeted toward ECA systems. This system allows synthesizing and mixing different face and body motions on the fly. At the core of this animation engine is a blending library independent of the animation type. This library defines a basic data structure and tools for performing different operations on animations (face or body).

In the future MIRALAB will work on a novel method to increase realism of communicative body animations from an animation defined only by a few joints. Depending on the emotional state of the ECA, different idle motions (such as balance shifts and small posture variations) will be generated. A realistic body animation will be obtained by mixing the idle motions with gesture motions.

In a first test version of this system, the animation engine was integrated with a dialogue manager and a speech synthesizer/recognizer. Currently, the gesture motions are planned by manually placing tags in the response text from the dialogue manager. In collaboration with other HUMAINE partners the integration with other high-level representation-languages will be investigated, future work will focus on integrating the animation engine with other representation languages. This will not only allow other researchers in the field to test their systems with an animation engine; it will also allow testing the flexibility of the system.

INESC is working on a system that allows the use of 3D characters with gestures and facial expression. The purpose of this system is to allow researchers to focus on higher level problems without having to worry about graphical implementation details. This contains both editors to create facial animation parameters and players to receive animation parameters. At least one model will be supplied that allows animation with FACS and MPEG-4 Facial Animation Parameters.

3.5.3.2   Expressive speech

The work on expressive speech in WP6 concentrates on the improvement of expressivity in synthesized speech. Traditionally this work is divided into the aspects of (1) specifying the input parameters that characterize a certain expressive content and of (2) developing and improving the technical means for actually allowing control over these parameters in speech synthesis.

The following topics will be tackled:

Blending of emotions

As seen from natural data [Cowie 03], emotional expression can rarely be classified as one clear expression of a so-called « full blown » emotional state, but will consist of a blending of several emotions. Nonetheless to our knowledge there have been not many attempts yet to simulate emotion-blending in speech-synthesis. An emotional speech simulator [Burkhardt 05] is to be used, that modifies prosodic, voice qualitative and aritculatory parameters to simulate emotional expression. T-S will experiment on emotion blending by interpolation of the rules representing to emotions to be blended.

Control of voice quality in speech synthesis

The lack of ability to not only control prosody but also voice quality parameters in (high quality) speech synthesis, is one of the real obstacles to improve on the current state of expressive speech synthesis. This problem will be tackled in two different ways. FT will work on improved models of speech-parameterisation, namely new versions of the source-filter

model, that allow for the manipulation of both parameters of the source (e.g. open quotient) and the articulators.

DFKI is working on novel methods for creating and manipulating the databases (e.g. diphone-inventories) used in concatenative synthesizers. In their line of research they are experimenting with methods for interpolating the spectral envelope of different input-samples in order to manipulate the voice-quality parameters of these samples.

Copy synthesis of emotional speech

Copy synthesis is one of the most valuable test-beds for actually developing and evaluating methods for expressive speech synthesis. By using parameters derived from natural speech it provides the means to de-couple the problem of synthesis proper from the problem of obtaining input parameters for the synthesis component. T-S will work on improved tools for copy synthesis and prosody transplantation.

Subjective tests will be led by FT in order to evaluate the different techniques of copy-synthesis of emotional speech available in WP6 (not restricted to TD-PSOLA, HNM and formant synthesisers). From an expressive/emotional point of view, the speech sequences will be evaluated by a panel of listeners in the laboratory according to their similarity and the fidelity between each synthetic and target sequences, acceptability naturalness and preference.

3.6 Element 6: Emotion Annotation and Representation Language

3.6.1 Leader

DFKI: Marc Schröder

3.6.2 Main participants

University of Paris 8: Myriam Lamolle, Catherine Pelachaud

OFAI: Hannes Pirker, Brigitte Krenn

Università di BARI: Nadja De Carolis

University of Twente: Dennis Reidsma

DFKI: Marc Schröder

3.6.3 Main steps towards producing element 6

As the main project in Europe addressing emotions in technical environments, HUMAINE is optimally placed to propose a versatile, re-usable technical format for representing and annotating emotions, which should be usable with a large variety of technical environments and theoretical models of emotion. We are proposing such a markup and representation language within WP6, in interaction with the other WPs.

Starting from a thorough collection of use cases and resulting requirements (see Section 4.6), we have formulated a first draft specification of such a language with the working title

"Emotion Annotation and Representation Language (EARL)". The draft specification is now sufficiently complete to be presented for comment to members of the other WPs. Minor points in the specification are still under discussion.

Feedback from the wider HUMAINE group will be used to improve the language syntax, before a first public proposal is made.

Subtask	Carried out by	Start / end dates
Get feedback from other WPs on usability in their environment	DFKI	months 23-28
Report on a full specification of the representation language, including use cases and examples	DFKI, UP8, OFAI, DI-BARI	months 23-29
Define and implement mapping mechanisms between different emotion representations within EARL	DFKI, OFAI, UP8, DI-BARI	30-42

3.7 Steps to ensure co-ordination

Coordination within WP6 will be ensured using means available within Humaine: use of the portals, exchange programs between partners, phone meeting, integration of the work using mock up modules. A coordination hierarchy has been established to provide an intermediate layer of coordination and collaboration for more effective management of WP6 tasks. This layer consists of the workgroup leaders of WP6 and is placed between the workpackage leader and the numerous Humaine partners. When working with individual partners, workgroup leaders provide coordination on a micro-scale with respect to individual capabilities. When liasing with the workpackage leader, they enable effective communication for macro-level planning of integration between capabilities.

Several partners are collaborating together on projects. Collaboration may take various forms, such as the share of the same data (e.g. EmoTV), evaluation studies on different cultures represented by the partners (e.g. polite studies), same tools (ECA system).

3.8 Steps to ensure dissemination

Works done within WP6 will be published in international journals and conferences. Special journal issues will be foreseen. Workshop on the topic of a capability (e.g. backchannel) will be organized and is to be organized (e.g., workshop on "multimodal corpora" at LREC'2006 (May 2006)).

4   Research achievements to date

4.1   Achievement 1: Cognitive Influences on Action

4.1.1   Participants

University of Augsburg (UA): Elisabeth André, Matthias Rehm

University of Paris 8 (UP8): Christopher Peters

University Sheffield (USFD): Peter Wallis, Yorick Wilks

4.1.2   Background of achievements to date

We have made steps towards proposing a theoretical architecture for this capability (Figure 2), which involves a large element of perception (both synthetic agent perception and real-world user perception studies). It also features planning and action selection mock-up modules for providing some form of behaviour feedback for driving agent animation. Here, 'social' is used to encompass the dual notions of emotion and interaction. The capability integrates as follows: Emotion-related attention shifts seek to provide the agent with a sense of social presence by allowing it to react or attend to emotional stimuli in a more natural manner. Corpus-based analysis from real users is vitally important for obtaining real-world data with which computational modelling of social perception can be enhanced, while BDI based studies are useful for analysing and improving the perception that humans have of resulting agent behaviours.
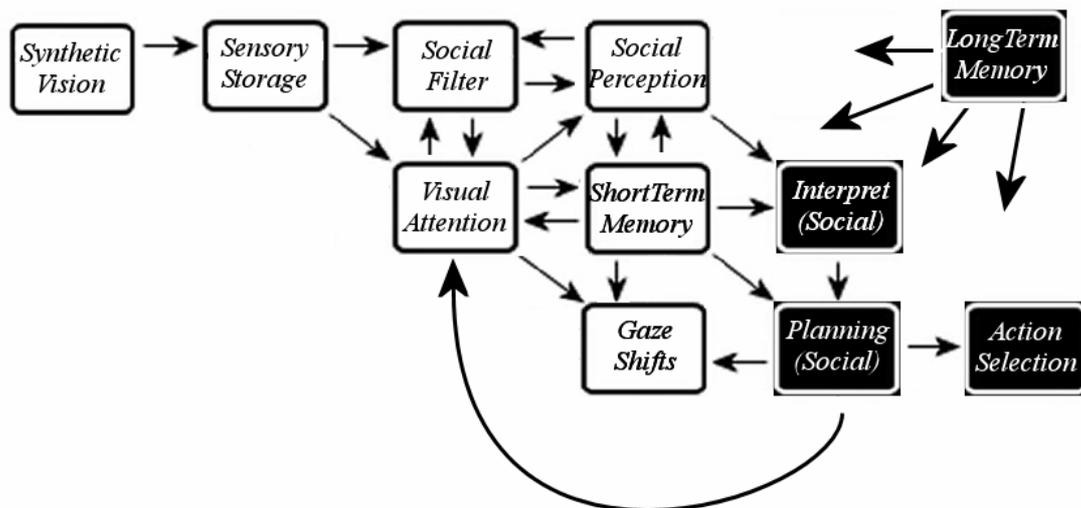

Figure 2: Proposed theoretical architecture for cognitive influences on action capability

Emotion-related attention shifts (UP8)

A fundamental aspect in social interaction is how the interaction is established in the first place, as well as the behaviours that may occur when one of the parties does not wish to interact. Our work is inspired by examinations of human behaviour described by Kendon 1990 and Goffman 1963, where conversation opening involves perceiving cues and sending a sequence of cues including directing of attention, gestures and facial expressions. It is further noted that the process may be complicated by the fact that we often seek to avoid the social embarrassment of engaging in interaction with an unwilling participant. This suggests a grace and subtly as both parties try to figure out whether the other will reciprocate in any conversation opening attempts. Based on completed work featuring synthetic vision, bottom-up attention and short-term memory in order to direct an agents gaze during idle viewing. 'Social attention' here refers to attention drawn due to socially relevant behaviours, in particular emotional behaviour linked to gaze and the eyes and of importance to interaction (Peters 2005). An initial gaze and attention driven real-time embodied agent has been created in a virtual environment to test and allow user perception studies to be conducted.

Corpus-based analysis of eye gaze behaviours to measure engagement of human conversational partners (UA)

The politeness maker is based on Brown & Levinson's theory of politeness. According to their theory, people maintain positive (self image) and negative face (wants and desires), which are continuously threatened during interactions, e.g., by commands or criticism on one's behaviour. Such speech acts are called face threatening acts (FTAs). People try to mitigate such face threats by addressing the addressee's positive face, negative face or both.

The politeness maker requires the following information 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.

Relying on the Gamble system as an evaluation platform, we investigated the level of users engagement with one another and with an ECA in the emotion rich context of a small game of dice that forces the participants to trust or try to deceive each other sometimes during the game.

Based on the results and experience gained in the first user study, a module for initiating an interaction has been developed: So far, users were instructed about the game and then started playing with the agent. It is inevitable that the agent exhibits a kind of initiating behaviour to attract the users for the interaction.

In Rehm and André (2005), we described a corpus study to shed light on the co-occurrence of gestures and verbal politeness strategies in face threatening situations.

BDI with Goal Tagged Activities (USFD)

BDI was originally devised to balance reactive and deliberative behaviour, but has also been used (Norling 2004) as a model of human cognition. Using BDI to drive agent action means the agent does what we humans expect it to, for the reasons we expect. The agent behaves in

accordance with the intentional stance (Dennett) - in accordance with our theory of other minds. Sometimes however human actions do not appear rational and our folk psychological explanation often involves an emotional component (de Rosis 04). A child cries because he is tired or a colleague says such and such because she is angry. Whatever the underlying reasons we humans behave the way we do, we humans reason about other minds in terms of mental attitudes and the assumption is that a synthetic character that actually uses mental attitudes to drive action will have realistic and believable behaviour.

As part of the quest for believable behaviour, we need a reliable method of collecting and implementing human strategies for dealing with the world. In the past we (Wallis et al) have used cognitive task analysis (CTA) to collect conversational strategies in a Wizard of Oz scenario. Simply sitting down and introspecting about the way we think we think is notoriously unreliable and the CTA approach provides model of human reasoning based on metal attitudes that is at least systematic.

Using CTA to collect beliefs, goals, plans, and so on highlights the fact that we do not use mental attitudes "all the way down." We do not (usually) reason about where to place each foot for instance when we walk down the street. Similarly we do not (usually) reason about being polite, when to gossip, tell a joke, or ostracise, nor about status and power relations. We do not (usually) reason about when to get angry or frustrated or abusive and yet it seems these types of behaviour play a role in maintaining the social fabric within which we operate. These types of behaviour are (usually) automatic and goal tagged activities (Wallis) provide a simple extension to the BDI mechanism that captures the unthinking nature of these behaviours.

4.1.3 Publications

C. Peters, Towards Direction of Attention Detection for Conversation Initiation in Social Agents, Proceedings of the AISB '05 Symposium on Virtual Social Agents: Social Presence Cues for Virtual Humanoids, Hatfield, England, April 2005.

M. Rehm and E. Andre. Informing the Design of Embodied Conversational Agents by Analyzing Multimodal Politeness Behaviours in Human-Human Communication. In: AISB Symposium for Conversational Informatics, 2005.

C. Peters, Direction of Attention Perception for Conversation Initiation in Virtual Environments, IVA'05 International Working Conference on Intelligent Virtual Agents, September 2005, Greece.

P. Wallis, Mixed Initiative Dialog using Goal Tagged Activities, proceedings of Embodied Conversational Agents: Balanced Perception and Action (AAMAS workshop) New York, 2004

4.1.4 Other output (demonstrations, resources, etc)

Demonstration of direction of attention detection system running on a desktop PC allowing two agents to move around an environment, sense each other using synthetic vision capabilities and engage in interaction behaviours based on goals and perception of looking behaviour of the attended.

Greta Gamble system used an evaluation platform for investigating the level of users engagement with one another and with an ECA in the emotion rich context of a small game of dice that forces the participants to trust or try to deceive each other sometimes during the game.

4.1.5 Follow-up in progress

Evaluation study will soon be undertaken to test some facets of the emotion-related attention shift model, in particular, users' perception of interest and attention based on the behaviours of the agent. These studies will provide invaluable information for the next iteration of design and development of the model.

4.2 Achievement 2: Creating Affective Awareness

4.2.1 Participants

Università Degli Studi di Genova (DIST): Antonio Camurri, Ginevra Castellano, Gualtiero Volpe (DIST)

University of Augsburg (UA): Elisabeth André, Matthias Rehm

University of Paris 8 (UP8): Catherine Pelachaud, Christopher Peters

University of Sheffield (USFD): Daniela Romano, Marylyn Walker, Swati Gupta, Orn Gurnasson, Jorge Arroyo Palacios, Arturo Arroyo Palacios, Ahmad Sharizan

4.2.2 Background of achievements to date

4.2.2.1 Creating Affective Awareness

Relying on the Gamble system as an evaluation platform, **University of Augsburg** investigated the level of users engagement with one another and with an ECA in the emotion rich context of a small game of dice that forces the participants to trust or try to deceive each other sometimes during the game. Two main characteristics of the system make it especially suitable as an evaluation platform for the current purpose.

- Voluntary emotional displays: Trying to deceive the other players, it is often necessary to hide one's felt emotion, (e.g. worry if the lie succeeds), with another emotion which is generally expressed by a smile. This smile is then said to be masking one's real emotion. To evaluate the impact of such masking expressions on the user, we modelled a number of different masks according to Ekman's findings (e.g., 1992).

- Multi-party interaction: In the Gamble system two users play the game with one agent, offering each user a choice of interaction partners and thus allowing for a direct interaction-inherent comparison of the users engagement behaviour towards another human and towards an agent. Evaluating the users' behaviours in such a scenario allows us to formulate a repertoire of strategies for supporting the engagement process.

The results of the first evaluation are described in detail in Rehm and André (2005a) and Rehm and André (2005b). 12 groups of 2 subjects participated, each playing two rounds of 15 minutes with the agent. The attention of the users towards the other human and towards the

agent was compared distinguishing between two conditions: the human/agent as speaker vs. the human/agent as listener. No significant difference was found for the listener condition. Thus, users accepted the agent as a competent participant during the game. In the speaker condition on the other hand, users showed significantly increased attention towards the agent.

4.2.2.2 Detection and Imitation

USFD have studied detection and imitation for the realm of believable social interaction with synthetic characters, the Paris 8 character *Greta* has been used to act as synthetic customer service able to engage in verbal interactions with a human user performing bank-related tasks for an on-line banking system. The system is web-based and makes use of a speech recognition system. In addition a system able to generate generates conversations with different strengths of politeness and variation, has been produced.

We are also currently building a automatic multimodal analyser of user's arousal. The system aims to analyse user's behaviour (facial expressions, gesture, postures) and physiological responses in order to understand his/her emotional and cognitive status, and the technology that would allow such automatic detection.

4.2.2.3 User's engagement

DIST From the beginning of the project, DIST have been attempting to analyse the relationship among music, emotion and movement.

We performed, in collaboration with Geneva Emotion Research Group, a pilot experiment in which participants were asked to move a laser pointer on a wall in front of them during listening to music excerpts (see related publications). With this experiment, we aimed at investigating the power of music in inducing emotions in subjects and a new way to measure their emotional engagement (the movement of the laser).

University Augsburg have been studying user engagement. According to Sidner et al. (2004), engagement is the process by which participants start, maintain, and end their perceived connection during an interaction. Thus, studying engagement processes is key to creating and maintaining a relationship with an ECA. By taking the user's emotional state and reaction into account, such a relationship will become to some level affective and help raising the user's awareness of the agent on the one hand and creating a kind of longer-term connection between user and agent on the other hand.

Based on the results and experience gained from a first user study, a module for supporting the different engagement behaviours is developed:

- Initiating an interaction: So far, users were instructed about the game and then started playing with the agent. It is inevitable that the agent exhibits a kind of initiating behaviour to attract the users for the interaction. This will be based on Svennevig's analysis (1999).

- Maintaining an interaction: The result of the first user study informs a gaze-model for the agent to respond appropriately to the increased attention of the user. Moreover, the comments of the participants during the game showed, that the agent's emotional behaviour was very welcome and increased the interest in the interaction. A thorough analysis of these comments will help in developing strategies to improve the user's affective engagement.

- Ending an interaction: We will need strategies to deal with users' decreased engagement in an interaction to re-initiate or end the interaction.

At **University of Paris 8**, we have already implemented the establish phase of interaction between two agents in a virtual environment using the Torque Engine. It uses the model of attention as well as of perception. The 3D world is continuously sensed by the agent using the perception and bottom-up visual attention modules. The sensed information is sent to the social attention module that detects which extracts socially relevant objects from the scene (e.g. agents and faces). The perception module then decomposes the image of the sensed agent into parts: the eye, the head, trunk, etc. The direction of these body parts are computed from the point of view of the agent that looks around. The attention and interest levels are then computed. The looking agent will decide to start interacting with the looked agent or not if these levels are high enough otherwise it will not attend to start a conversation.

The maintain phase has been developed within the Greta engine. Currently, it models only the gaze behaviours of the speaker.

4.2.2.4 Adaptation

At **USFD**, a personality model for a believable adaptable intelligent synthetic 3D character has been developed. A character driven by the model is able to engage in believable social affective interactions with the users and the other characters in the world. The character is able to detect the events in the world and their emotional content and respond to them also adapting in the long term its personality to the events that happen during its synthetic life.

DIST are designed a layered approach (see related publications) that, starting from multimodal inputs (audio, video, data from sensors, etc.), may allow to interpret the emotional state of users and the emotional communicative intent of artists such as dancers or musicians.

University of Paris 8 has used the Greta system as an early mock-up to allow an ECA to adapt its gaze behaviours based on a limited perception of the gaze and interest of another agent. The gaze of the listener is fixed ahead by few parameters specifying the overall gaze behaviour of the listener. The algorithm is as follows: the system computes the level of interest demonstrated by the listener from the point of view of the speaker; to do so it determines the gaze direction of the listener and computes its level of attention. The level of interest is obtained by integrating over time this value. The effectiveness of conversation is then computed. It corresponds to the evaluation by the speaker of the listener's attitude in the conversation. If the listener is showing low level of interest in participating to the conversation, the speaker will decrease its perception of the effectiveness of the conversation. It will start gazing more at the listener as a way to try to attract the listener's gaze and have him back to be engaged in the conversation. When this value is lower than a given threshold, the speaker decides to stop the conversation.

4.2.3 Publications

Camurri, G. Castellano, B. Mazzarino, G. Varni, G. Volpe *Expressive gesture analysis and expressive gestural control strategies* in Proc. Stockholm Science and Music Science Symposium, Stockholm, Sweden, 2004

A. Camurri, G. Castellano, P. Coletta, A. Massari, B. Mazzarino, G. Varni, G. Volpe *Multimodal interfaces for expressive interaction*, in Proc. HCItaly 2005, Italian Symposium on Human Computer Interaction, Roma, September 2005.

G. Castellano *Experiments, Analysis, and Models of Motor Activation as a Component of an Emotional Process*, Master Thesis, DIST-University of Genoa, InfoMus Lab, October 2004.

A. Camurri, G. Castellano, M. Ricchetti, G. Volpe, *Subject interfaces: measuring bodily activation during an emotional experience of music,* 6th International Gesture Workshop, Vannes, France, May 2005.

Romano, D.M., Sheppard, G., Hall, J., Miller, A., Ma, Z. (2005). BASIC: A Believable, Adaptable Socially Intelligent Character for Social Presence. PRESENCE 2005, The 8th Annual International Workshop on Presence, 21-22 September 2005, University College London, London, UK.

Gupta, S., Romano, D.M., Walker M. A. (2005). Politeness and Variation in Synthetic Social Interaction, H-ACI Human-Animated Characters Interaction Workshop, British HCI 2005 The 19th British HCI Group Annual Conference Napier University, Edinburgh, UK 5-9 September 2005

Dourish, P. and Bly, S., (1992), Portholes: Supporting awareness in a distributed workgroup, *Proceedings of ACM CHI 1992*, 541-547.

C. Peters, C. Pelachaud, M. Mancini, E. Bevacqua, Engagement Capabilities for ECAs, workshop ``Creating bonds with ECAs'', Fourth International Joint Conf. on Autonomous Agents & Multi-Agent Systems, Utrecht, July 2005.

C. Peters, C. Pelachaud, M. Mancini, E. Bevacqua, I. Poggi, A model of attention and interest using gaze behaviour, IVA'05 International Working Conference on Intelligent Virtual Agents, September 2005, Greece.

4.2.4 Other output (demonstrations, resources, etc)

The Greta agent (Paris8) has been used to act as synthetic customer service able to engage in verbal interactions with a human user performing bank-related tasks for an on-line banking system. The system is web-based and makes use of a speech recognition system. In addition a system able to generate generates conversations with different strengths of politeness and variation, has been produced.

4.2.5 Follow-up in progress

Various experiments are planned to study and understand the effect of emotions on the voluntary and involuntary human responses to emotional stimuli.
Moreover the personality model is being installed on a robot able to express emotions with its facial expressions and posture.

4.3 Achievement 3: Backchannel properties and architecture

4.3.1 Participants

University of Paris8: Catherine Pelachaud, Elisabetta Bevacqua, Christopher Peters

University of Twente: Dirk Heylen, Ruben Kooijman

OFAI: Hannes Pirker, Brigitte Krenn

ISTC: Isabella Poggi

DFKI: Marc Schröder

4.3.2 Functions, forms, and timing of backchannel feedback

4.3.2.1 Backchannel functions

Conversation is the principal means used by human beings to communicate and interact with each other. Each sentence is an action (a speech act) that aims at a goal; through words a speaker transmits his goal to the interlocutor (or listener) in order to obtain something (to reach his goal) [Poggi81].

During conversations two flows of information are established between the speaker and the listener. The first one concerns the topic, what the speaker is saying and the message he wants to transmit, the latter, emitted by the listener, is necessary to monitor the conversation. We will refer to the latter as signals in the back-channel. The term back-channel has been used in different ways by different authors. To get an idea of some of the discussion about the use of the term we have include the following quote from [Roger&Bull89].

Yngve (1970) introduced this term to refer to brief utterances (such as 'mmm', 'uh huh', 'yes', etc.) which are used to signal to the speaker the continued interest and attention of the listener. Duncan (1972) identified five forms of back-channel – sentence completions, requests for clarification, brief phrases such as 'uh huh' and 'right', and head nods and head shakes. Duncan (1969), although critical of what he calls 'external variable' approaches, none the less makes extensive use of this classification of back channels as the basis for subsequent quantitative analysis (see Duncan & Fiske, 1977). […] Heritage, in Chapter 2, takes issue with the conventional treatment of 'back-channels' as signals of continued attention, arguing that the role of what he calls 'response tokens' has been substantially underestimated by the use of this classification. Response tokens, Heritage maintains, may serve a whole variety of communicative functions, such as indicating a desire to shift topic (Jefferson, 1981b), acknowledging receipt of information (Heritage, 1984) or to promote telling of 'news' (Jefferson, 1981a).

This shows that authors differ in what behaviours they count in as backchannels and in what they assume the function of these behaviours should be to be called a back-channel. One could say in general, that listeners produce all kinds of behaviours that the speaker can interpret in various ways. Some of this is not explicitly made to "communicate" though it could signal certain information to the speaker (e.g. when the listener sneezes, the speaker may interpret this as a symptom of a cold), another part may be a signal of the emotions or mood of the listener. In most definitions such signals are excluded from the concept of backchanels. Backchannels are typically behaviours that are intended as explicit signals as some kind of feedback on the process of communication. They are determined by what Castelfranchi and Paris call control goals [Castelfranchi&Parisi80]. In Clark's terminology, back-channels are signals in track 2 [Clark96]. In this deliverable we will use the term back-channel quite liberally as feedback of the listener on the different levels processing of what is being said. This deliverable should be seen as a first characterisation of the phenomenon we want to study. As the research progresses we will have occasion to refine the terminology and provide more precise definitions.

The back-channel is essentially used as a cooperative way of exchanging information about the successfulness of communication [Allwood92, Yngve70, Goodwin89, Dittman68, Brunner79, Schegloff82] and about the listener's intention to adopt or refuse the speaker's

goals [Poggi81]. From the literature we can deduce three backchannel functions that correspond to the three human communication levels: establishing and maintaining engagement, comprehension and reaction.

1. Establishing and maintaining engagement

At this level the backchannel provides information about the possibility to start and bring on a conversation; so, it gives feedback about:

Contact: whether the listeners is willing and able to start and maintain a conversation,

Perception: whether the listener is willing and able to perceive the Speaker's speech, for example if he is deaf, or if he can't understand the speaker's language, or if an external event appears to alter the quality of perception and make the conversation impossible.

Attention: whether the listener is willing to pay attention to the speaker.

To generate backchannel at this level of the conversation it is not necessary that the listener understands the meaning of the speaker's words. Here, the signals emitted by the listener are only about the establishment of a conversation and its maintenance.

2. Comprehension.

At the level of comprehension signals in the back-channel provides information about:

Understanding: whether the listener is willing and able to understands the content of the speaker's words,

Interest: whether the listener is interested and involved (engaged) in what the other participant is saying.

Here, a real semantic comprehension of the speaker's discourse is necessary.

3. Reaction

Once the message is understood (or misunderstood) the listener can show his reaction. At this level the backchannel aim to convey the listener's intentions and goals that can or cannot be consistent to speaker's intentions and goals. Here the backchannel provides information about:

Believability: whether the listener believes what the other participant is saying,

Attitude: what are the emotions/attitudes the speaker's discourse is eliciting in the listener (Allwood92),

Agreement: whether the listener agrees and is willing and able to adopt the speaker's goals.

At this level the listener's reaction and then his backchannel are strongly bound to social and cultural aspects, they depend on the listener's knowledge, beliefs and goals, included his ethical-moral values.

Backchannels are not always produced consciously by the listener [Allwood92], sometimes he emits backchannel signals without awareness, in an instinctive manner. For example, some

gestural feedbacks can be provided even without eye contact between the interlocutors [Allwood & Cerrato03].

4.3.2.2 Representation and generation of backchannel signals

Backchannel signals may be generated in different ways depending on three different dimensions:

cognitive / reactive: that distinguishes between what can be done without explicit planning and what needs or resorts to it.

sincere / deceptive: concerning whether the Listener is sincere or has the goal to deceive about his reception and reaction to the speaker's talk: the reactive mode only allows sincere backchanneling, while the cognitive mode is necessary when one has the goal to deceive.

imitation / dictionary: Imitating the speaker's expression and movements is a generic way to provide positive backchannel, that is, of communicating one's alignment to the speaker. But beside this, a more specific way to provide backchannels consists of providing different possible positive or negative signals, differing from each other for the communication level: contact, perception, attention, understanding, interest, believability, attitude, agreement. This implies that in the Speaker's mind it is represented a list of signal – meaning pairs that form a "backchannel dictionary" [Poggi04].

4.3.2.3 Backchannel forms

Backchannel consists of a set of verbal and non-verbal signals emitted by the listener during conversation. Verbal signals are single words, short sentences, simple sounds (like "yeah", "I see", "mh" and "mhm") whereas non-verbal signals are provided by the listener through gaze, facial expressions, head, hand and body movements [Allwood92, Cerrato03, Heylen05a, Heylen05b].

Often the same signal can be produced with different significance according to the voice pitch and tune [Gardner98], the speaker's previous sentence [Allwood93] and the concomitant facial expression; for example, the word "yeah" can express understanding and agreement but can also imply the opposite meaning if uttered with irony.

The message provided by backchannel depends also on the combination of signals that the listener decides to transmit at the same time. For example, head movements [Allwood&Cerrato03], gaze direction and facial expression can reinforce, weaken or contradict the meaning of a verbal feedback.

Notwithstanding this polysemy of backchannel signals, it is possible to single out a systematic correspondence of each backchannel form with a specific meaning or a class of meanings [Poggi02].

4.3.2.4 State of the art

Theoretically, a backchannel signal appears when the speaker is able to perceive it more easily and when, through it, the listener can show his understanding of the conversation so far. Such a moment in an interaction corresponds to a grammatical completion, usually accompanied by a region of low pitch. Many backchannel models are based on this result:

Ward and Tsukahara modelled the location of backchannel in Japanese and English conversation simply by inserting them where the speaker produced a region of low pitch lasting 110 ms [Ward&Tsukahara00], Cassell and Bickmore implemented a backchannel model for REA that puts a signal at each pause that lasts more than 500 ms [Cassell&Bickmore00], Cathcart enhanced this kind of model by inserting a backchannel also every n words, where n is a number deriving from the analysis of a corpus of data [Cathcart03].

Maatman presented a different model that is based on the speaker's behaviour. He presented a list of rules useful to decide which kind of feedback a listener provides as a consequence of a particular action of the speaker [Maatman05]. Those rules, deriving from a corpus of data, are useful to predict when a feedback can occur. For example, backchannel continuers (like head nods, verbal responses) appears at a pitch variation in the speaker's voice; frowns, body movements and gaze shifts are produced when the speaker shows uncertainty; facial expressions, postural and gaze shifts are provided to reflect those made by the speaker (mimicry).

Thórisson developed a talking head capable of producing real-time feedback analysing the user's behaviour [Thorisson94]. The system can perceive user's intonation (obtained with automatic intonation analysis) and hand gestures and gaze (provided by a human observer in a Wizard of Oz manner). This information is used to automatically control the gaze of the talking head, its backchannel and the turn-taking behaviour (which consist of asking question at appropriate points in the dialogue).

4.3.3 Publications

D. Heylen, A Closer Look to Gaze, workshop "Creating bonds with ECAs", Fourth International Joint Conf. on Autonomous Agents & Multi-Agent Systems, Utrecht, July 2005.

D. Heylen, Challenges Ahead. Head Movements and other social acts in conversation, in AISB 2005 - Social Presence Cues Symposium, U. of Hertforshire, 2005.

C. Peters, C. Pelachaud, M. Mancini, E. Bevacqua, Engagement Capabilities for ECAs, workshop "Creating bonds with ECAs", Fourth International Joint Conf. on Autonomous Agents & Multi-Agent Systems, Utrecht, July 2005.

C. Peters, C. Pelachaud, M. Mancini, E. Bevacqua, I. Poggi, A model of attention and interest using gaze behaviour, IVA'05 International Working Conference on Intelligent Virtual Agents, September 2005, Greece.

4.3.4 Follow-up in progress

We are now in the process of defining a number of simplified instantiations of the "optimal" architecture for a backchanneling ECA (Figure 1). These instantiations differ in complexity, from the most simple to slightly more ambitious ones. This will lead to a number of mock-up and proof-of-concept implementations as indicated in Section 3.3.

4.4 Achievement 4: Coordination of signs in multi modalities

4.4.1 Participants

CNRS-LIMSI: Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian

University of Paris8: Catherine Pelachaud, Maurizio Mancini

University of Twente: Zsofia Ruttkay

OFAI: Hannes Pirker, Brigitte Krenn

ISTC: Isabella Poggi

EPFL: Alejandra Garcia Rojas

ICCS: Amaryllis Raouzaiou

4.4.2 Background of achievements to date

From multimodal emotional corpora to models of coordination between modalities

We have defined a copy-synthesis approach grounded on a corpus of TV interviews and an expressive ECA (CNRS-LIMSI, UP8). CNRS-LIMSI has studied the relations between real-life emotions and multimodal behaviours in TV interviews collected in the EmoTV corpus. The manual annotation of a video sample of real-life emotions extracted from the EmoTV corpus enabled to compute realistic values of expressivity parameters that were used for driving the specification of the Greta expressive ECA developed by UP8. This copy-synthesis approach, which is still at an exploratory stage, enabled to identify the relevant levels of representation for studying the complex relation between emotions and multimodal behaviours in non acted and non basic emotions. It allowed the partners to refine the ECA system, the expressivity model and the annotation scheme.

 ICCS has analysed automatically facial expressions and hand gestures from videos of acted behaviours ; it is producing expression profiles for intermediate expressions and expressivity profile based on quantitative measures in gestures. In particular, it computes facial/hand expressivity based on discrete and dimensional models, synchronization of expressive cues in facial expressions and hand gestures, as well as quantitative parameters to render expressivity on neutral hand gestures.

Gesture repositories

A "gesticon" is a repository where the information on communicative gestures, postures, facial expressions is bundled. A gesticon developed along these lines would thus be a component with the following functionality:

- Input: A collection of features describing the envisaged communicative function.

- Output: A set of candidate gestures together with a description of their physical properties, which are to be interpreted, modified and rendered by an animation engine.

OFAI and UP8 have developed a first draft of a representation format for encoding both the form (i.e. the physical properties) and the function (i.e. semantics, pragmatics) of gestures and facial expressions, and thus provide a means for mapping functions to forms within the Gesticon. They have designed the representation language for a symbolic description of communicative gestures, to be used for specifying the entries of a gesture repository. A first draft of the Gesticon can be found on the wiki page of Humaine web site.

4.4.3 Publications

Abrilian, S., Martin, J.-C. and Devillers, L. (2005b). A Corpus-Based Approach for the Modeling of Multimodal Emotional Behaviours for the Specification of Embodied Agents. 11th International Conference on Human-Computer Interaction (HCII'2005), Las Vegas, Nevada, USA, 22 - 27 July

Abrilian, S., Devillers, L., Buisine, S. and Martin, J.-C. (2005a). EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. 11th International Conference on Human-Computer Interaction (HCII'2005), Las Vegas, Nevada, USA, 22 - 27 July LEA

Devillers, L., Abrilian, S. and Martin, J.-C. (2005). Representing real life emotions in audiovisual data with non basic emotional patterns and context features. First International Conference on Affective Computing & Intelligent Interaction (ACII'2005), Beijing, China, October 22-24 Spinger-Verlag Berlin.519-526. 2005.

M. Mancini, B. Hartmann, C. Pelachaud, A. Raouzaiou, K. Karpouzis, Expressive Avatars in MPEG-4, IEEE International Conference on Multimedia & Expo (ICME) 2005, Amsterdam, July 2005.

Martin, J.-C. (2005) Annotation, Interpretation and Synthesis of Cooperation between Modalities in Multimodal Communication. Invited talk at the summer school of the Similar Network of Excellence. August 5th 2005.

Martin, J.-C., Abrilian, S. and Devillers, L. (2005). Annotating Multimodal Behaviours Occurring during Non Basic Emotions. 1st International Conference on Affective Computing & Intelligent Interaction (ACII'2005), Beijing, China, October 22-24 Spinger-Verlag Berlin.550-557. 2005.

Martin, J.-C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M. and Pelachaud, C. (2005). Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs. 5th International Working Conference On Intelligent Virtual Agents (IVA'2005), Kos, Greece, September 12-14. 2005

4.4.4 Other output (demonstrations, resources, etc)

Description and illustrations of models of multimodal emotional behaviours (e.g. emotional/expressivity profiles, different time-scales) computed from the manual annotations of one video sample. Animation file corresponding to several replays of this video sample.

4.4.5 Follow-up in progress

Subtask:

- Copy synthesis: Annotation of other videos involving masked emotions will be undertaken. Currently, we do not use all the annotations provided by the EmoTV corpus. The context annotations include other information related to "appraisal" dimensions such as the time of the event, the implication of the person, etc. which might be interesting to consider in the model of the agent. Other levels might also be relevant (head movements) so as to generate different behaviours with different levels of fidelity. Perceptual tests evaluating if the contextual cues, the emotion and the multimodal behaviours are perceptually equivalent in the original video and in the simulation of the corresponding behaviours by the ECA, thus revealing how much such a technique is successful. These perceptual tests will also help finding out if differences of quality and of level of details between the real and the simulated multimodal behaviours have an impact on the perception of emotion. The copy-synthesis approach will benefit from validations of manual annotations of multimodal behaviour and computation of intercoder agreement.

- Gesture repository: We are creating a database of emotional reflex movement animations and Gesticon Entries. These animations are directly stored in a knowledge base system under emotional parameters. We are refining the first draft of the Gesticon Representation Language

4.5   Achievement 5: Expressivity

4.5.1 Participants

DFKI: Marc Schröder

EPFL: Alejandra Garcia Rojas

FT-RD: Valerie Maffiolo, Olivier Rosec

INESC: Guilherme Raimundo

CNRS-LIMSI : Laurence Devillers, Jean-Claude Martin

University of Genève MiraLab: Arjan Egges, Nadia Magnenat-Thalmann

OFAI: Hannes Pirker, Brigitte Krenn

University of Paris8: Catherine Pelachaud, Maurizio Mancini

T-S: Felix Burkhardt

Università Degli Studi di Genova (DIST): Antonio Camurri, Ginevra Castellano, Gualtiero Volpe

4.5.2 Achievements in Expressive Behaviour

The motion synthesis library that is developed by MIRALab uses motion captured animations to reproduce new motions. The motion captured animations that are at the basis of the system are recorded according to different emotional states and different people [Egges etal 04b]. The effect of portraying emotional motions or postures is thus achieved by selecting the appropriate motions from the pre-recorded database and adapting them. These motions are

then combined with facial expressions. The facial expressions can be modelled by hand or they can be pre-recorded using a motion capture system.

INESC has developed a graphical tool for supporting the animation of ECAs. This application allows the use of several graphical techniques such as key framed animation, of skeleton poses, blending of these animations, inverse kinematic animation of the skeleton, morph targets, definition and use of control parameters such as FAPs and AUs of FACS and integration with festival text to speech with viseme/phoneme synchronization. All these techniques are working in real time.

We have been working on the analysis of 2D/3D human-full body movement both as conveyor of emotional content and as component of an emotional process, in terms of analysis of its expressive qualities. We designed a layered approach that, starting from multimodal inputs (audio, video, data from sensors, etc.), may allow one to interpret the emotional state of users and the emotional communicative intent of artists such as dancers or musicians.

UP8 is animating an agent on the basis of annotations obtained from LIMSI's EmoTV-video corpus. [Martin05]. The manual annotations are transformed to APML tagged text plus values for the expressivity dimensions, which are then used in order to animate the an ECA using the Greta player.

Some studies [Scherer98] deal with more complex emotions (e.g. several modalities or combinations of basic emotions, or real-life emotions). Simultaneous emotions are referred to as blends of emotions. After discussing with Suzanne Kaiser, WP3, on theoretical foundation of facial expressions arising from blends of emotion, UP8 has started to develop a computational model. This work is being done in close relation with the development of Capability 4. We called these expressions 'complex facial expressions'. UP8's model is based on the research by Ekman [Ekman75]: the different blends of facial expressions can be distinguished depending on the type of emotions, their apparition in time (sequence, superposition) as well as if they are felt or fake. For the moment, two different types of blending have been developed: masking and superposition. Both of them are based on fuzzy inference.

UP8 proposes a model of gesture manner, called *gesture expressivity*, that acts on the production of communicative gestures. Its model of expressivity is based on studies by researchers such as (Wallbott, 1998 ; Wallbott and Scherer, 1986 ; Gallaher, 1992). UP8 describes expressivity by a set of 6 dimensions. Each dimension acts on a characteristic of communicative gestures. A detailed description of the implementation of this model can be found in (Hartmann et al, 2005).

UP8 has conducted two studies to evaluate its gesture expressivity model. The goal of the first study was to test the following hypothesis: The chosen implementation for mapping single dimensions of expressivity onto animation parameters is appropriate ? a change in a single dimension can be recognized and correctly attributed by users. The hypothesis tested in the second study was to test if combining parameters in such a way that they reflect a given communicative intent will result in more believable overall impression of the agent. The results of these tests confirm that the general approach for expressivity modelling is worthwhile pursuing. However, only a subset of parameters and a subset of expressions were recognized well by users.

KTH and UP8 have developed a real-time application GretaMusic of an expressive synthetic face displaying nonverbal behaviours in relation with expressive music. As the quality of the music is changed, the associated facial expressions vary. The variation happens at two levels:

at the signal levels and at the level of movement expressivity. that links music and emotional facial expressions. The model of gesture expressivity was extended to head movements.

4.5.3 Achievements in Speech Expressivity

Several projects have been initiated in this area: HUMAINE-members so far have contributed to both of these areas as follows.

T-S developed an open source emotional simulation speech synthesizer based on the freely available Mbrola [Dutoit et al 1996] speech synthesizer [Burkhardt05]. It aims at both scientists and students who want to study the effect of acoustic variation on emotional perception as well as at programmers who want to add emotional expression to their software. The distribution of the software will help to boost the interest in emotion related speech synthesis world-wide. The program might also be used as a module in implementations of other capabilities in WP6 as well as in other Workpackages.

FT has developed a new analysis-synthesis method which enables the separation of glottal source and vocal tract information. This method is based on an ARX model of the speech combined with an LF model of the glottal excitation and is described in [Vincent etal 05]. Given this analysis scheme, the following information are available: F0, vocal quality features (open quotient, asymmetry coefficient, etc.), vocal tract information and a source residue due to mismatch of the actual source to the LF model. The synthesis algorithm has been designed so that all parameters can be modified independently.

DFKI, in cooperation with partners outside HUMAINE, has developed a new method for addressing one of the key obstacles to emotional speech synthesis, viz., the control of voice quality in synthetic speech [Turk etal 05]. Interpolated diphone databases were created by mixing the spectral envelopes of different recorded voice qualities, and it was showed that the voice qualities of the resulting diphone voices were perceived as intended. This currently allows to differentiate the degree of vocal effort, one key voice quality correlate of the arousal dimension of emotions, more finely than what was possible before [Schröder&Grice03].

4.5.4 Publications

[Albrecht et al 05] Albrecht, I., Schröder, M., Haber, J., & Seidel, H. (2005). Mixed feelings: expression of non-basic emotions in a muscle-based talking head. Virtual Reality: 8 (4), 201-212.

[Buisine et al 06] S. Buisine, B. Hartmann, M. Mancini, C. Pelachaud, Conception et Evaluation d'un Modele d'Expressivité pour les Gestes des Agents Conversationnels, Revue en Intelligence Artificielle RIA Édition spécial ``Interaction Emotionnelle'', 2006.

[Burkhardt05] Burkhardt F., Emofilt: the Simulation of Emotional Speech by Prosody-Transformation, Proc. of Interspeech 2005, Lisboa, Portugal.

[Camurri et al 04] A. Camurri, G. Castellano, B. Mazzarino, G. Varni, G. Volpe, Expressive gesture analysis and expressive gestural control strategies in Proc. Stockholm Science and Music Science Symposium, Stockholm, Sweden, 2004

[Camurri et al 05] A. Camurri, G. Castellano, P. Coletta, A. Massari, B. Mazzarino, G. Varni, G. Volpe, Multimodal interfaces for expressive interaction, in Proc. HCItaly 2005, Italian Symposium on Human Computer Interaction, Roma, September 2005.

[Dias05] Dias, J.: FearNot!: Creating Emotional Autonomous Synthetic Characters for Empathic Interactions. Universidade T´ecnica de Lisboa, Instituto Superior T´ecnico, Lisboa, MSc Thesis, 2005

[Dias etal 05] Dias J., Paiva A.: Feeling and Reasoning: a Computational Model for Emotional Agents, EPIA Affective Computing Workshop, 2005.

[Egges etal 04b] Egges A., Molet T., Magnenat-Thalmann N. Personalised Real-time Idle Motion Synthesis (Seoul, Korea, 2004). Pacific Graphics 2004.

[Hartmann etal 05] B. Hartmann, M. Mancini, and C. Pelachaud, Implementing Expressive Gesture Synthesis for Embodied Conversational Agents, in *Gesture Workshop*, Vannes, May 2005.

[Mancini etal 05] M. Mancini, R. Bresin, C. Pelachaud, From acoustic cues to expressive ECAs, Gesture Workshop, Vannes, May 2005.

[Martin etal 05] Martin J.-C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M. and Pelachaud, C. Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs. 5th International Working Conference On Intelligent Virtual Agents(IVA'2005), Kos, Greece, September 12-14 Springer. pp 405-417. 2005

[Ochs etal 05] M. Ochs, R. Niewiadomski, C. Pelachaud, D. Sadek, Intelligent Expressions of Emotions, 1st International Conference on Affective Computing and Intelligent Interaction ACII, Chine, October 2005.

[Schröder 04] Schröder, M. (2004). Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. Proc. Workshop on Affective Dialogue Systems: Lecture Notes in Computer Science (pp. 209–220). Kloster Irsee, Germany.

[Schröder 04] Schröder, M. (2004). Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis (Ph.D thesis). Vol. 7 of Phonus, Research Report of the Institute of Phonetics, Saarland University, Germany.

[Trouvain&Schröder 04] Trouvain, J., & Schröder, M. (2004). How (not) to add laughter to synthetic speech. Proc. Workshop on Affective Dialogue Systems (pp. 229-232). Kloster Irsee, Germany.

[Turk etal 05] Turk, O., Schröder, M., Bozkurt, B. & Arslan, L. (2005). Voice Quality Interpolation for Emotional Text-to-Speech Synthesis. Proc. Interspeech 2005, Lisboa, Portugal, pp. 797-800.

[Vincent etal 05] D. Vincent, O. Rosec & T. Chonavel, "Estimation of LF glottal source parameters based on an ARX model", Interspeech'2005, pp. 333-336, Lisboa, Portugal.Follow-up in progress

4.5.5 Follow-up in Progress

Behaviour

In collaboration with CNRS-Limsi, UP8 aims to work with the masked behaviours observed in the EmoTV corpus. Indeed, in one the video of the corpus, a lady masks her

disappointment by a tense smile. This could be modelled by blending the smile of the faked happiness and the tenseness of the felt disappointment.

Evaluation of the GretaMusic is being set-up by UP8 and KTH. Perceptual tests are being designed to understand how subjects of the experiment will perceive the displayed expressions of the ECA based on the music. Mix-match cases will be included in the tests.

In collaboration with OFAI, MIRALab will investigate the integration of their ECA animation engine with the other mark-up languages, such as RRL. Representative example outputs will be constructed in order to test various aspects of the control of the virtual character.

Speech

There will be further improvements in T-S's EmoFilt-tool [Burkhardt05], which will. focus on the simulation of a larger set of discrete emotions, a more subtle performance, and experimenting with the blending of emotions.

FT's future work will be directed to characterisation of the residual signal in their ARX-model [Vincent etal 05]. More precisely they plan to have a better decomposition of the glottal source signal into a deterministic part and a stochastic part.

At DFKI an extension of the spectral interpolation algorithm from [Turk etal 05] is planned. Instead of off-line interpolation for the creation of diphone databases, they will attempt to use on-line interpolation, requiring less data (because no interpolated databases need to be created) and allowing for more flexibility (because the interpolation ratio can be selected dynamically). DFKI in addition will also interact with the speech group in WP4 to evaluate the usability of their feature extraction tools for unit selection voice database preparation. A first test with a small database will be carried out.

4.6 Achievement 6: Emotion Annotation and Representation Language

4.6.1 Participants

DFKI: Marc Schröder

OFAI: Hannes Pirker, Brigitte Krenn

University of Paris8: Myriam Lamolle, Catherine Pelachaud

University of Bari: Nadja De Carolis

University of Twente: Dennis Reidsma

4.6.2 Use-cases, Requirements, and Specification

Starting from the question of potential use cases for a standardised Emotion Annotation and Representation Language (EARL), we have compiled a list of requirements that such a language needs to fulfil in order to be useful. We have then proceeded to the formulation of an XML dialect addressing these specifications, and have devised a family of XML Schemas defining the syntax of correct EARL documents.

4.6.3 Publications

Schröder, M., & Breuer, S. (2004). XML representation languages as a way of interconnecting TTS modules. Proceedings of the 8th International Conference on Spoken Language Processing. Jeju, Korea.

4.6.4 Follow-up in progress

Draft documents defining requirements, informal and formal specification, and use cases, continue to be worked on in the restricted section of the HUMAINE portal. After a maturing phase consisting of discussion with other WPs, these documents will be made available to the public as Deliverable D6e, in month 29. We will also investigate means for promoting the resulting language within and beyond HUMAINE.

4.6.5 Links to other WPs

The emotion markup language has links to most other workpackages. One type of link is related to requirements motivated by use cases; these include the manual annotation of databases (link to WP5 and WP3), the automatic recognition of emotion (link to WP4), and the generation of emotional behaviour in ECAs (WP6 itself). On another level, WP3 and WP5 recommend useful emotion representations that should be encoded in the language. In the future, these recommendations will be extended to cover, in particular, the issue of mappings between emotion representations.

5 Conclusion

5.1 Obstacles encountered or foreseen

The integration of such a voluminous and varied body of work from different Humaine members was a foreseen obstacle. We have overcome this obstacle through the enumeration of specific ECA capabilities that advance the state-of-the-art, clustering the work of different partners into similar spheres that are crucial to emotion agent research in general, and in functional terms, we have specified their positioning in a theoretical generic agent architecture and are working on representation languages to advance the unification of the inputs and outputs between components (see previous deliverable). At workgroup granularity, cross-links have been established with other Humaine workpackages of relevance in order to ensure project-wide integration.

5.2 Relation to the state of the art

We have taken care to enumerate a number of key areas in ECA research for focusing our efforts in the Humaine project. These *capabilities* make it especially clear for the reader to view how our work is contributing to the field and making broad advances beyond the state-of-the-art in the perceptive, cognitive, active, and evaluative aspects of emotion-oriented agent research. The work proposed in many of these capabilities has hardly been addressed in emotion agent literature and in some cases, work done within the Humaine context represents the first steps taken in totally new directions (1a. Emotion-related attention shifts, for example).
Note: any other cases that we can list where sub-capabilities are completely new for conversational agents? Do you think we should go through each capability here and say how it relates/improves on the state-of-the-art? This might be hard…

5.3 Evidence of esteem

A special issue of the International Journal of Humanoid Robotics on "Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids" edited by Catherine Pelachaud and Lola Cañamero will be published in 2006. Several workshops hosted by main conferences have been organized by Humaine partners (AAMAS 2004 on "Balanced Perception and Actions in ECAs", AAMAS 2005 on "Creating bonds with ECAs", ABUSE 2005, AISB 2005 on "Social Presence Cues for Virtual Humanoids", LREC 2004 on "Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces", HCI 2005 on "Human-animated characters interaction"). These workshops have gathered many international researchers. Participants of WP6 have also been invited to several seminars to present their work related to Humaine.

6 References

[Allwood92] Allwood J, Nivre J, Ahlsén E, "On the semantics and pragmatics of linguistic feedback". In Journal of Semantics. 1992. Vol. 9. No. 1, pp. 1-26

[Allwood&Cerrato03] Allwood J., Cerrato L., "A study of gestural feedback expressions", First Nordic Symposium on Multimodal Communication, Paggio P. Jokinen K. Jönsson A. (eds), Copenaghen, 23-24 September 2003, pp.7-22

[Brown&Levinson87] P. Brown and S. C. Levinson. Politeness --- Some universals in language usage. Cambridge University Press, 1987.

[Brunner79] Lawrence J.Brunner. Smiles can be back channels. Journal of Personality and Social Psychology, 35 (5): 728-734, 1979.

[Burkhardt01] F. Burkhardt, Simulation emotionaler Sprechweise mit Sprachsynthesever-fahren, Dissertation an der TU-Berlin, Shaker Verlag 2001.

[Burkhardt05] Burkhardt F., Emofilt: the Simulation of Emotional Speech by Prosody-Transformation, Proc. of Interspeech 2005, Lisboa, Portugal.

[Cassell&Bickmore00] Cassell, J., T. Bickmore, et al. (2000). Human conversation as a system framework: Designing embodied conversational agents. Embodied Conversational Agents. J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Boston, MIT Press: 29-63.

[Castelfranchi&Parisi80] Castelfranchi, C. and Parisi, D. Linguaggio, concoscenze e scopi Bologna: Il Mulino, 1980.

[Cathcart03] Cathcart N., Carletta J. and Klein E., "A Shallow Model of Backchannel Continuers in Spoken Dialog". EACL 2003: 51-58

[Cavalluzzi.etal04] A. Cavalluzzi, V. Carofiglio and F. de Rosis, Affective Advice Giving Dialogs, June 2004.

[Cerrato03] Cerrato L. and Skhiri M., "Analysis and measurement of communicative gestures in human dialogues". Proc. of AVSP 2003, St. Jorioz, France, pp. 251-256

[Clark96] Clark, H. Using Language. Cmapridge University press, 1996.

[Cowie03] Cowie, R. and R. Cornelius. Describing the emotional states that are expressed in speech. Speech Communication 40, 5-32, 2003.

[DeMeijer89] DeMeijer, M. "The contribution of general features of body. 1989.

[Dennett87] D. C. Dennett, The Intentional Stance, 1987.

[DePaulo.etal96] B. M. DePaulo, D. A. Kashy, S. E. Kirkendol, M. M. Wyer and J. A. Epstein. Lying in Everyday Live. Journal of Personality and Social Psychology, 70/5, 979--995, 1996.

[Dittman68] Allen.T.Dittman and L.G.Llewellyn. Relationship between vocalizations and head nods as listener responses. Journal of Personality and Social Psychology 9 (1): 79-84, 1968.

[Dias05] Dias, J.: FearNot!: Creating Emotional Autonomous Synthetic Characters for Empathic Interactions. Universidade T´ecnica de Lisboa, Instituto Superior T´ecnico, Lisboa, MSc Thesis, 2005

[Dias etal 05] Dias J., Paiva A.: Feeling and Reasoning: a Computational Model for Emotional Agents, EPIA Affective Computing Workshop, 2005.

[Douglas-Cowie.etal03] Douglas-Cowie, E., Campbell, N., Cowie, R. and Roach, P. "Emotional speech; Towards a new generation of databases." Speech Communication(40). 2003.

[Dutoit96] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. van der Vrecken, "The MBROLA project: Towards a set of high quality speech synthesisers free of use for non commercial purposes," in Proc. 4th ICSLP, Philadelphia, USA, 1996.

[Egges etal 04a] Egges A., Kshirsagar S., Magnenat-Thalmann N. Generic Personality and Emotion Simulation for Conversational Agents. Computer Animation and Virtual Worlds. 15(1): pp. 1-13, January 2004

[Egges etal 04b] Egges A., Molet T., Magnenat-Thalmann N. Personalised Real-time Idle Motion Synthesis (Seoul, Korea, 2004). Pacific Graphics 2004.

[Ekman&Friesen75] P. Ekman P. and W.V. Friesen W.V., Unmasking the face. A guide to recognizing emotions from facial clues, Prentice-Hall Inc., Englewood Cliffs, N.J., 1975

[Ekman92] Ekman P., Telling Lies, Norton and Company. 1992.

[Ekman99] Ekman P., Basic emotions. Handbook of Cognition & Emotion. T. Dalgleish and M. J. Power. New York, John Wiley: 301–320. 1999.

[Ekman03] Ekman P., The Face Revealed, London: Weidenfeld & Nicolson. 2003.

[Gallaher92] Gallaher P., Individual differences in nonverbal behaviour: Dimensions of style. Journal of Personality and Social Psychology, 1992(63) 133?145.

[Garcia-Rojas.etal06] A. Garcia-Rojas, F.Vexo and D.Thalman, Semantic Model for Virtual Human Reflex Animation. Submitted to IJHR 2006

[Gardner98] Gardner, R. (1998). Between speaking and listening: The vocalisation of understandings. Applied Linguistics, 19, 204–224.

[Goffman63] Goffman, E., Behaviour in public places: notes on the social order of gatherings, The Free Press, New York, 1963.

[Goodwin81] Goodwin, C. (1981) Conversational Organization: Interaction between Speakers and Hearers. New York: Academic Press.

[Hartmann etal 05] B. Hartmann, M. Mancini, and C. Pelachaud, Implementing Expressive Gesture Synthesis for Embodied Conversational Agents, in Gesture Workshop, Vannes, May 2005.

[Heylen05a] D. Heylen, Challenges Ahead. Head Movements and other social acts in conversation, AISB 2005 - Social Presence Cues Symposium.

[Heylen05b] Dirk Heylen, A Closer Look at Gaze, AAMAS Workshop on Creating Bonds 2005.

[Kendon90] Kendon, A., Conducting interaction: patterns of behaviour in focused encounters, Cambridge University Press, New York, 1990.

[Kipp04] Kipp, M. Gesture Generation by Imitation. From Human Behaviour to Computer Character Animation. Florida, Boca Raton, Dissertation.com. 1581122551. 2004.

[Kshirsagar etal 00] Kshirsagar S. Garchery S., Magnenat-Thalmann N.: Feature point based mesh deformation applied to MPEG-4 facial animation. In Proceedings of Deform2000 (Geneva, Switzerland, November 2000), pp. 23–34.

[Maatman05] R. M. Maatman, Jonathan Gratch and Stacy Marsella, "Natural Behaviour of a Listening Agent, " in 5th International Conference on Interactive Virtual Agents, Kos, Greece, 2005

[Martin etal 05] Martin J.-C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M. and Pelachaud, C. Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs. 5th International Working Conference On Intelligent Virtual Agents(IVA'2005), Kos, Greece, September 12-14 Springer. pp 405-417. 2005

[Norling& Sonenberg04] E. Norling and L. Sonenberg, Creating interactive characters with BDI agents, 2004.

[Peters05] Peters, C., Direction of Attention Perception for Conversation Initiation in Virtual Environments, IVA 2005, pp. 215-228, 2005.

[Peters.etal05] C. Peters, C. Pelachaud, M. Mancini, E. Bevacqua, Engagement Capabilities for ECAs, workshop "Creating bonds with ECAs", Fourth International Joint Conf. on Autonomous Agents & Multi-Agent Systems, Utretch, July 2005.

[Poggi81] Isabella Poggi, C.Castelfranchi and D.Parisi. Answers, Replies and Reactions. In H.Parret, M.Sbisà & J.Verschueren (Eds.) Possibilities and Limitations of Pragmatics. Asterdam: John Benjamins, pp.569-591, 1981.

[Poggi02] Poggi I.: "Towards the Alphabet and the Lexicon of Gesture, Gaze and Touch". In Multimodality of Human Communication. Theories, problems and applications. Virtual Symposium edited by P.Bouissac. http://www.semioticon.com/virtuals/index.html, 2001-2002.

[Poggi04] Poggi I. "Towards a Lexicography and Phonology of Nonverbal Communication Systems". In: L.Payratò, N.Alturo & M.Payà (Eds.) Les fronteres del llenguatge. Promociones y Publicaciones Universitarias, Barcelona 2004, pp. 247-279.

[Rehm&André05c] Rehm, M., and André, E., Informing the Design of Embodied Conversational Agents by Analyzing Multimodal Politeness Behaviours in Human-Human Communication. In: AISB Symposium for Conversational Informatics, 2005.

[Rehm&André05a] Rehm, M., and André, E., Where do they look? Gaze Behaviours of Multiple Users Interacting with an ECA, Intelligent Virtual Agents (IVA), Springer, 241—252. 2005.

[Rehm&André05b] Rehm, M., and André, E., Catch me if you can – Exploring Lying Agents in Social Settings, Proceedings of AAMAS 2005, 937—944. 2005.

[Roger&Bull] Roger, D. and Bull, P. (eds.) Conversation : an interdisciplinary perspective. Clevedon: multilingual matters, 1989.

[Romano elal 05] Romano, D.M., Sheppard, G., Hall, J., Miller, A., Ma, Z. (2005). BASIC: A Believable, Adaptable Socially Intelligent Character for Social Presence. PRESENCE 2005, The 8th Annual International Workshop on Presence, 21-22 September 2005, University College London, London, UK.

[Scherer98] Scherer K.R., Analyzing Emotion Blends, in /10th Conference of the International Society for Research on Emotions/. 1998, Würzburg, Germany: Fischer, A., pp. 142-148.

[Schegloff82] Emanuel A. Schegloff. Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In: D.Tannen (Ed.), Analyzing Discourse: Text and Talk. (Georgetown University Round Table on Languages and Linguistics). Georgetown University Press, Washington, DC, pp. 71-93, 1982.

[Scherer84] Scherer K.R., "On the nature and function of emotion: a component process approach", in K.R. Scherer & P. Ekman (Eds.), Approaches to emotion (pp.293-317). Hillsdale, NJ: Erlbaum. 1984

[Schröder&Grice03] Schröder, M., Grice, M., Expressing vocal effort in concatenative synthesis. Proceedings of the 15th International Conference of Phonetic. Barcelona, Spain. 2003.

[Svennevig99] Svennevig, J., Getting Acquainted in Conversation. John Benjamins. 1999.

[Thorisson94] Thorisson, K., "Face-to-Face Communication with Computer Agents". In Working Notes, AAAI Spring Symposium on Believable Agents, pp. 86--90, 1994.

[Ting-Toomey99] S. Ting-Toomey. Communicating Across Cultures. The Guilford Press, 1999.

[Turk etal 05] Turk, O., Schröder, M., Bozkurt, B. & Arslan, L. (2005). Voice Quality Interpolation for Emotional Text-to-Speech Synthesis. Proc. Interspeech 2005, Lisboa, Portugal, pp. 797-800.

[**Vincent etal 05**] D. Vincent, O. Rosec & T. Chonavel, "Estimation of LF glottal source parameters based on an ARX model", Interspeech'2005, pp. 333-336, Lisboa, Portugal.

[Wallbott98] Wallbott H.G., Bodily expression of emotion/./ European Journal of Social Psychology, 1998. *28* 879-896

[Wallbott&Scherer86] H.G. Wallbott and K.R. Scherer, Cues and Channels in Emotion Recognition. Journal of Personality and Social Psychology, 1986. *51*(4) 690-699.

[Wallis.etal01] P. Wallis, H. Mitchard, D. O'Dea, and J. Das. Dialogue Modelling for a Conversational Agent, 2001.

[Wallis04] P. Wallis, Mixed Initiative Dialog using Goal Tagged Activities, 2004.

[Ward&Tsukahara00] N. Ward and W. Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics, 23:1177-1207.

[Yngve70] Victor H. Yngve. On getting a word in edgewise. In: Papers from the Sixth Regional Meeting, Chicago Linguistic Society, pp. 567-578. Chicago Linguistic Society, Chicago, IL, 1970.

[Karpuzis etal] K. Karpouzis., Sarris, M. Strintzis, (eds.), 3D Modeling and Animation: Synthesis and Analysis Techniques, pp. 175-200, Idea Group Publ.