

INTEGRATED MODEL OF EXPRESSIVE AND ATTENTIVE CAPABILITIES

Conveying Affectiveness in Leading-edge
Living Adaptive Systems

CALLAS

Project IST-34800

Deliverable D133 WP1.3

Programme Name: IST
Project Number: 34800
Project Title: CALLAS
Partners: Coordinator: ENG (IT)
 Contractors:
 VTT Electronics, BBC, Studio Azzurro, XIM,
 Digital Video, Humanware, Nexture, University
 of Augsburg, ICCS/NTUA, University of Mons,
 University of Teesside, Helsinki University of
 Technology, Université Paris 8, Scuola Normale
 Superiore di Pisa, University of Reading,
 Fondazione Teatro Massimo, HITLaboratory
 New Zealand

Document Number: callas.D133.PAR8.WP1.3.V1.0
Work-Package: WP1.3
Deliverable Type: Description of prototype
Contractual Date of Delivery: 31 October 2008
Actual Date of Delivery: 31 October 2008
Title of Document: Integrated model of Expressive and Attentive
 Capabilities
Author(s): Nikolaus Bee, Sylwia Julia Hyniewska, Maurizio
 Mancini, Radosław Niewiadomski, Catherine
 Pelachaud, Chistopher Peters, Jerome Urbain

Approval of this report

Summary of this report: Description of CALLAS components of WP13

History:

Keyword List:

Availability This report is: public

Table of Contents

EXECUTIVE SUMMARY	1
1. INTRODUCTION: OBJECTIVES OF WP 1.3	2
2. ARCHITECTURE OF AFFECTIVE EMBODIED CONVERSATIONAL AGENT	3
2.1 BACKGROUND	3
2.2 FML-APML LANGUAGE	5
2.2.1 <i>BML: Behavior Markup Language</i>	5
2.2.2 <i>FML-APML</i>	6
2.3 GRETA ARCHITECTURE	7
2.3.1 <i>Listener Intent Planner</i>	8
2.3.2 <i>Behavior Planner</i>	8
2.3.3 <i>Behavior Realizer</i>	9
2.3.4 <i>FAP-BAP Player</i>	9
2.3.5 <i>Synchronization</i>	9
3. EXPRESSIVE AGENT	10
3.1 BACKGROUND	10
3.1.1 <i>Expressions of emotions blend</i>	10
3.1.2 <i>Multimodal expression of emotion</i>	11
3.2 STATE OF ART ON EXPRESSIVE AGENTS	12
3.3 ANNOTATION	14
3.3.1 <i>Observation of emotional behavior</i>	14
3.3.2 <i>Annotations for CALLAS</i>	15
3.4 EXPRESSIONS OF GRETA	16
3.4.1 <i>Multimodal expressions of emotions</i>	16
3.4.2 <i>Model of complex facial expressions</i>	21
3.4.3 <i>Facial expressions of different intensities</i>	25
3.4.4 <i>Full-body expression of communicative intensions</i>	27
3.4.5 <i>Laughter production</i>	31
4. ATTENTIVE AGENT	33
4.1 THE GAZE AWARENESS MODULE	33
4.1.1 <i>State of art</i>	33
4.1.2 <i>Modeling attention and interest</i>	33
4.1.3 <i>Implementation</i>	35
4.2 PROOF OF CONCEPT: A COMPARISON BETWEEN HUMAN-GRETA AND HUMAN-ROBOT INTERACTIONS	35
4.2.1 <i>Data acquisition</i>	35
4.2.2 <i>System architecture</i>	36
4.2.3 <i>Project outcomes</i>	37
5. AFFECTIVE AND COGNITIVE THEORY OF MIND FOR ATTENTIVE AGENTS	39
5.1 BACKGROUND	39
5.2 STATE OF ART	39
5.3 THEORY OF MIND	40
5.4 APPROACH FOR AFFECTIVE INTERACTION	40
5.5 MIMICRY VS. ROLE-TAKING	41
5.5.1 <i>Model for mimicry</i>	41
5.5.2 <i>Model for role-taking</i>	42
REFERENCES	43

Executive Summary

In this document we describe works we have undertaken to create an expressive and attentive embodied conversational agent. In this deliverable we present our model and algorithm. This work is being used in the Showcase Interactive TV. Its integration in the showcase is described in the Deliverable D341. In the deliverable D212 we motivate our choice of emotion representation for our computational models.

We first describe the main architecture of our agent technology, Greta. It is compliant with the SAIBA framework and works in real-time. We give a description of each of the main modules of agent architecture.

We then turn our attention to the algorithm we have developed to enhance the expressivity of our agent. We follow two approaches. In one approach we base our model on theoretical models reported in the literature. In a second approach we use an observational approach. We annotated data from a video corpus and base our model from these annotations. We report works done following both of these approaches in Sections 3.1, 3.2, and 3.3.

To increase the expressiveness of our agent we have elaborated algorithms to:

- express emotion dynamically through various modalities,
- compute the facial expression corresponding to complex emotion,
- modulate facial expression depending on the intensity of an emotion,
- characterize how agents differ in expressing emotional state and communicative functions.

For each of these expressive qualities, an algorithm has been implemented. Representation languages to describe the link between an emotional state and its dynamic expression across modalities have been designed.

Noticing that the agent expressivity would be largely improved in certain situation by giving it the ability to laugh, we are also developing methods to generate human-like laughs. The first steps accomplished towards natural laughter production are described in Section 3.

We also present, in Section 4, how the agent is endowed with perceptive capabilities and how it uses them to be attentive. This model uses two metrics, the attention and interest ones. This model has been used in a shared attention scenario between a single user and an ECA. The section is concluded by a description of the integration of our agent technology within an eINTERFACE project. This project aimed to generate in real-time acoustic and visual backchannels to be displayed by a virtual agent (Greta) and a robot (AIBO). A large and freely available corpus of storytelling behaviors (and backchannels) was also recorded during this project.

The last section of this deliverable deals with a model of an empathic listening agent that responds to the user's emotive and attentive state. Two response types have been considered: mimicry (linked to affective empathy) and role-taking (linked to cognitive empathy).

1. Introduction: objectives of WP 1.3

The objective of the workpackage WP1.3 is to model an emotional multimodal ECA. ECAs are autonomous agents with a human-like appearance and communicative skills. They have shown their potential to allow users to interact with the machine in a natural and intuitive manner through human communicative means. We believed the following capabilities are necessary to build a truly emotional multimodal ECAs:

- **Emotional communication:** Communication is done via verbal and nonverbal means. Gesture, facial expression, gaze, body, prosody, speech are at work to convey meaning. They provide information on the emotional state of the emitter, her mood, personality, etc. Communication is not simply significant by which verbal and nonverbal signals are displayed but also by how they are executed. The expressivity of behaviors, the choice of the words are an integrant part of the communication process. Moreover emitted signals are highly synchronized with each other. ECAs ought to reflect these human qualities. They ought to be able to display expressive synchronized multimodal behaviors.
- **Emotional expression in a social context:** When conversing, a person may decide to express an emotion different from the one she actually felt because she has to follow some socio-cultural norms or she is pursuing some others of her goals. Ekman (1975) refers to the former as display rules. Blend of emotions may be due by rapid sequences, superposition of two or more emotions or by masking one from another one. We need to go beyond the facial expression of basic emotions and to take into account blending of emotions.
- **Surrounding awareness:** From the age of 9 months, human infants engage in a range of joint visual attention behaviors, the most obvious of which are gaze monitoring and following and the "protodeclarative" pointing gesture; In the latter case, the infant alternates his or her gaze between the adult's eyes and the object at which they are both attending (Bates et al., 1979). Adults also engage in such behavior. The ability to detect and engage in shared attention behaviors with a user is therefore of importance for natural interaction between users and ECAs that are situated in and must make reference to the real environment.

Thus, one of the crucial issues in the creation of ECAs is to enhance them with social intelligence and communicative abilities to give them the capacity to interact with the user in natural way and to display complex and subtle expressions. In this deliverable we describe models we have developed to create an emotional multimodal ECA endowed with expressive and attentive capabilities:

- Model of emotional multimodal nonverbal behaviors,
- Model of facial expression of complex emotions,
- Model of surrounding awareness.

2. Architecture of affective embodied conversational agent

In this section we describe the architecture of an affective embodied conversational agent (ECA) called Greta. This agent architecture is of general purpose use and modular that works in real-time. The 3D agent is able to communicate using verbal and nonverbal channels like gaze, head and torso movements, facial expressions, and gestures. It follows the SAIBA framework that defines functionalities and communication protocols for ECA systems and the MPEG4 standard of animation. The agent system is optimized to be used in interactive real-time applications. In this deliverable we present the technical details of our system as well as several applications that use it.

2.1 Background

In this section we present some existing ECA systems. We describe also the standardization efforts made so far by the ECA community. We conclude this section by presenting the first architectures that were developed according to the SAIBA platform.

The humanoid Rea was one of the first ECAs. It was developed by Cassell and Bickmore (Cassell and Bickmore, 1999) and was designed to work as a virtual real estate agent. It is able to understand the user's multimodal behaviors and to respond with appropriate speech and intonation accompanied by various types of nonverbal behaviors. It is a three dimensional full-body animated character that is displayed on a large projection screen, in front of which the user stands. Rea is a real-time system involving user's speech recognition and movement detection, a dialog manager and behavior planner. Later on, Cassell's team further developed this agent technology. Their aim was to create a tool that could be used by various platforms and in many applications. The BEAT toolkit (Cassell, 2001) is a modular and extensible animation tool working in real-time that selects and schedules nonverbal behaviors of a virtual character. It extracts linguistic and contextual information from the input text, chooses adequate gestures, eye gaze, and other nonverbal behaviors. The BEAT offers a synchronization scheme between all behaviors. For this purpose it uses a set of rules derived from psychological research on nonverbal behaviors. The BEAT can be integrated with various animation systems. It generates as output a set of instructions in a proprietary format that can be then interpreted by an animation system or can be edited by a human animator.

Max (Kopp et al., 2003) is another example of multimodal interactive agent that works in real-time. It is a three-dimensional human size embodied agent that inhabits a virtual environment. Max allows multimodal bidirectional communication and can be integrated with various input devices and in different application settings. Among others it was used to help the user in a construction task (Kopp et al., 2003); in another application Max can dialog with users on various topics such as museum description. In a highly interactive situation Max is able to communicate with the human user in a face-to-face manner using prosodic speech, various gestures, gaze, and facial expressions. The user communicates with Max using natural language and gestures (analyzed by data glove). Max is able to have deliberate and reactive conversation with the user. It displays also facial expressions of emotions and is able to generate feedback-driven reactive behaviors like gaze tracking of the current interlocutor. Finally it schedules and executes all verbal and nonverbal behaviors in synchrony.

The architectures presented above were advanced and powerful but they still used some proprietary protocols and architectures. Huang et al. proposed GECA (Huang et al., 2006) - a generic framework for building ECAs. It is programming language independent, real-time, distributed and general purpose architecture. It can be used to create ECAs able to capture and interpret various inputs and to generate synchronized verbal and nonverbal output. This

framework is composed of three layers: the communicating platform, XML-based communicating protocol, and an API for the modules creators. The framework uses a blackboard that integrates different ECA components like speech recognition module or motion capture module. The platform is able to exchange both sensor data streams and command messages between all the components. Huang et al. (Huang et al., 2006) proposed also a XML-based high-level communication protocol to be used between the components. The messages are ordered in a hierarchical structure (for example, *input.speech.text*, or *output.body.gesture*) and each message type has a specified set of elements and attributes, e.g. *intensity* or *duration*. The implementation of the framework written in Java, uses OpenAir communication protocol and generates animation in MPEG-4 standard. It was used to build Multicultural Tour Guide agent (Cerekovic et al., 2007). The application uses 11 components, among which speech recognition and motion capture components for input and animation character for output.

SAIBA (Kopp et al., 2007; Vilhjálmsón et al. 2007) is an international research initiative whose main aim is to define a standard framework for the generation of a virtual agent behavior. It defines a number of levels of abstraction (see Figure 1), from the computation of the agent's communicative intention to behavior planning and realization.

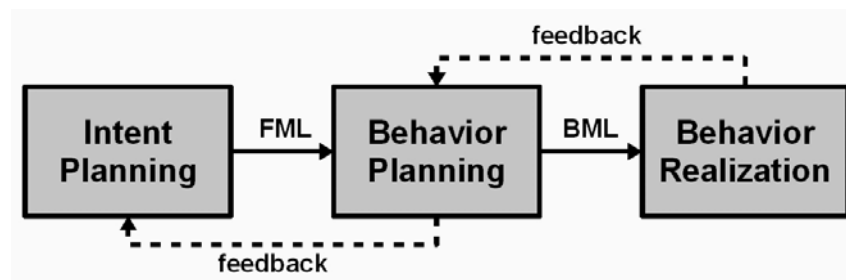


Figure 1. SAIBA framework (Kopp et al., 2007).

The Intent Planning module decides the agent's current goals, emotional state and beliefs, and encodes them into the Function Markup Language (FML) (Heylen et al., 2008). To convey the agent's communicative intentions, the Behavior Planning module schedules a number of communicative signals (e.g., speech, facial expressions, and gestures) which are encoded with the Behavior Markup Language (BML). It specifies the verbal and nonverbal behaviors of ECAs (Vilhjálmsón et al., 2007). Finally the task of the third element of the SAIBA framework, the Behavior Realization, is to realize the behaviors scheduled by the Behavior Planning. It receives input in the BML format and it generates the animation.

There exists several implementations like SmartBody (Thiebaux et al., 2008) and BMLRealizer (Arnason and Þorsteinsson, 2008) that are SAIBA compatible. SmartBody (Thiebaux et al., 2008) is a modular, distributed open-source framework for animating ECAs in real time. It is based on the notion of animation controllers. The controllers are organized in a hierarchical structure. In SmartBody two types of controllers are distinguished. Ordinary controllers manage the separate channels, e.g. head pose or gaze. Then the meta-controllers manipulate the behaviors of subordinate controllers allowing the synchronization of the different modalities to generate consistent output from the BML code. SmartBody corresponds to the Behavior Realization module of the SAIBA architecture. It takes as input the BML code (including speech timing data and the world status updates) and composes multiple behaviors and generates character animation synchronized with audio. The verbal content is generated by an external TTS system. BML used within SmartBody is a subset of the standard, but it offers extensions that introduce interruptions and predefined animations.

SmartBody can be used with the Nonverbal Behavior Generator (Lee et al., 2006) that corresponds to the Behavior Planning in the SAIBA framework. It is a rule-based module that generates BML annotations for nonverbal behaviors from the communicative intent and speech text. On the other hand, SmartBody can be used with different characters, skeletons

and even different rendering engines. It was used in many applications for example in Virtual Patient (Kenny et al., 2008) to realize a virtual teenager with psychological disorders.

BMLRealizer (Arnason and Porsteinsson, 2008) created in the CADIA lab is another implementation of the Behavior Realization layer of the SAIBA framework. It is an open source animation toolkit for visualizing virtual characters in 3D environment that is partially based on the SmartBody framework. As input it also uses BML; the output is generated with the use of the Panda3D rendering engine.

2.2 FML-APML language

In this section we describe two languages we implemented in our system: BML and FML-APML. Right now, in the SAIBA initiative, the BML language is more specified than the FML language. Even though work on BML is not completely finalized our system implements the current existing version. We propose several extensions of BML that allows one to exploit better the capabilities of our agent. While the final specification of BML is on its way, the second language called FML is not yet defined. FML is at its infancy. There has been a workshop at last AAMAS 2008 for which we participated in its organization that dealt with describing the scope of information FML should encompass. No work on defining FML tags have started yet. We propose a temporary solution that we called FML-APML - the language that is based on the previous language APML but has the features of the future FML.

2.2.1 BML: Behavior Markup Language

BML language is not yet a standard, however researchers agreed on a “common” BML syntax specification to allow one to exchange BML files and engines between different systems, as described in (Kopp et al., 2007; Vilhjálmsón et al., 2007). The BML language allows us to specify the nonverbal signals that can be expressed through the agent communication modalities. Each BML top-level tag corresponds to a behavior the agent is to produce on a given modality: head, torso, face, gaze, body, legs, gesture, speech, lips. In the current version for each modality one signal can be chosen from a short fixed list. Each signal has defined a duration and a starting time. This temporal information can be absolute (in seconds) or relative, in relation to the other verbal or nonverbal signals.

The BML language version we have implemented in our agent contains some extensions which allow us to define labels to use a larger set of signals which can be produced by the agent and to specify the expressivity of each signal.

Signal label. In the common BML syntax it is possible to specify just a small set of signals for the agent. For example we can specify only 4 mouth shapes: flat, smile, laugh and pucker. This is a limitation, since some agents capable of performing complex actions could not be fully exploited. A parameter called *reference* was introduced to specify the name of the facial signal that the Behavior Realizer has to produce. Thus, in our version of BML we have two types of information about a signal: the *type* attribute, which is mandatory and refers to the small set of signals defined in the BML common version, and the *reference* attribute, which is used by our agent to perform a nonverbal behavior from a larger set of signals. In our system, the Behavior Realizer always prefers to perform the signal specified by the *reference* attribute, if present. But still, we can give the BML code computed by our system to any other Behavior Realizer, as it also contains the first parameter, the *type* attribute, which is mandatory and can be interpreted by other Behavior Realizers.

Expressivity parameters. Our agent can dynamically modulate multimodal signals using a small set of high level parameters, that we call *expressivity parameters* (see Section 3.4.4). They influence the quality of movement: for example, the gesture of raising a hand can be performed quickly or slowly, with more or less energy, reaching a point further or nearer in space, and so on. Expressivity parameters are not included in the common BML syntax but can be interpreted by our Behavior Realizer. Thus, in the implementation of BML in our system, we can specify not only *which* signals the agent has to perform but also *how*. For

example, a beat gesture (with a vertical up-down movement of the hand/arm) can be performed in different ways: quickly or slowly, smoothly or jerkily, etc.

2.2.2 FML-APML

FML encodes communicative and emotional functions the agent aims to transmit. Our version of this language, FML-APML is an XML-based markup language for representing the agent's communicative intention and the text to be uttered by the agent. The communicative intentions of the agent correspond to what the agent aims to communicate to the user: e.g., its emotional states, beliefs and goals. It originates from the APML language (de Carolis et al., 2004) which uses Isabella Poggi's theory of communicative acts (Poggi, 2007). FML-APML uses a similar syntax as BML one. It has a flat structure and allows defining explicit duration for each communicative intention.

Each tag represents one communicative intention; different communicative intentions can overlap in time. We consider the following tags (taken from (Poggi, 2007)):

- **certainty**: is used to specify the degree of certainty the agent intends to express;
- **performative**: represents the agent's performative e.g. suggest, approve, or disagree;
- **theme/rheme**: represents the topic/comment of conversation; that is, respectively, the part of the discourse which is already known or new in the participants' conversation;
- **belief-relation**: corresponds to the metadiscursive goal, i.e. the goal of stating the relationship between different parts of the discourse;
- **turntaking**: models the exchange of speaker turns;
- **emotion**: describes the emotional state of the agent. We can define simple emotions using emotional labels (e.g. anger or sadness) but also complex emotional states like masking (i.e. the agent has a certain emotion but it hides it by showing another, fake, one) or superposition of two emotions (see Section 3.4.2 for details);
- **emphasis**: is used to emphasize the agent's verbal or nonverbal message;
- **backchannel**: Through backchannels the listener provides information about its communicative intentions, in particular about its will and ability to continue, perceive, understand the interaction and its attitude towards the speaker's speech (if it believes or not, likes or not, accepts or refuses what is being said) (Allwood et al., 1993);
- **world**: refers to objects of the world.

We can remark that this language allows us to describe the agent's communicative functions when it is either the speaker or the listener.

The attributes of FML-APML tags are:

- **name**: the name of the tag, representing the communicative intention modeled by the tag. For example, the name *performative* represents a performative communicative intention;
- **id**: a unique identifier associated to the tag; it allows one to refer to it in an unambiguous way;
- **type**: this attribute specifies the communicative meaning of the tag. For example, a performative tag has many possible values for the type attribute e.g. suggest, propose, approve, etc. Depending on both the tag name (performative) and type (one of the above values), our Behavior Planning module determines the nonverbal behaviors the agent has to perform;
- **start**: starting time of the tag, in seconds. It can be absolute (time 0 corresponds to the start of the FML-APML message) or relative to another tag. It represents the point in time at which the intention specified by the tag starts to be communicated;

- end: duration of the tag. It can be a numeric value (in seconds) relative to the beginning of the tag or a reference to the beginning or end of another tag (or a mathematical expression involving them). It represents the duration of the communicative intention modeled by the tag;
- importance: a value between 0 and 1 which represents the probability that the communicative intention encoded by the tag is communicated through nonverbal behavior;
- intensity: the emotions can be expressed with different intensities (Ekman, 1975)
The intensity of an emotional state is described by a value from the interval [0..1].

Our agent uses the FML-APML language. FML-APML is based on APML (de Carolis et al.; 2004), allowing the specification of the agent's communicative intentions, emotions and belief states (such as performatives, emotions, or belief-relations). In comparison with APML, the FML-APML language is simpler to use and tag nesting (i.e. tags had to be nested one in another; no partial overlap was allowed) is not required any more. The duration of each communicative intention can be specified explicitly (in seconds) or in relation to a speech act. The other novelty is the possibility to define not only the speaker's intentions but also the listener's ones. Finally, in FML-APML information on the world can be specified to communicate some physical or abstract properties of objects, persons, events.

2.3 Greta architecture

Figure 2 illustrates the architecture of our agent. The Behavior Planner receives as input the agent's communicative intention encoded in FML-APML and generates as output a set of BML signals. These signals are sent to the Behavior Realizer that generates the agent's animation following the MPEG-4 standard. Finally, the animation is played by the Player. Our architecture has also a Listener Intent Planner that belongs to the SAIBA Intent Planner module. Such a planner is able to generate in real-time the agent's behavior while in the role of the listener.

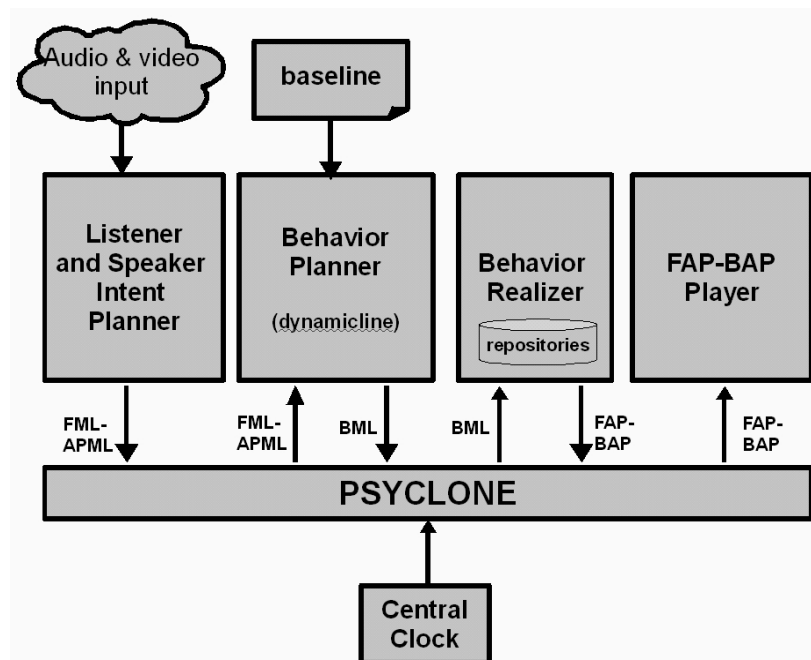


Figure 2. The architecture of Greta agent.

All modules in the architecture are synchronized by the Central Clock and communicate with

each other through a whiteboard. For this purpose we use the Psyclone messaging system (Thórisson et al., 2005) which allows modules and applications to interact together, even if they are running on separate machines connected through TCP/IP. The system has a very low latency time and is suitable for real-time applications.

In the following subsections we describe each module of our agent.

2.3.1 Listener Intent Planner

The Listener Intent Planner module is in charge of the computation of the agent's behaviors while being a listener when conversing with a user. This component encompasses three modules called reactive backchannel, cognitive backchannel, and mimicry.

Research has shown that there is a strong correlation between backchannel signals and the verbal and nonverbal behaviors performed by the speaker (Maatman, 2005; Ward and Tsukahara, 2000). Models have been elaborated that predict when a backchannel signal can be triggered based on a statistical analysis of the speaker's behaviors (Maatman, 2005; Morency, 2005; Ward and Tsukahara, 2000). We use a similar approach and have fixed some probabilistic rules to prompt a backchannel signal when our system recognizes certain speaker's behaviors; for example, a head nod or a variation in the pitch of the user's voice will trigger a backchannel with a certain probability. Probabilities are set based on studies from the literature (Maatman, 2005; Ward, 2000). The reactive backchannel module takes care of this predictive model. On the other hand, the cognitive backchannel module computes when and which backchannel should be displayed using information about the agent's beliefs towards the speaker's speech. We use Allwood's taxonomy of communicative functions of backchannels (Allwood et al., 1993): contact, perception, understanding, and attitudinal reactions.

We have elaborated (Bevacqua et al., 2007; Heylen et al., 2007) a lexicon of backchannels based on perceptive studies. The cognitive module selects which signals to display from the lexicon depending on the agent's reaction towards the speaker's speech. The third module is the mimicry module. When fully engaged in an interaction, mimicry of behaviors between interactants may happen (Lakin et al., 2003). This module determines when and which signals would mimic the agent. So far we are considering solely speaker's head movement in the signals to mimic. A selection algorithm determines which backchannels to display among all the potential signals that are outputted by three modules.

The Listener Intent Planner can be seen as part of a larger module that has the task of computing the agent's intentions when being a listener or a speaker. So far only the module to compute the listener's intentions has been implemented and the Speaker Intent Planner, the module charged with the calculation of the speaker's intentions, is still under construction. Together, these two modules will correspond to the Intent Planner in the SAIBA framework.

2.3.2 Behavior Planner

The Behavior Planner takes as input both the agent's communicative intentions specified by the FML-APML language and some agent's characteristics. The main task of this component is to select, for each communicative intention, the adequate set of behaviors to display. The output of Behavior Planner is described in the BML language. It contains the sequence of behaviors with their timing information to be displayed by our virtual agent.

Our agent is characterized by its general tendency to behave. These characteristics are at the level of behaviors and not at the emotional or personality level, even though both levels are intrinsically correlated. The agent's general behavior tendency is represented by the agent's *baseline*. This last one contains information on the preference the agent has in using its communicative modalities (head, gaze, face, gesture and torso) and on the expressive quality of each of them. Expressivity is defined by a set of parameters that affect the qualities of the agent's behavior (e.g. wide vs. narrow gestures). The system uses the agent's baseline to compute how a given communicative intention is shown. Our system enables to have agents defined with different baselines to communicate accordingly. It allows us to give some

coherency in the agent's behaviors throughout their interaction with users. The details of this algorithm are presented in Section 3.4.4.

2.3.3 Behavior Realizer

This module generates the animation of our agent following the MPEG-4 format (Ostermann, 2002). The input of the module is specified by the BML language. It contains the text to be spoken and/or a set of nonverbal signals to be displayed. Each BML tag is instantiated as a set of key-frames that are then smoothly interpolated. Facial expressions, gaze, gestures, torso movements are described symbolically in repository files. The Behavior Realizer solves also eventual conflicts between the signals that are scheduled to happen on the same modality at the same time. The Behavior Realizer uses repository files of predefined facial expressions, gestures, torso movements and so on. The agent's speech, which is also part of the BML input, is synthesized by an external TTS system. The TTS system provides the list of phonemes and their respective duration. This information is used to compute the lip movements.

When the Behavior Realizer receives no input, the agent does not remain still. It generates some idle movements. Periodically a piece of animation is computed and is sent to the Player. This avoids unnatural "freezing" of the agent. For some modalities, like the head or gaze, the Behavior Realizer manages also signals with "infinite" duration i.e. signals with an a priori unknown ending time. In this way we can force the agent to keep the head turned till a BML command arising from a new communicative intention is generated by the Intent Planner module.

2.3.4 FAP-BAP Player

The FAP-BAP Player receives the animation generated by the Behavior Realizer and plays it in a graphic window. The player is MPEG-4 compliant. The animation is defined by the Facial Animation Parameters (FAPs) and the Body Animation Parameters (BAPs) (Ostermann, 2002). The FAPs define the shape deformation or movements of a set of 68 fundamental points on a synthetic face with respect to their neutral position; the BAPs represent rotations of body parts around specific joints. Facial and body configurations are described through respectively FAP and BAP frames.

Each FAP or BAP frame received by the Player carries also the time of its visualization computed by the Behavior Realizer; such a time is calculated from the Central Clock (see next section). In case the Player receives more than one frame with the same timestamp it displays the latest one it receives.

2.3.5 Synchronization

Each component can send to the whiteboard message types; it can read them from the blackboard when they are published by another component. Each component can also generate output and publish it to the blackboard. For example the Behavior Realizer sends messages of the type *Agent.Data.FAPFrame* for facial animation, *Agent.Data.BAPFrame* for body animation, and *Agent.Data.Wav* for speech. It also receives two types of messages: *Agent.Data.BML* containing BML commands and *Agent.Data.Clock* used for synchronization purpose.

The synchronization of all modules in the distributed environment is ensured by the Central Clock which broadcasts regularly timestamps through Psyclone. All other components are registered in the whiteboard to receive timestamps.

3. Expressive agent

3.1 Background

A growing interest in developing virtual characters expressing emotions has been observed in recent years. This is motivated by an attempt to enhance human-machine interaction. To be able to express emotions, an agent needs to have the nonverbal communication skills and access to data on how to communicate in a way to be understood by humans.

Although humans communicate through several modalities at the same time, it is the face that is a privileged place for the expression and the decoding of emotions, as suggested by interdisciplinary theorists (e.g. Kaiser and Wehrle, 2001; Ekman, 1972), starting with early work by Darwin (1872/1998) and Duchenne (1876/1999).

When working with virtual agents, it is important to keep in mind that the interaction of the humans with virtual characters is similar to the human-to-human interaction (Schilbach et al. 2006; Brave et al., 2005). Therefore, it seems plausible that for the emotional expression synthesis in ECAs it would be appropriate to apply a psychological model of human behavior.

Today emotions can be understood in the psychology domain in two major ways: as a dynamic cognitive evaluation of a situation (componential appraisal theory) or as a discrete and automatic reaction to a situation (discrete emotions theory).

According to the componential appraisal theory, an emotional state is created by the significance given to different elements of an event. According to Scherer, a researcher from the componential approach, an emotion would arise from a series of sequential evaluation checks of the surrounding stimuli (Scherer, 1992). This evaluation is subjective, with respect to the well being of the individual. Thus the mental state is function of the subjective appraisal and not a preprogrammed reaction. Furthermore, each step of this subjective appraisal is linked to a facial response. Those facial movements are defined by Scherer in terms of action units (AU) which represent the position of particular muscles during an expression. The appraisal theory states that it is the cumulation of the AU resulting from the step by step evaluation that creates the final emotional expression. The number of facial emotional expressions is thus very large, as the various elements of the facial expressions (AUs) can co-occur in different patterns.

On the other hand, the discrete emotion theorists believe there are a limited number of fundamental emotions which often are called “*primary*” or “*basic*” emotions (e.g. Izard, 1977; Plutchik, 1980). They claim that emotions are common to different cultures and at least some of them occur also in some other species (Ekman, 1999). Each of these prototypical emotions is characterized by a specific adaptive function, expression (e.g., specific facial behavior), physiological pattern, distinctive conscious experience (a feeling), and instrumental action (Keltner and Buswell, 1997; Manstead et al., 2005).

3.1.1 Expressions of emotions blend

Some discrete emotion theorists like Ekman (2003b) believe that the very wide range of emotional expression can be explained by six basic emotions (i.e. anger, fear, joy, sadness, surprise, disgust) and their blends. Blends denominate the expressions in which more than one emotion is involved. The different types of blends can appear in the form of *superposition* of emotions or *masking* of one emotion by another. Ekman observes that they occur often in everyday life (Ekman and Friesen, 1969) along a *simulation*, which is faking an emotion, and *inhibition* of one emotion.

According to Ekman the complex facial expressions are obtained by the composition of

expressions over different face areas. For instance in the case of a superposition of two emotions, the display is composed of one emotional expression for the upper face area and a different one for the lower face (Ekman, 1975). The boundary between the upper and the lower face is not precisely defined: for certain pairs of emotions (e.g., anger and sadness) the eyes are included in the upper face area, while for other pairs they are not (Ekman, 1975). Ekman described eighteen different expressions of superposition for pairs involving six emotions (Ekman (1975, 2003b)). However, not every possible combination of the upper and the lower faces is plausible e.g. sadness in the superposition with happiness is expressed by the upper face region, and happiness by the lower face. The opposite case does not occur. Researchers have also shown that humans distinguish the expression of a felt emotion from the expression of a fake one (Ekman and Friesen, 1969; Frank et al., 1995; Gosselin et al., 1995). A list of deception cues, i.e. features of expression that are useful in distinguishing between fake and felt expressions, have been proposed (Ekman, 1975; 1985; 2003a). Since, humans are not able to control voluntarily all their facial muscles, the observation of facial movements that are accomplished only with difficulty in conscious expressions can lead to the differentiation between “genuine” and “fake” expressions by the perceiver. Expressions of particular felt emotions may thus be associated with specific facial features like sadness brows (Ekman, 1975) or *orbicularis oculi* activity in the case of joy (Ekman, 2003b).

Not only the reliable features lack in fake expressions, but they also are not easily and fully inhibited in the felt emotions. Moreover, felt and fake expressions can also be distinguished by their variation of symmetry, synchronization, and timing (Ekman, 1975; 1985). Fake expressions are more often asymmetric (Ekman, 2003a), more abrupt (Frank et al., 1995; Ekman and Friesen, 1982) and are often displayed for longer durations than felt ones (Ekman, 2003a).

Summarizing, the discrete emotion approaches provide concrete predictions on several emotional expressions. They have been applied to Greta's platform. The idea of universality of the most common expressions of emotions was particularly sought to enable the generation of “well recognizable” facial displays. What is more, the unitary nature of the expressions was particularly attractive for its simplicity: the different elements (e.g. AU) of each expression are predicted to have a common development, with only one starting and one ending point and a common apex. However easy to categorize in terms of evoked emotions, the expressions based on discrete theory are still oversimplified.

3.1.2 Multimodal expression of emotion

In line with the appraisal theory, which claims that an emotion is a *dynamic episode* that produces a sequence of response patterns on the level of gestures, voice and face (Scherer and Ellgring, 2007), it could be advantageous for Greta's believability to include more temporal variations and multimodality. Although most of the studies concentrate on the face, some studies show that emotions can be also related to body movements (Wallbott, 1998; Pollick et al., 2001). To create a believable multimodal expression for the agent, more information is needed on the sequence of appearance of different components and on the complexity of real life displays.

A way to obtain more such data is by direct observation, whether guided by theory or not. Some observational studies have explored the complexity of emotional expressions in terms of their dynamics and/or multimodality. Thus, Keltner (1995) studied the sequence of facial and gestural movement in embarrassment. She relied on the analysis of their appearance frequencies in audio-visual data. Shiota and colleagues on the other hand, studied three positive emotions: awe, amusement, pride (Shiota, et al., 2003). They showed that the three have expressions that are more than prototypical static facial expressions as described in Ekman's work (1975). They would rather be expressed by a set of possible signals, sometimes with asynchronous onsets, offsets and apices. The expression is not to be seen as categorical and not all elements have to be present in an expression at the same time, for such to be recognized as a display of a particular emotional state.

In the expression of awe (Shiota, et al., 2003), for example, Shiota observes raised inner

eyebrows (AU1), widened eyes (AU 5), an open mouth with a slight drop of the jaw (AU 26+ AU27), a forward jutting of the head and one visible (deep) inhalation. Although the eyebrow movement, the eye widening and mouth opening are clearly present in the great majority of expressions, the other two appear in less than one third of the cases. Shiota has analyzed in a similar way the expression of amusement and pride, providing detailed information on the exact presence of facial movements in term of AU, but also on their frequency of appearance. Shiota differentiated these three positive emotions, going beyond the one well recognized positive expression of happiness. Hence, the Duchenne smile¹, which is often considered the only reliable expression of a positive affect, cannot be used as the only expression of various positive internal states.

Hence, theory and annotation of audiovisual corpora has provided substantial guidance in the creation of synthetic facial expressions for Greta, and was applied for the improvement of the agent's expressivity and believability. The link between expressions and internal states of the ECA is being reinforced.

3.2 State of art on expressive agents

Several models of facial expressions have been proposed to enrich the agent's facial behavior. The existing solutions usually compute new expressions "averaging" the values of the parameters of the expressions of "basic" emotions (Ekman, 1975, Ekman 2003b). The model called Emotion Disc (Ruttkay et al., 2003) uses a bi-linear interpolation between two basic expressions and the neutral one. In the Emotion Disc six expressions are spread evenly around the disc, while the neutral expression is represented by the centre of the circle. The distance from the centre of the circle represents the intensity of expression. The spatial relations are used to establish the expression corresponding to any point of the Emotion Disc. Models of Tsapatsoulis et al. (Tsapatsoulis et al., 2002) and Albrecht et al. (Albrecht et al., 2005) can be used to compute expressions. Both use the expressions of two "neighboring" emotions to compute the facial expressions for non-basic emotions. For this purpose they use different multidimensional spaces, in which emotional labels are placed. In both approaches new expressions are constructed starting from the six Ekman's expressions: anger, disgust, fear, happiness, sadness, and surprise. In more detail, in Tsapatsoulis et al. (Tsapatsoulis et al., 2002) two different approaches are used. First of all, a new expression can be derived from a basic one by "scaling" it. In the second approach a new expression is generated by looking for the spatially closest two basic emotions as defined within the dimensional space proposed by Whissell (Whissell, 1989) and Plutchik (Plutchik, 1980). Then the parameters of these expressions are weighted with their coordinates. Albrecht et al. (Albrecht et al., 2005) proposed an extended approach. The authors use a three dimensional space of emotional states defined by activation, evaluation, and power as proposed in (Cowie et al., 1999). Bui (Bui, 2004) uses a set of fuzzy rules to determine the blending expressions of six basic emotions based on Ekman's findings (Ekman, 1975). A subset of rules is attributed to each pair of emotions. The fuzzy inference determines the degrees of muscles contractions of the final expression as a function of the input emotions intensities. Finally, different types of facial expressions were considered by (Rehm and André, 2005). In a study on deceptive agents, they showed that users were able to differentiate between the agent displaying an expression of felt emotion versus an expression of fake emotion (Rehm and André, 2005)). For this purpose they manually defined facial expressions according to Ekman's description of expressions for fake emotion. These expressions are more asymmetric and miss reliable features.

Ruttkay (Ruttkay, 2001) proposed the system for the differentiation of facial expressions of an ECA. It allows the designer to modify by hand the course of a facial expression animation which is defined par default by a trapezoid attack-hold-delay. The plausibility of the final

¹ According to Duchenne, the genuine smile includes the contraction of the zygomaticus major (the muscle pulling the corners of the lips upwards) with that of the orbicularis oculi (which is mostly perceived as a contraction of the lower eyelid and wrinkles near the eyecorners).

animation is reassured by a set of constraints. In more details the system allows the user, for any single facial signal, to define manually the course of the animation. The user has in disposal an editor of facial displays using which he can deliberately modify the animation curve for any FAP parameter. The system verifies the feasibility of the animation generated in this way. For this purpose it checks the consistency of the new animation with a set of constraints. The constraints are defined on the key-points of the animation and concern facial animation parameters. One can, for example, force the facial expressions to be symmetric (i.e. all FAPs have same values for each key-point) or choose that only in the first and last key-point the expression is symmetric while it doesn't have to be on the other key-points. If the new animation is not consistent with the constraints the system will report it to the user. In that case he can change the animation or modify the constraints.

Xueni Pan et al. (Pan et al., 2007) proposed an approach to display emotions that cannot be expressed by static facial expressions but are expressed by certain sequences of signals (facial expressions and head movements). First of all, certain sequences of signals were extracted from the video-corpus. From this real data Pan et al. built a directed graph (called a motion graph) in which the arcs are the observed sequences of signals and nodes are possible transitions between them. Then different paths in the graph correspond to different expressions of emotions. Thus, new animations can be generated by reordering the observed displays.

Another system for generation of nonverbal behavior was proposed by Michael Kipp (Kipp, 2006). This work focuses on nonverbal behavior that is synchronized with the verbal content. The system allows for both automatic and manual definition of nonverbal behaviors in four different modalities. In the manual mode the user adds commands that trigger a signal into the description of the scene by hand. In the second approach the signals are inserted automatically by the system using a set of predefined rules. Rules determine triggering conditions of the signal as the function of the text. Thus a signal can be triggered for example by the particular word, sequence of the words, type of sentence (e.g. question) or when the agent starts the turn. The system also resolves eventual conflicts between signals i.e. when the signals defined manually coincide with the signals added by the system. According to the conflict resolver the signals that overlap but use different modalities can be both displayed. Otherwise, the signals manually defined are preferred over those which were automatically generated. The system offers also the possibility to discover/learn new rules.

The possibility to produce realistic nonverbal acoustic content would also improve the agents' expressivity. Among the various nonverbal signals used (consciously or not) by humans to communicate feelings, laughter occupies a central place thanks to its frequent use, its broad range of shapes, its various significations (even if it is generally associated to positive feelings) and its contagious properties. Its analysis receives growing attention and there were interesting attempts to synthesize laughter. Lasarczyk and Trouvain (Lasarczyk and Trouvain, 2007) compared two systems for generating laughter sounds, inspired by works done in speech synthesis. The first is a 3D simulation of the vocal tract, the second consists in concatenating diphones from a small laughter database. Sundaram and Narayanan (Sundaram and Narayanan, 2007) tackled the problem under a different angle. Noticing that most of voiced episodes exhibit an oscillatory behavior, they modeled the envelope of the laughter waveform with a mass-spring analogy and synthesized the vowel-like sounds of laughters using Linear Prediction. Despite the quality of the models and the adaptation of speech synthesis techniques to laughter issues, the produced laughter samples do not sound natural. The proposed methods are indeed missing one important characteristic of human laughters: variability. To conclude, we would like to emphasize that laughter synthesis is not a sub-problem of emotional speech synthesis but a distinct research field. This point of view is corroborated by the fact that effective speech synthesis systems cannot produce human-like laughters.

3.3 Annotation

We aim to create an expressive multimodal agent. We took two approaches to reach this aim. On one hand we gather data from the literature and on the other hand we annotated ourselves visual data. We report below our work especially for the second approach. Work done for the first approach has been reported in the background section 3.1.

Besides gathering data from observational studies, a work of annotation has been done for specific emotions needed by the agent. Five emotional states (tension, relief, joy, sadness, anger) have been annotated with at least two cases per state. The audiovisual clips were extracted from TV series, from EmoTv corpus, Belfast Naturalistic Emotional Database audiovisual corpus and Humaine database. These clips were described in term of facial and gestural changes. The face has been annotated with the *Facial Action Coding System* (FACS, Ekman and Friesen, 1978) when appropriate and gestures have been described in a way to enable fast comprehension in order to prepare the generation of the movements on Greta .

For the generation of emotional expressions for ECAs, a great quantity of detailed information is needed on the actual movements, on their co-occurrences as well as on the link between the internal states and their expression, and so on. Some can be resolved simply by automatic synthesizing of captured data, others by theoretical studies. However the first method does not enable generalizations (a movement is not broken down into its elements and the elements cannot be compared in different cases) nor the understanding of a movement as an expression of a state. As for the last method, today detailed predictions and information on dynamical emotional expression is still scarce in research literature, particularly concerning the integration of different modalities of expression. One complementary way to obtain such data is the direct study of human behavior.

3.3.1 Observation of emotional behavior

When studying emotional behavioral and more generally nonverbal communication, one needs specific tools for the description of what is physically **expressed and perceived**. The face being one of the crucial modes of communication, as mentioned above (Section 3.1), techniques to measure facial expressions objectively and on a micro-analytic level are indispensable. Anatomically based coding systems like the *Facial Action Coding System* FACS (Ekman and Friesen, 1978) lend themselves to this purpose. This system is particularly interesting for exploratory studies as it is free of theoretical assumptions. Independent of any interpretation, it does not rely on prototypical expressions. It is more precise than any other existing facial coding technique, as in the case of EmFACS (Ekman and Friesen, 1975) which was developed by the same authors to tackle exclusively the elements considered to contribute to basic emotional expressions. FACS describes all the physically feasible facial movements in term of **action units** (AU). The action units stand for the minimal changes that can be visually perceptible and not for the contraction of one muscle, as in some cases more than one muscle is involved in a minimal facial movement (some muscles tend to act together). These minimal action units have each a numeric code. The code has to be attributed when the AU is perceived, along with its intensity.

Numerous coding schemes have been developed for the coding of gestures; however they are all focusing on specific research questions and are not a standardized tool. They are the work of individual research teams and they are not fit for all kind of studies. A general coding scheme including all the possible postural and body part movements would be cumbersome. As for the measure and description of gestures in the context of emotion, no well known and appropriate coding scheme is available. The researchers have to define their description level by themselves in relation with their needs.

3.3.2 Annotations for CALLAS

To complete the data obtained from theory and literature, audio-visual recordings of emotional displays have been annotated. The video extracts have been chosen from various sources, mostly from TV shows (from TV series, from EmoTV corpus, QUB Belfast Naturalistic Emotional Database and Humaine database). The observed people were non actors, placed in an emotional situation. By these measures the appearance of natural and not stereotyped behavior was encouraged.

The video clips were annotated with Anvil v4.7.6 (Kipp, 2001), a software enabling the description of audio visual recordings. This annotating tool is free for research and educational purposes. It allows frame by frame, as well as normal speed viewing of the annotated video. The criteria of annotations can be defined by the user in an XML file and the number of tracks can vary. Each track is to be dedicated to an element of annotation, such as a modality (face, body parts, voice, see Figure 3) or the evaluation of a finer grained element of that modality (one track for the evaluation of head movement to the left, one track for the head movement to the right). Each track can either have an open or a closed list of attributes. The attributes are defined in the XML file.

For the annotation of the multimodal recordings, five tracks have been defined: emotion, facial expression, head movement, gaze and gestures. The attributed emotion was defined mostly from the situation, while the evaluation of intensity was an overall impression from the situation and all the expressive modalities combined. The facial expression was described in term of AU from the FACS. Head, gaze and gestures were described verbally, in a way enabling the choice of pertinent elements for the generation of that expression on Greta's platform. An important aspect of that annotation based on the five tracks was the possibility of comparing the different emotional displays along the same modalities, e.g. to see that in joy there is forward movement of the torso while there will be none in tension, relief or sadness. The tracks do not have a predefined list of answer modalities and can be filled freely by the annotator.

So far, five emotional states have been chosen (tension, relief, joy, sadness, anger). At least two clips have been annotated per state.

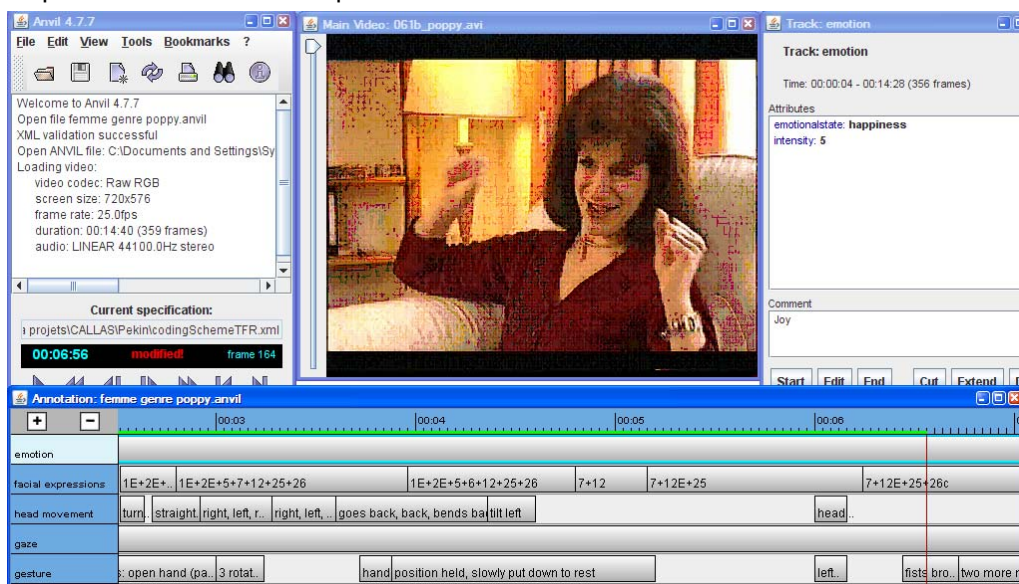


Figure 3. Illustration of video annotations with the Anvil software: multimodal display of joy from the Belfast Naturalistic Emotional Database, Cowie et al., 2003)

In the annotated displays, one can see that the face is often not the only source of information on the internal state of the perceived person. Thus in relief, it is the body's drop of

tension that is most spectacular, with its efferent movements, such as jaw dropping, backward projection of the head or thrust of the hands into the air. Often the facial mimicry is minimal and mostly insufficient for the understanding of the affective changes, e.g. closing of the eyes and mouth opening per se do not lead to an attribution of the label “relief”.

What is more, some movements seem to be an expression of an “undifferentiated” arousal: the general shaking of body parts (mostly of the head) is present in the strong emotional states of joy, anger and tension. In some states, however, one can see some movements typical of that emotion (see Figure 4 for the arm movements of joy).

To conclude, the manual annotation of the particular clips depicting emotional behavior seems a useful method for obtaining relevant data for the creation of repertoire of expressions for an ECA. With the individualized selection of clips, with among criteria the element feasibility on the Greta platform, the annotations complete the theoretical and observational studies and procure additional information concerning emotional nonverbal behavior.



Figure 4. Multimodality of the expression of joy: facial, torso and hand movements (extracts from a video clip from Naturalistic Emotional Database, Cowie et al., 2003)

3.4 Expressions of Greta

3.4.1 *Multimodal expressions of emotions*

As reported above, we have elaborated our model of multimodal expressions of emotions from two approaches, namely from the annotation of data and from data reported from the literature. For this latter one, we have particularly looked at the works of Dacher Keltner (Keltner, 1995; Keltner and Buswell, 1996; Haidt and Keltner, 1999; Keltner, 2005), Shiota, et al. (Shiota, et al., 2003) and Harrigan and O'Connell (Harrigan and O'Connell, 1996). From the analysis of expressions of emotions like embarrassment (Keltner, 1995), awe, amusement, pride (Shiota, et al., 2003) or anxiety (Harrigan and O'Connell, 1996) (see also Section 3.1) it was shown that certain emotions are expressed by a set of signals which are arranged in certain interval of time rather than by a static facial expression. The expressions of emotional states are dynamic and they can be displayed over different modalities like face, gaze and head movement, gestures, or even the posture. Interestingly, these signals do not have to occur simultaneously (Keltner, 1995).

To go beyond agents showing simply static facial expression of emotion, we have defined a representation scheme that encompasses dynamicity of facial expressions of an emotion. The main task of our algorithm is to generate the multimodal expressions of emotions, i.e.

expressions that are composed of different signals (or behaviors) partially ordered in time and with the use of different nonverbal communicative channels. These multimodal expressions can be of any duration while the respective signals have fixed durations (e.g. facial expressions of emotions usually are not longer than four seconds (Ekman, 1975) and gestures often have at least a minimum duration). In more details, we define for each emotional state a *behavior set* – a set of signals through which the emotion is displayed and a *constraint set* that defines relations between the signals in the behavior set. These two sets are defined from literature (see Section 3.1) and from annotations (see Section 3.3).

The single signals are described in the repositories (see Figure 2) while the behavior sets are described in the central database of the behaviors called *lexicon* (see Figure 5). It is XML-based file that contains mapping between the communicative intentions of the agent and behavior sets. Also the relations between signals (i.e. constraint sets) are described in XML-like format. Thus, from the single label of an emotional state (e.g. anger or embarrassment) our system generates sequences of multimodal expressions, i.e. the animation of a given duration composed of a sequence of signals on different modalities. It does so by choosing a coherent subset of signals from the behavior set, their durations, and order of display.

The algorithm can be seen as part of the Behavior Planner layer of the SAIBA architecture (see Section 2). The emotional state of the agent is described in FML (or FML-APML) language. Our model translates it to a set of behaviors described in BML. In the following subsections we present details of this process.

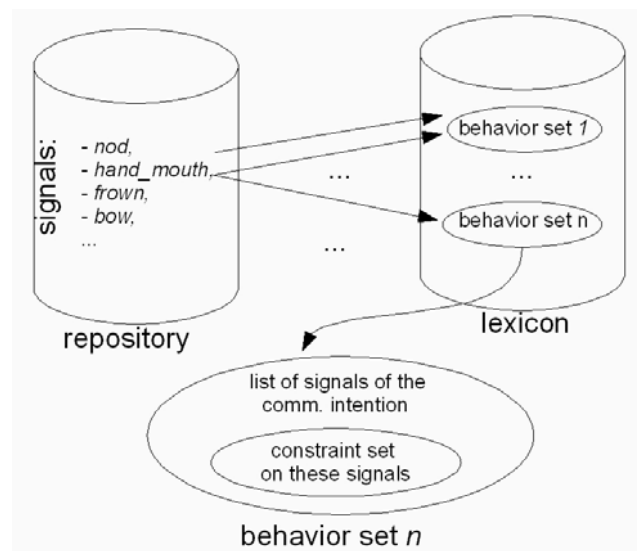


Figure 5. Multimodal expressions of emotions

Multimodal behavior sets

The *lexicon* file introduced in (Mancini and Pelachaud, 2008) is the central database describing the behaviors of the Greta agent. Each communicative function is defined by its type (e.g. performative) and value (e.g. announce or deny). In the *lexicon*, the communicative functions are mapped to behavior sets. The *behavior set* contains a set of signals of different modalities e.g. *head_nod*, *shaking-hand gesture* or *smile* to be displayed by Greta. Thus, one can define in the lexicon, for example, the communicative function *greet* of the type *performative* to be displayed by a head nod and/or shaking-hand gesture. The *lexicon* allows also one to precise certain temporal relations between signals belonging to one behavior set like *simultaneity* (two signals start and end at the same time) or *exclusion* (only one signal from *n* can be displayed).

We use the syntax of the lexicon file to describe the multimodal expressions of emotions. For example the behavior set for the expression of the emotion of embarrassment described by

Dacher Keltner (Keltner, 1995) can be defined by the following signals:

- two head movements: *head_down* and *head_left*,
- three gaze direction: *look_down*, *look_right*, *look_left*,
- three facial expressions: *smile*, *non-Duchenne smile*, and *neutral expression*,
- gesture: *open flat hand on mouth*,
- torso movement: *bow*.

Keltner observed a number of regularities in expressions that concern the signal duration and the order of displaying. For this reason we add a number of variables to the description of signals in the behavior set to describe these features. For each signal one can also define:

- *probability_start* - the probability of the occurrence at the beginning of a multimodal expression (a value in the interval [0..1]),
- *probability_end* - the probability of the occurrence at the end of a multimodal expression (a value in the interval [0..1]),
- *min_duration* – the minimum duration of the signal (in seconds),
- *max_duration* – the maximum duration of the signal (in seconds),
- *repetitivity* – certain signals should not be repeated during one expression (a value from a set {0,1}).

Below we present an example of the behavior set for the emotion of embarrassment.

```
<multimodal emotion="embarrassment">
<signals>
<signal id="1" name="head=head_down" repetitivity="0" min_duration="2" max_duration="4" probability_start="0.8"
probability_end="0.3"/>
<signal id="2" name="head=head_left" repetitivity="0" min_duration="5" max_duration="9"
probability_start="0.1" probability_end="0.8"/>
<signal id="3" name="gaze=look_down" repetitivity="0" min_duration="2" max_duration="4" probability_start="0.9"
probability_end="0.3"/>
<signal id="4" name="gaze=look_right" repetitivity="0" min_duration="1" max_duration="2" probability_start="0.5"
probability_end="0.5"/>
<signal id="5" name="gaze=look_left" repetitivity="0" min_duration="1" max_duration="2" probability_start="0.5"
probability_end="0.5"/>
<signal id="6" name="affect=smile" repetitivity="1" min_duration="2" max_duration="4" probability_start="0.6"
probability_end="0.6"/>
<signal id="7" name="affect=not_duchenne_smile" repetitivity="1" min_duration="2" max_duration="4"
probability_start="0.6" probability_end="0.6"/>
<signal id="8" name="affect=neutral" repetitivity="1" min_duration="1" max_duration="2" probability_start="0.3"
probability_end="0.3"/>
<signal id="9" name="emotional=hand_mouth" repetitivity="0" min_duration="3" max_duration="5"
probability_start="0.2" probability_end="0.7"/>
<signal id="10" name="bow" repetitivity="0" min_duration="5" max_duration="7" probability_start="0.4"
probability_end="0.9"/>
</signals>
```

Figure 6. An example of the behavior set for the emotion of embarrassment.

In this example the emotion of embarrassment can be displayed by 10 different signals. Only three of them (smile, non-Duchenne smile and neutral) can be repeated during one expression (the value of the parameter *repetitivity* for these signals is 1). The signals

head_down and *gaze_down* will occur much more often at the beginning of the multimodal expression. On the contrary the signals *head_left*, *touch_face*, and *bow* will occur much more often at the end of the expression.

Constraint sets

The signals in multimodal expressions of emotions do not occur totally by chance ((Keltner, 1995), see also Section 3.1). We define for each emotional state a constraint set that describes all plausible configurations of signals. This set introduces a set of limitations on the occurrence and duration (i.e. on the values for start and end time) of the signal in relation to others signals. Typical examples of these relations that were observed in multimodal expressions of emotions are:

- S1) signals s_i and s_j occur contemporarily (i.e. they start and end together),
- S2) only one signal $s_i \in S$ of a set S occurs at a time,
- S3) signal s_i starts (ends) before the signal s_j starts (ends),
- S4) signal s_i starts and ends before (after) the signal s_j starts (ends),
- S5) only one signal s_i of a set S occurs during the animation,
- S6) signal s_i cannot (has to) start (end) the animation etc.

Defining the constraints we use a low-level notation. Each signal is described by two variables that correspond to its start time and end time (in seconds). Then arithmetic operators and relations of (in)equality are used to describe the relations between signals. We distinguish between two types of constraints. The first type describes the relation between two signals. It is denoted by a tag *constraint* which has one parameter: *type*. The *type* is used to define the arithmetic operation. Two nested tags *arg* precise two signals for which the constraint is defined. The last tag of the block defines the arithmetic relation between them. It can be one of the following tags: *morethan*, *lessthan*, or *equal*. For example the constraint:

```
<constraint type="minus">
  <arg id="1" type="end"/>
  <arg id="2" type="start"/>
  <morethan value="0"/>
</constraint>
```

It means that the signal 2 (i.e. *head_left*) can start only after the signal 1 (*head_down*) ends ($S2.start - S1.end > 0$).

The second type of constraints serves to introduce the limitations on the start and the end time of any signal. We can for example introduce numerical constraints on these variables. In this case the tag *constraint* has only one nested tag *arg* and value of the *type* parameter is "oneargument" E.g.:

```
<constraint type="oneargument">
  <arg id="1" type="start"/>
  <morethan value="1"/>
</constraint>
```

It means that the start time of the signal 2 (*head_left*) has to be longer than 1 second (i.e. the expression cannot start with this signal). Using the syntax presented above it is possible to describe, among others, the cases S1 – S6.

Algorithm

Our algorithm works as follow. Let A be the animation to be displayed by Greta. A can be seen as a set of triples $A = \{(s_i, st_i, sp_i)\}$ where s_i is the name of the signal, st_i is the start time of the signal s_i and sp_i is its stop time. The input to the system is an emotional label, e , and its expected duration, t . At the beginning A is empty. In the first step the algorithm chooses a set of multimodal behaviors $S_e = \{s_i\}$ corresponding to the emotional state e . Then the algorithm decides the number of time stamps, n , as a function of the duration t and the values of the expressivity parameters (see Section 3.4.4). Next, at each time step, t_j , the system chooses randomly a signal-candidate s_c between the signals of the set S_e considering their probabilities of occurrence. For this purpose it manages a table of probabilities that contains, for each signal s_i , its current probability value $p_i(t_j)$. Obviously, at the first time stamp, $t_0 = 0$, the values of this table are equal to the values of the variable *probability_start*, while at the last time stamp $t_{n-1} = t$ the probabilities are equal to the *probability_end* value. At each time stamp, t_j , the probabilities $p_i(t_j)$ of each signal s_i are updated.

The candidate for a signal to be displayed s_c in a turn t_j is chosen using the values $p_i(t_j)$. Next, the start time st_c of the candidate s_c is randomly chosen from the interval $[t_j, t_{j+1})$ and the consistence of (s_c, st_c, \dots) with the partial animation A is checked. If all the constraints are satisfied for the partial animation A and (s_c, st_c, \dots) stop time st_c of S_c , is randomly chosen between two values:

$$sp_{c1} = min_{sc} + \frac{R \cdot (max_{sc} - min_{sc})}{2}, sp_{c2} = max_{sc} - \frac{R \cdot (max_{sc} - min_{sc})}{2}$$

where R is a value from the interval $[0..1]$, $max_{sc} = max_duration$ of S_c while $min_{sc} = min_duration$ of s_c . Otherwise the other signal from S_e is chosen as a candidate.

The consistency of the triple (s_c, st_c, sp_c) with the partial animation A is checked again. If all the constraints are satisfied the triple (s_c, st_c, sp_c) is added to A . The table of probabilities is updated (if repetitivity of s_c is 0 then $p_c(t_{j+z}) = 0$, $z = 1..n-j$) and the algorithm chooses another signal, moves to next the time stamp, or finishes generating the animation.

The algorithm presented above is able to generate a number of animations that are consistent with the constraints. In this way we avoid the schematization of the agent behavior - a common problem of other algorithms generating repetitive behavior for ECAs.

Example. In Figures 7 and 8 two examples of the animation for the expression of embarrassment are shown. In Figure 7 the following images present the frames of animations of Greta displaying respectively the signals: a) *look_right*, b) *head_down* and *gaze_down*, c) *gaze_left*, d) *gaze_left* and *non-Duchenne_smile*, e) *gaze_left*.

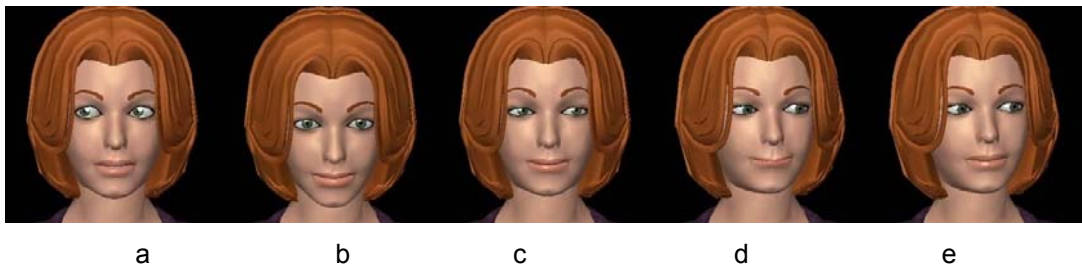


Figure 7. An example of multimodal expression of embarrassment.

In Figure 8 the following signals are displayed: a) *neutral expression*, b) *smile*, c) *smile* and *gaze_right*, d) *gaze_left*, e) *gaze_down* and *head_down*, f) *touching face gesture*.

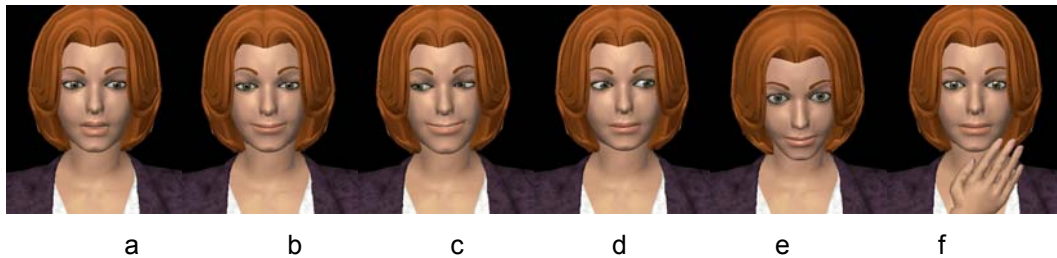


Figure 8. An example of multimodal expression of embarrassment.

3.4.2 Model of complex facial expressions

Our model of complex facial expressions is based on Paul Ekman's studies (Ekman, 1975; Ekman, 2003a; Ekman, 2003b) (see also Section 3.1). We define complex facial expressions using a face partitioning approach. Each facial expression is defined by a set of eight facial areas (i.e., brows, upper eyelids, eyes direction, lower eyelids, cheeks, nose, lips, lips tension). An expression is a composition of these facial areas, each of which can display signs of emotion. For complex facial expressions, different emotions can be expressed on different areas of the face; e.g., in sadness masked by happiness, sadness is shown on the eyebrows area while happiness is displayed on the mouth area.

The main task of our algorithm is to assign expressions of emotion to different parts of the face. For this purpose we define for each type of complex facial expressions a set of rules that describe the composition of the facial areas. These rules, based on the description proposed by Ekman, refers to six emotions, namely: anger, disgust, fear, joy, sadness, and surprise. Let's call a set of the expressions of these emotions BASEXP. Then, for an input emotion for which the facial expression is not defined explicitly by our rules (e.g. expression of contempt or disappointment) our algorithm chooses the most appropriate solution. For this purpose we use the algorithm FS based on fuzzy similarity (see Niewiadomski and Pelachaud, 2007b). In this approach each facial expression is described by fuzzy sets. Fuzzy similarity is used to compute the degree of visual similarity between them. Our algorithm compares any two facial expressions and outputs a value of similarity in the interval $[0..1]$ (0 meaning "not similar at all" while 1 means identical expressions). Once the most similar expression (chosen among the 6 ones) is known we can apply corresponding rules to the input expression. The rules determine which elements of the input expression are used in the complex facial expression.

In the following sections we present our algorithm for different types of complex facial expressions: superposition, masking, fake or inhibited expressions².

Superposition

Superposition happens when two emotions are felt at the same time. The resulting expression has some features of the expressions of both felt emotions, like in the expression of superposition of joy and sadness described by Paul Ekman where the raised brows of sadness are accompanied by a smile (Ekman (1975, 2003b)). The superposition of two emotions is usually expressed by a combination of the upper part of one expression with the lower part of the other one (Ekman (1975, 2003b)).

Superposition schemes. The six emotions analyzed by Ekman give us 30 different ordered

² For the detailed description of the algorithm see also (Niewiadomski and Pelachaud 2007a, Niewiadomski and Pelachaud 2007b)).

pairs of emotions. We classified them according to their face partition. It allows us to distinguish 10 different superposition schemes. By superposition scheme SS_i we mean a particular division of the eight face areas between any two emotions e.g. the facial areas F_1, F_2, F_3, F_4 (forehead, brows, eyelids, and eyes) belong to the expression of the first emotion and the facial areas F_5, F_6, F_7, F_8 (nose, cheeks, and lips) to the second one.

By Z we denote a set of superposition schemes of Ekman's expressions. Two different pairs of emotions can share the same superposition scheme. It means that two different ordered pairs of emotions divide the face in the same way e.g. both pairs: sadness and fear as well as anger and happiness create superposition expression in which the F_1, F_2, F_3, F_4 are taken from the first expression (sadness or anger respectively) while the F_5, F_6, F_7, F_8 are taken from the second element of the ordered pair (fear or happiness respectively).

Algorithm. The algorithm generates the expression of superposition for any two expressions $Exp(E_i)$ and $Exp(E_j)$ by choosing one superposition schema SS_i from the set Z of superposition schemes. In the first step, for each input (i.e. simple) expression $Exp(E_i)$ we establish its values of similarity with the expressions in the BASEXP set. Any simple expression can be represented by a vector with values in the interval $[0..1]$ that correspond to the degrees of similarity between that expression and the one from BASEXP (Niewiadomski and Pelachaud, 2007b). A set of rules SFR_{sup} is used to create an expression of superposition. For each pair of expressions from BASEXP we define a rule that associates it with one SS_i . The output of the system of rules is an SS_i according to which the final expression is composed.

Notation. The complex facial expressions can be described in BML language by using the *reference* extension tag: (see Section 2.2.1). The syntax for the expression of superposition is the following:

`<reference>affect=complex:em1_and_em2</reference>`

where *complex* is a keyword while *em1* and *em2* are the names of emotions to be superposed.

Example. Figure 9 presents an example of the superposition expression computed by our model. Figures 9a and 9b show the expressions of anger and sadness respectively. Figures 8c and 8d show the superposition as a composition of face areas of both input expressions. In the Figure 9d we can see which parts of the face correspond to sadness and which ones to anger. In that image the areas F_5, F_6, F_7 , and F_8 (expressing sadness) are marked out with by the yellow circles while areas F_1, F_2, F_3 , and F_4 (expressing anger) by a red color.

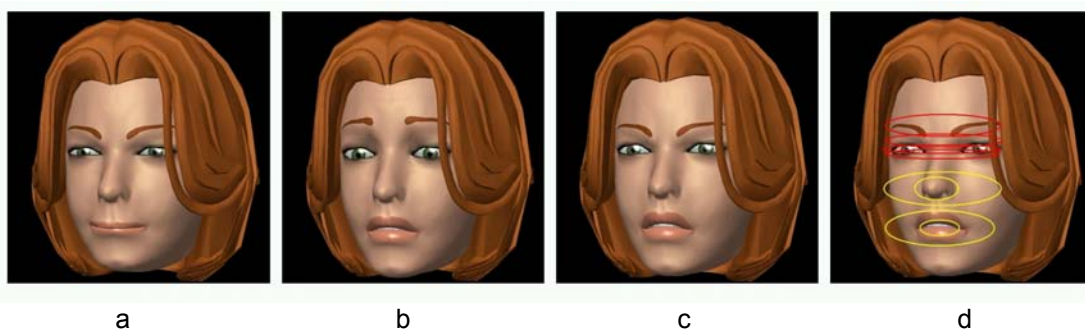


Figure 9. Superposition of anger and sadness. From the left to right: anger a), sadness b), superposition of anger and sadness c) superposition of anger and sadness with significant areas marked d).

Masking

Masking occurs when a person decides for some reason not to display her felt emotion and prefers to display a different emotional expression. The expression of masking is influenced

by deception clues described in Section 3.1. The felt emotion leaks over the mask according to the inhibition hypothesis. On the other hand, the fake expression is not complete as it lacks the reliable features.

Rules. For each deception clue we define in our model a separate set of rules. The SFR_{inh} describes which elements of the felt facial expression are expected to appear even if the expression is concealed. Then the SFR_{rf} specifies the face areas that do not occur in a fake expression. In order to create the facial expression of masking we use both sets of rules.

SFR_{rf} is a set describing rules for reliable features of expressions from BASEXP. Each rule indicates the reliable features of an expression. We map each reliable feature to the facial areas F_1, \dots, F_8 . Each output variable of the rule corresponds to one facial region of the resulting expression and expresses the possibility of occurrence (POS) of it. For example the rule RF_4 is: the more the input expression is (similar to) the expression of *happiness*, the more the possibility of occurrence of the lower eyelids area of the input expression is *low*, the possibility of occurrence of brows and upper eyelids areas is *may occur*, and the possibility of occurrence of other areas is *high*.

Another set of rules, SFR_{inh} , is defined on the basis of the inhibition hypothesis. Each rule of SFR_{inh} indicates the elements of facial expressions that leak over the mask. For this purpose we map each leaking feature to facial area from the set F_1, \dots, F_8 . Activity in these facial areas can be observed even if the expressions are inhibited.

Algorithm. The algorithm generates the expression of masking for expressions of any two emotions: the felt one (E_i) and the fake one (E_j). First, the values of fuzzy similarity FS are established for their expressions and the elements of BASEXP. Similarly as it was in the case of superposition, each facial expression $Exp(E_i)$ (resp. $Exp(E_j)$) is represented by a six elements vector $[a_1, \dots, a_6]$ (resp. $[b_1, \dots, b_6]$) of real values from the interval $[0..1]$. Then the elements of the final expression are processed separately. The vector $[a_i]$ of the felt expression E_i is processed by SFR_{inh} , while the vector $[b_i]$ of the fake expression E_j is processed by SFR_{rf} .

SFR_{inh} and SFR_{rf} are complementary. Both of them return the predictions about the occurrence of certain F_k . For each facial area the results of SFR_{rf} and SFR_{inh} are combined in order to obtain the masked expression. In particular, for each facial area the following can happen:

- C1) possibility of occurrence of k -th facial area of the felt expression $Exp(E_i)$ is high and the possibility of occurrence of k -th facial area of the fake expression $Exp(E_j)$ is also high. It means that k -th facial areas of both expressions: $Exp(E_i)$ and $Exp(E_j)$ are candidates to be shown. In this case the felt expression should be expressed as it is difficult to inhibit it voluntarily.
- C2) possibility of occurrence of k -th facial area of the felt expression $Exp(E_i)$ is high and possibility of occurrence of k -th facial area of the fake expression $Exp(E_j)$ is low. Then the k -th facial area of $Exp(E_i)$ is used.
- C3) possibility of occurrence of k -th facial area of the felt expression $Exp(E_i)$ is low and possibility of occurrence of k -th facial area of the fake expression $Exp(E_j)$ is high. Then the k -th facial area of $Exp(E_j)$ is used.
- C4) possibility of occurrence of k -th facial area of the felt expression $Exp(E_i)$ is low and possibility of occurrence of k -th facial area of the fake expression $Exp(E_j)$ is low. It means that neither k -th facial area of $Exp(E_i)$ nor of k -th facial area of $Exp(E_j)$ can be used. In this situation k -th facial area of neutral expression is used instead.
- C5) The possibilities of occurrence of k -th facial area of $Exp(E_i)$ and k -th facial area of $Exp(E_j)$ are somewhere between high and low. It means that both may occur. The facial area is chosen randomly between $Exp(E_i)$ and $Exp(E_j)$.

Thus, the final expression is composed of facial regions of the felt, the fake, and the neutral emotion.

Notation. The complex facial expressions can be described in BML language by using the *reference* extension tag: (see Section 2.2.1). The syntax for the expression of masking is the following:

`<reference>affect=complex: em1_maskedby_em2</reference>`

where *complex* is a keyword while *em1* is the name of felt emotion, while *em2* is the name of the fake one.

Example. Figure 10c presents the agent displaying the expression of disappointment masked by a fake happiness. Let us explain how we obtain the complex expression displayed by the agent. We applied our similarity algorithm and found that disappointment has a facial expression very similar to sadness. In our model the features of felt sadness that leak over the masking expression are: forehead, brows, and upper eyelids. These elements are represented by the facial areas F_1 (forehead and brows) and F_2 (upper eyelids). According to the inhibition hypothesis, they can be observed in masked sadness. The expression of disappointment (Figure 9a) is found to be very similar to the expression of sadness according to the similarity algorithm. So the rules of sadness will be applied also in the case of disappointment expression. In the expression of disappointment masked by fake joy (Figure 10c) we can notice the movement of brows, which is a characteristic of disappointment. On the other hand, the mouth area displays a smile (sign of happiness).

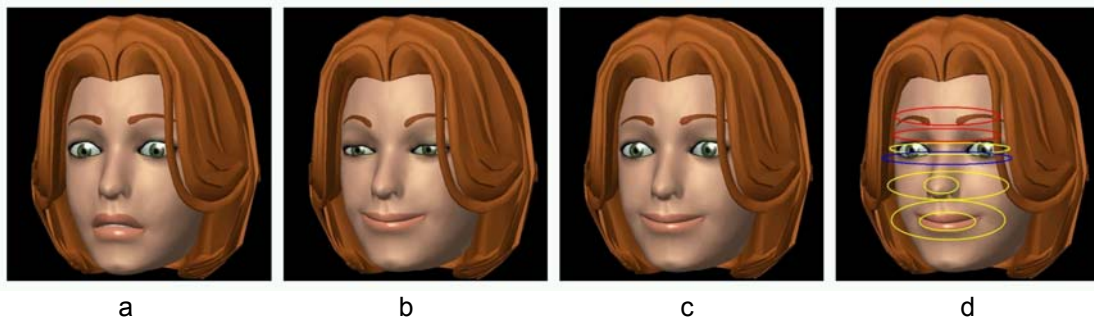


Figure 10. Disappointment masked by happiness. From the left to right: disappointment a), happiness b), disappointment masked by happiness c) disappointment masked by happiness with significant areas marked d).

Fake and inhibited expressions

Similarly to other cases of facial expression management, the fake and inhibited expressions can be detected by masking clues. The felt emotion leaks over the facial mask according to the inhibition hypothesis while a fake expression is incomplete as it lacks reliable features. It means that SFR_{inh} needs to be used for inhibited expressions, while SFR_{rf} is used for fake expressions. The expression of inhibition can be seen as hiding the felt emotion under the “mask” of the neutral expression. Similarly, making a fake expression means changing the neutral facial expression to some fake expression. Thus we can use the same procedure that we used in the case of masking for fake or inhibited expressions. We introduce a slight modification to the algorithm presented in the previous section: we add rules for the neutral expression to the sets SFR_{rf} , SFR_{inh} . We assume that “false neutral expression” can easily be made deliberately. Thus we add to SFR_{rf} a new trivial rule (RF_7): “the more the input expression is (similar to) the neutral expression, the more the possibility of occurrence of any area is high”. Then we add also a new trivial rule (INH_7) to SFR_{inh} . It represents a fact: “the more the input expression is (similar to) the neutral expression, the more the possibility of occurrence of any area is low”. It is so as the neutral expression does not involve any particular facial movement.

Notation. The complex facial expressions can be described in BML language by using the *reference* extension tag: (see Section 2.2.1). The syntax for the fake expression is the

following:

```
<reference>affect=complex:neutral_maskedby_em1</reference>
```

where *complex* is a keyword while *em1* is the name of emotion. The inhibited expression is generated by the command:

```
<reference>affect=complex:em1_maskedby_neutral</reference>
```

where *complex* is a keyword while *em1* is the name of emotion.

Example. In Figure 11 we can see another complex facial expression of non-basic expression, i.e. the inhibited expression of contempt (Figure 11c and 11d). We can compare it with the felt expression of contempt (Figure 11a) and the neutral expression (Figure 11b). The facial expression in Figure 11b is different from the one in Figure 11c: the eyebrows and nose wrinkling in Figure 11c. The expression of contempt is considered as very similar to the expression of disgust. Then the facial areas F_1 (eyebrow) and F_5 (nose) cover the features of felt disgust that leak over the mask. As a consequence, they can be observed in inhibited disgust and thus they can be observed also in inhibited contempt. These facial areas are signaled in Figure 11d.

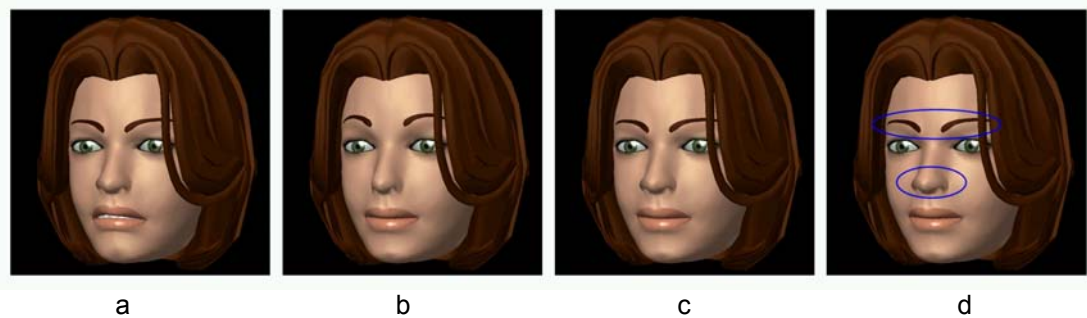


Figure 11. Inhibited contempt. From the left to right: contempt a), neutral expression b), inhibited contempt c), inhibited contempt with significant areas marked d).

3.4.3 Facial expressions of different intensities

Our model of facial expressions of different intensities is based on Paul Ekman's studies (Ekman, 1975; Ekman, 2003a) (see Section 3.1). According to him the intensity of a felt emotional state can be expressed in at least two different ways: by modulating the intensity of facial muscles contractions or by so called partial expressions i.e. facial expressions that involve only certain regions of the face e.g. the mouth area, the forehead, or the eyes area. Both approaches are implemented in Greta.

Algorithm. The following algorithm (see Figure 12) is used to generate facial expressions of emotions of different intensities. The input to the system is the label of the emotional state, *e*, and the intensity value, *int*, from the interval [0..1]. The different intensities can be expressed in two different ways in Greta:

1. by partial facial expressions, or
2. by increasing/decreasing all the values of the facial animation parameters (FAPs).

The default approach is inspired by Becker et al.'s experiment (Becker et al, 2005). According to this result the low intensity displays of emotions generated with the second approach often were not recognized by the users. Thus, the first method is preferred in our algorithm if the value of intensity of the emotional state is low. On the contrary if the emotional state is characterized by a high intensity more probably the second approach will be chosen. The method of expression generation can also be chosen explicitly by the user in Greta's configuration file. It can also be chosen randomly by the system.

When the facial expressions are generated using the first method (i.e. by partial expressions) the flow of our algorithm is the following. First, the algorithm checks in the repository of facial expressions if there is an explicit description of partial facial expressions of the emotion e . From the literature we know the partial facial expressions for six facial expressions (see Section 3.1). Certain emotions like anger have many different partial expressions. Each facial expression that occurs in the repository file is associated with a sub-interval of the interval of possible values of the variable int . For example the emotion of *anger* has four partial expressions that are associated with the intervals $[0..0.25]$, $[0.26..0.5]$, and so on. The partial expression such that value of int belongs to its interval is displayed. In the case the explicit definition is not present in the repository the partial facial expression depends on the valence of its emotion. If the emotion e is positive the lower face of its full-blown (default) expression is displayed, the upper face is used instead when the emotion e is negative. The algorithm uses the values of valence from (Albrecht et al, 2005).

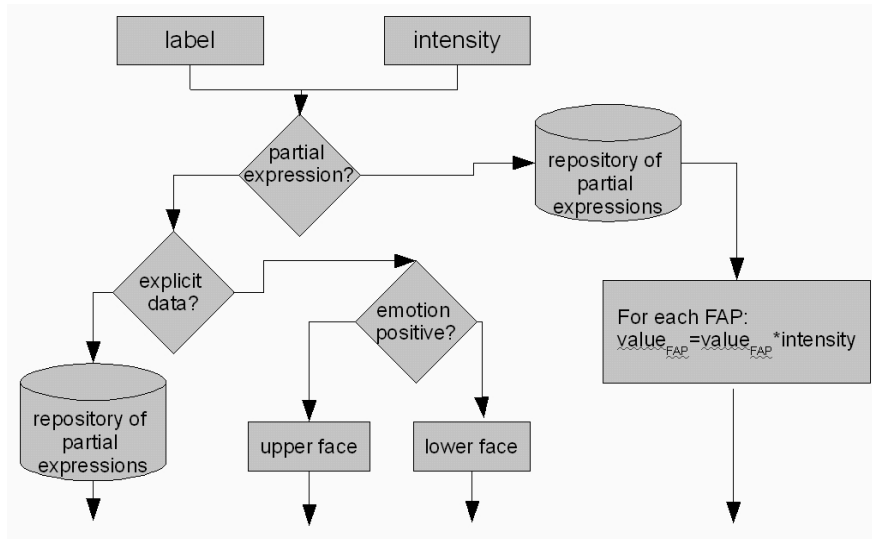


Figure 12. Generation of facial expressions of different intensities.

In the second approach the algorithm looks for the full-blown expression of e in the repository file. Then it modifies the values for each FAP according to the following formula:

$$FAP_{value} = FAP_{value} \cdot \sqrt{intensity}$$

Notation. The intensity of the emotion can be defined in the FML language by using the parameter *intensity* of the tag *emotion* (see Section 2.2.2 for details). The default value is 1.

Example. In Figures 13 and 14 facial expressions of fear of different intensities are presented. In Figure 13 facial expressions are created with the first approach. In particular, in Figure 13a the fear expression is expressed only by eyes and eyebrows, in Figure 13b, by the mouth area, while in Figure 13c the full-blown facial expression of fear is presented.

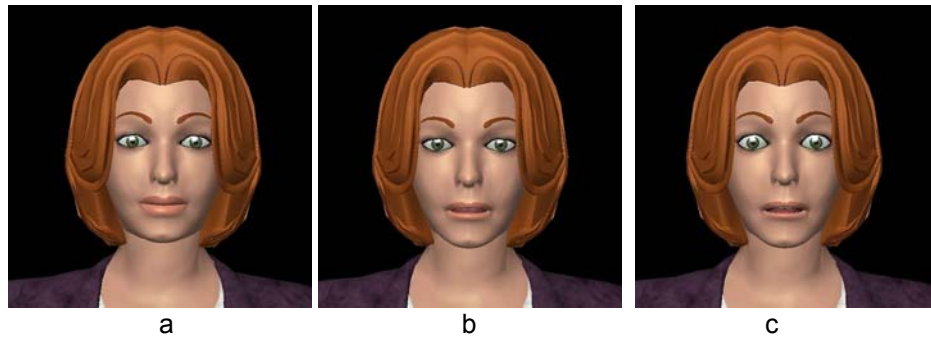


Figure 13. Facial expressions of fear generated with the first approach.

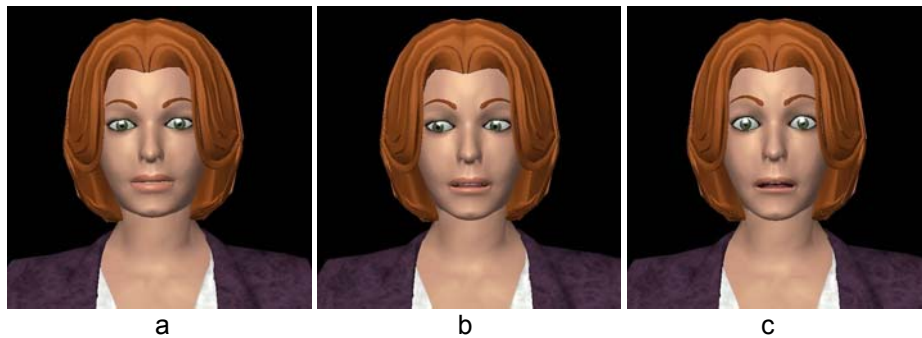


Figure 14. Facial expressions of fear generated with the second approach.

In Figure 14 the same emotion is expressed using the second approach.

3.4.4 Full-body expression of communicative intentions

From the research on human behavior, we know that people differ in the way they use their modalities: one can be very expressive on the face, another can gesture a lot. The concept of *modalities preferences* encompasses this variability in the modalities use. People can also differ in the quality of their behavior. For example, one can have the tendency to do large hand gestures. Both concepts are implemented in Greta. Thus it can display the same communicative intention in a number of different ways:

- using different modalities and signals,
- using the same signals but with different expressivity.

Modalities preferences

Greta can communicate to the user multimodally, that is by using many modalities at the same time. It produces signals on the following modalities:

- face (eyebrows/eyelids/mouth/cheek movements),
- head movement (head direction and rotation, such as nods and shakes),
- gestures (arms and hands movements),
- body posture (upper part of the body movements).

The modalities preferences represent the agent's degree of preference of each available modality. If for example we want to specify that the agent has the tendency to mainly use hand gestures during communication we assign a high degree of preference to the *gesture*

modality, if it uses mainly the face, the face modality is set to a higher value, and so on. For every available modality (face, head movement, gesture, posture), we define a value between 0 and 1 which represents its preferences. Agents can also use two or more modalities with the same degree of preference. This means that the agent will communicate with these modalities equally.

Expressivity of behavior

Expressivity of behavior is an integral part of the communication process as it can provide information on the current emotional state, mood, and personality of the agents. We have defined and implemented (Hartman et al., 2005; Mancini et al., 2007) a set of parameters that affect the qualities of the agent's behavior such as its speed, spatial volume, energy, fluidity. Thus, the same gestures or facial expressions are performed by the agent in a qualitatively different way depending on the following parameters:

- **Overall Activity:** amount of activity (e.g., passive/static versus animated/engaged). This parameter influences the number of single behaviors occurring during the communication. For example, as this parameter increases, the number of head movements, facial expressions, gestures and so on, increases. Its value is a floating point number ranging from 0 to 1 where a value of zero corresponds to no activity, and a value of one corresponds to maximum activity;
- **Spatial Extent:** amplitude of movements (e.g., expanded versus contracted). This parameter determines the amplitude of, for example, head rotations and gestures. The attribute, like all the following, is a real number defined in the interval $[-1..1]$. A value of zero corresponds to a neutral behavior, that is, the behavior of the agent without any expressivity control; in such a case, the agent performs nonverbal signals with the amplitude that was defined by the system designer. A value of -1 corresponds to the reproduction of very small and contracted movements (e.g. head rotations), while value of 1 corresponds to very wide and large movements;
- **Temporal Extent:** duration of movements (e.g., quick versus sustained actions). This parameter modifies the speed of execution of movements. They are slow if the value of the parameter is negative, or fast when the parameter is positive;
- **Fluidity:** smoothness and continuity of movement (e.g., smooth, graceful versus sudden, jerky). Higher values allow smooth and continuous execution of movements while lower values create discontinuity in the movements. Figure 15(a) shows the same movement executed with different fluidity values.

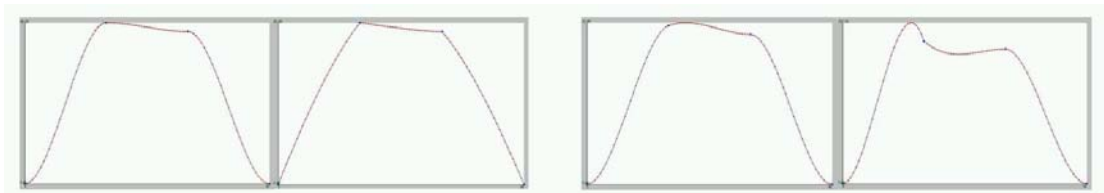


Figure 15: (a) Fluidity variation: left diagram represents high fluidity, right diagram represents low fluidity for the same behavior. (b) Power variation: left diagram represents movement executed with low power; while the right diagram represents the same movement with high power.

- **Power:** dynamic properties of the movement (e.g., weak/relaxed versus strong/tense). Higher (resp. lower) values increase (resp. decrease) the acceleration of the head or limbs rotation, making the overall movement look more (resp. less) powerful. Increasing this parameter also produces movement overshooting. Figure 15(b) shows some examples of curves with different power.
- **Repetitiveness:** this parameter permits the generation of rhythmic repetitions of the

same rotation/expression/gesture. For example, a head nod with a high repetitiveness becomes a sequence consisting of very fast and small nods.

Baseline and dynamicline

In our model we want to capture the idea that people have tendencies that characterize globally their behavior, but these tendencies can change in some situations, due to some events. To encapsulate this global and local qualities we have introduced the concepts of *baseline* and *dynamicline*, which both contain information on the agent's modalities preferences and expressivity but with different time span: while the *baseline* is the overall definition of how the agent behaves in general situation, the *dynamicline* is the *local* specification of the agent's behavior (for example during a given agent's emotional state). In our model, baseline and dynamicline do not only differ by their meaning (global vs. local behavior tendency) but also by the fact that the baseline is an input parameter, that is, it is used to define some characteristics of an agent, while the dynamicline is automatically computed by the system at runtime, depending on the current agent's communicative intention and/or emotional state.

The **baseline** of an agent has to be defined manually before running the system. We define the baseline by the pair $(Mod, Expr)$ where: *Mod* represents the modalities preferences (a value from the interval $[0..1]$ for each modality) while *Expr* is the behavior expressivity. This is the set of expressivity values that represents the base behavior tendency of the agent. There is a set of expressivity parameters for each modality.

The **dynamicline** is computed at runtime from the agent's baseline and the agent's communicative act or emotional state. So, each time, a new communicative intention and/or emotional state arrives in input, our system will compute a new dynamicline for the agent. Dynamicline is modeled by the pair $(Mod-Dyn, Expr-Dyn)$ where:

- *Mod-Dyn*: the current agent's modalities preferences. It represents the agent's tendency to use its modalities given a certain communicative intention and/or emotional state. It is obtained by modulating the modalities preferences *Mod* of the baseline depending on the actual communicative act and/or emotional state.
- *Expr-Dyn*: the current agent's expressivity parameters. It represents the agent's expressivity of movements given a certain communicative intention and/or emotional state. It is obtained by modulating the expressivity parameters *Expr* of the baseline depending on the actual communicative act and/or emotional state.

Behavior qualifiers

Communicative intentions and emotional states may influence the way one tends to communicate nonverbally. We call *behavior qualifier* the set of *modulations* that, given an emotional state or a communicative intention, acts on the behavior tendency of a conversational agent. A *modulation* is defined as a variation over one of the parameters contained in the baseline of an agent. It is represented by the following components:

- **destination**: it is the parameter in which the result of the modulation will be stored. For example it can be one of the modalities preference, or an expressivity parameter.
- **operator**: it specifies which operation should be performed between the terms listed in the modulation definition. The actual operators implemented in our system are simple mathematical operations like an addition, subtraction, multiplication, division, scaling. We have also defined an assignment operator to copy values between parameters.
- **list of terms**: it is the list of the modulation terms. Each term can be one of the

modality preferences, an expressivity parameter, or a numeric value. The number of terms depends on the operator, for example a simple assignation (e.g., $SPC = 1.0$) will need just one term, while a sum (e.g., $SPC = SPC + 0.5$) will need two terms.

As an example, let us see how we define a behavior qualifier that represents the following description: a hot anger state (i) increases the degree of bodily activation and at the same time (ii) the speed, amplitude and power of movements will be very high. Let us notice that the modulations described in this behavior qualifier are of two kinds: relative and absolute. In the example, part (i) of the qualifier says that the degree of bodily activation increases. This is a relative variation since it gives an indication of the local behavior tendency (dynamicline) in terms of the global tendency (baseline). Instead, part (ii) of the qualifier indicates that speed, amplitude and power of movement should be very high: in this case it refers to absolute values; that is, the local behavior tendency (dynamicline) is explicitly defined, and it does not refer to the global tendency (baseline). 16 shows the code representing this behavior qualifier. The lines 3 - 28 describe the modulations that act on the Overall activation expressivity parameter of the agent's body, face and gesture modalities by multiplying it by 1.5. These relative variations are described in part (i) of the behavior qualifier of "hot anger". On the other hand, lines 29 - 48 describe the modulations that assign the value 0.9 in the Temporal, Spatial and Power expressivity parameters of the agent's gesture modality. These absolute variations are described in part (ii) of the example.

```

01 <qualifier name="hot anger">
02
03   <modulation>
04     <destination>body</destination>
05     <parameter>OAC.value</parameter>
06     <operator>MULT</operator>
07     <term1name>body</term1name>
08     <term1attribute>OAC.value</term1attribute>
09     <term2value>1.5</term2value>
10   </modulation>
11
12   <modulation>
13     <destination>face</destination>
14     <parameter>OAC.value</parameter>
15     <operator>MULT</operator>
16     <term1name>face</term1name>
17     <term1attribute>OAC.value</term1attribute>
18     <term2value>1.5</term2value>
19   </modulation>
20
21   <modulation>
22     <destination>gesture</destination>
23     <parameter>OAC.value</parameter>
24     <operator>MULT</operator>
25     <term1name>gesture</term1name>
26     <term1attribute>OAC.value</term1attribute>
27     <term2value>1.5</term2value>
28   </modulation>
29
30   <modulation>
31     <destination>gesture</destination>
32     <parameter>TMP.value</parameter>
33     <operator>VAL</operator>
34     <term1value>0.9</term1value>
35   </modulation>
36
37   <modulation>
38     <destination>gesture</destination>
39     <parameter>SPC.value</parameter>
40     <operator>VAL</operator>
41     <term1value>0.9</term1value>
42   </modulation>
43
44   <modulation>
45     <destination>gesture</destination>
46     <parameter>PWR.value</parameter>
47     <operator>VAL</operator>
48     <term1value>0.9</term1value>
49   </modulation>
50 </qualifier>

```

(i) { lines 3-28 } (ii) { lines 29-48 }

Figure 16. Example of behavior qualifier definition.

Dynamicline computation

The agent's baseline and the agent's communicative act are used to compute the *dynamicline*. During the process, the modalities preferences and the expressivity parameters contained in the baseline are modulated depending to the agent's actual communicative intention and/or emotional state and the resulting values are stored in the dynamicline. It means that a communicative intention and/or emotional state will have different impacts on the dynamiclines of two agents having different baselines. For example, if an agent has a general tendency (baseline) to perform movements with average speed/amplitude and to use hand gestures moderately then in a sad state it will do very few hand gestures with very low amplitude and speed. On the other hand, an agent with a general tendency of gesturing a lot with fast and large movements even when being sad, will continue doing gestures although less and with less expressivity (average speed and amplitude).

Example. Let us consider the baseline on Figure 17a, and the qualifier defined in Figure 16. We also suppose that the current emotional state of the agent corresponds to the one defined in the qualifier, i.e. *hot anger*. So, the *dynamicline computation* module decides to apply the

behavior qualifier to the baseline of Figure 17a, by performing the operations specified in the qualifier. The result of this computation is the dynamicline shown on the left column of Figure 17b.

The parameters affected by the qualifier are highlighted in bold. For example, lines 3, 11, and 19 represent the *Overall activation* expressivity parameter for the three modalities, and they have been obtained from the values in the baseline by multiplying them by a factor of 1.5, as defined in the qualifier (Figure 16). Instead, the gesture expressivity parameters in lines 20 - 22 are not related to the baseline, as they are explicitly determined by the qualifier.



Figure 17. Example of a) baseline and b) dynamicline.

Once the expressivity parameters and modalities of emotional expressions are changed according to the dynamicline the agent displays them.

3.4.5 Laughter production

The experiments conducted during the eINTERFACE workshop (see Section 4.4) strengthened our conviction that agents' expressivity is in certain situations considerably limited by the absence of laughter. In consequence, we began investigating methods to solve this problem.

We have acquired two databases containing a significant number of laughter episodes, recorded in different situations:

- The ICSI Meeting Corpus, recorded by the International Computer Science Institute (ICSI) of Berkeley. This database gathers audio recordings from 75 meetings occurring "naturally" (not influenced by the database recording project) (Janin et al., 2003). Each meeting participant wore a head-mounted microphone and 4 additional microphones were placed in the meeting room. Meetings involved from 3 to 10 participants, with an average of 6 (Janin et al., 2004). There were 53 participants in total. The resulting 85 hours of speech were fully transcribed and nonverbal events such as laughter were also annotated. The corpus contains 11515 segments marked as "laugh", 980 as "speech-laugh", 970 as "breath-laugh" and numerous less frequently occurring verbal episodes (Laskowski and Burger, 2007).
- We were able to obtain laughters recorded for the artistic installation "The world starts every second" (Lafontaine and Todoroff, 2007). The corpus gathers laughters from children and professional singers. Some singers portrayed different states of

mind like “lover laugh”, “hysteric laugh”, “obsessional laugh”, etc. The recordings took place in traditional recording rooms or in places with lower acoustics (echo, etc.). There are also several occurrences of group laughs. The laughter episodes from this database are typically long (over 20s) and some are obviously exaggerated, corresponding to stereotypes of what we consider as “free laughter”. They do not correspond to the large majority of natural laughter occurrences found in the ICSI Meeting Corpus, but they have a strong power of eliciting laughter to their listeners.

From the study of the state of the art, the analysis of the laughter databases and our desire to create a laughing Greta, the following questions arose:

- 1) How can a user find quickly a target utterance, corresponding to his imagination, in a large laughter corpus?

This issue is currently being investigated in a different context, in the framework of the Numediart research program (www.numediart.org) in which FPMs is involved. One project, called “Audio Cycle”, aims at providing the user a suited interface for browsing an audio database (containing audio segments used by DJs) and building an audio performance. A visual organization of the sounds is developed, based on similarity measures on different aspects of the sounds (rhythm, timbre, harmony), enabling a fast exploration of the database towards the desired effect. The user interface and sound analysis algorithms designed in this context could be adapted to a laughter corpus browsing,

- 2) How can the agent automatically answer to an input laughter with an appropriate laughter?

Extending the “Audio Cycle” concept, we will build a web-based application that will take the user laughter signal as input and find, thanks to suited similarity measures, the most appropriate answering episode in the database. The output laughter will possibly enhance the user’s initial laughter, in which case the system would bounce to another utterance in the database and so forth, creating a laughter loop. This will be the topic of another Numediart project, “Laughter Cycle”, that will be launched in April 2009 and from which results are expected in July 2009. This large-scale web approach will also enable us to enrich our laughter corpus with new utterances, while the underlying analysis system will give Greta the ability to laugh consistently with the user, as was desired in the case of the eNTERFACE’08 scenario (Section 4.4).

- 3) How can the resulting laughter be produced in an audio-visual way?

This question is still widely open and will be investigated when the first “audio only” produced laughters will be available.

4. Attentive agent

In this section we describe a first version of the component that adds the basic awareness of the user's behavior to the ECA. The ECA, in order to be believable, needs not only to display accurately its communicative intentions (see Section 3) but also to be aware of the user's behaviors. Different aspects of the awareness are related to different senses (vision, hearing, touch,...) At the moment we focus only on visual aspect and on the behavior of the user. We plan to add the notion of acoustic awareness in the next year of CALLAS project.

In the Section 4.1 we present the current version of the component that models the user's attention and interest using the information about her gaze and head movements. Then, in the Section 4.2, we describe an application that tracks the user's behavior to generate backchannels signals.

4.1 The gaze awareness module

The gaze awareness module represents an integration of previous work spanning a number of different domains (Asteriadis et al., 2007; Peters, 2006) inside CALLAS project. We focus on shared attention as it relates to interest in the other interactant, the scene, and particularly, the interaction itself, as signaled by gaze motions and gaze following. The agent attempts to track the state of the interaction, based on its interest and the theorized interest of the user. This will enable the agent to decide, for example, to halt ongoing behavior if the user is not interested, or explain an object in detail if the user is paying a lot of attention to it. Our aim is to outline the most important interconnected components, capabilities and metrics that will form the basis of the system to be used for a set of experiments investigating shared attention and engagement between a user and agent.

In order to build an attentive agent we use the component developed at ICCS (Asteriadis et al., 2007) within the CALLAS project (see the deliverable D122 for details). It employs facial feature analysis of images captured from a standard web-camera to determine the direction of the user's gaze, head movements and head poses. This information is then processed in a number of interpretation stages relating to user interest, allowing the agent to assess the state of the interaction and conduct shared-attention behaviors.

4.1.1 *State of art*

A number of researchers have considered eye gaze for HCI, either to communicate through a robot or computer with other humans or with virtual agents. Vertegaal et al. (Vertegaal et al., 2003) considered the significance of gaze and eye contact in the design of GAZE-2, a video conferencing system that ensures parallax-free transmission of eye-contact during multiparty mediated conversation. In work using conversational agents, some approaches have cast the ECA in the primarily role of a listener, for example, as a sensitive artificial listener (SAL) (Bevacqua et al., 2008), that provides feedback to a discourse conducted primarily by the user. In a similar vein to the current work, attentive presentation agents (Prendinger et al., 2007) rely on the eye gaze of the user to infer attention and visual interest, based on an algorithm presented in (Qvarfordt and Zhai, 2005) in order to alter their ongoing behavior in real-time.

4.1.2 *Modeling attention and interest*

The raw information about the user's head and eye directions obtained from the Detector Module is converted into 2D coordinates used to reference the virtual scene. There are two basic possibilities: the user is either looking inside or outside the screen area containing the 3D scene. In this work, we are not only concerned about where the user's gaze lands inside of the screen area, but also where it lands outside, as it can be an indicator of lack of

attention. In order to facilitate both of these possibilities, at the beginning of each interaction scenario with the user, a calibration process is invoked in order to find the corresponding maximum and minimum extents of the screen boundary in terms of raw head and eye direction values. After the conversion, the final coordinate data structure consists of a flag signaling gaze inside or outside the screen, accompanied by a 2D coordinate. If the flag indicates gaze within the scene boundary, the 2D coordinates correspond to the (x,y) screen position with respect to these boundaries. Otherwise, the 2D coordinate signals the screen boundary edge or corner that gaze fell outside.

The screen coordinates obtained from the Detector Module are used to compute the nearest virtual object falling under that gaze position. The attention that the user may have in particular objects, in the scene as a whole and/or in the interaction, is an important issue in this work. While the Detector Module detects the user's gaze direction (i.e. the eye/head direction of the user mapped into scene coordinates), this information must be converted into knowledge of what they are looking at for use in interaction understanding. Temporal integration is an important concept here: if at one time instant, the system detects that the user is looking outside of the screen, this does not necessarily imply that they are uninterested in what is happening - they may simply be glancing momentarily towards a distraction in their environment or staring upwards to think about what the ECA is saying. To aid in assessing the detection of the users attentive behaviors over different time-frames, we define a number of metrics.

Directness and level of attention

We use a directness metric to refer to the momentary orientating of the user eyes and head with respect to an area on the screen and record the ratio between them. For example, the user may have their head rotated directly towards an object in order to look at it - this would be considered a high degree of directedness. On the other hand, the user may have their head turned to the side, but be looking back at the object with their eyes - this would be considered a lower degree of directedness. Since metrics based on user gaze configuration during a single frame are highly unreliable indicators of attention, we define a level of attention metric. It refers to a clustering of a user's focus of interest in a single region over multiple frames.

Virtual attention objects

In order to simplify the analysis of what is being looked at in the scene, in a methodology similar to (Prendinger et al., 2007), we define virtual attention objects, or VAO's. A single VAO is attached to each object for which we wish to accumulate attention information - for example, one VAO is defined for the agent, one for each scene object, one for the scene background, and one to represent the area outside of the screen. If the screen-coordinate of the gaze fixation is located inside a VAO, then its corresponding level of attention is updated to reflect this. Thus, as the users gaze moves around the screen, each VAO maintains a history of how much and when the user has fixated it. The agent has access to the information of all VAOs in the scene. Since the agent is itself a VAO, it therefore has a full assessment of the users gaze around the scene and attention to specific objects.

Level of interest

Over a larger time-frame, and for a specific set of VAOs, the user's level of interest, *LIU*, for that set can be computed based on the stored attention levels for each member of the specified set. The possible values for LIU are: *low*, *medium*, and *high*. By defining a set of VAOs that contains only those objects currently relevant to the interaction, such as a recently pointed to or discussed object, and comparing the attention paid to these objects with the rest of the scene, we can obtain a measurement of how interested or engaged the user is with respect to the interaction itself, rather than superficial scene details.

Parameters for agent behavior generation

In addition to the metrics used for interpretation of the user's attention and interest, a level of interest is defined for the agent, LIA. Unlike the LIU, which is based on the user's detected behavior, the LIA helps define how the agent should generate its behavior. The LIA is determined by the agent's motivation in interacting.

4.1.3 Implementation

In practice, our system is comprised of two key modules: the Detector Module and the Shared Attention Player Module. These modules communicate via a Psyclone connection - a blackboard system for use in creating large, multi-modal A.I. systems (see Section 2). At the moment, the data about the user's gaze and head movements is used. Other the information about the user like posture could be used if suitable input components are available. The Detector Module does facial feature analysis of images captured from a standard web-camera in order to determine the direction of the users gaze. The Shared Attention Player Module is based on Greta agent (see Section 2). It contains the graphical representation of the agent and the scene, and receives updates of the users gaze from the Detector Module. It implements the agent interpretative capacities (metrics described in Sections 4.2).

In the first version of our module, the agent stands behind a table containing a number of simple objects, represented in this case by the rectangles. The user can see on the screen the cursor that indicates the simulation of his gaze attention. The level of interest for each object of the scenes (and the agent) is calculated in realtime and displayed on the screen in a form of the graph.

4.2 Proof of concept: a comparison between human-Greta and human-robot interactions

During an eINTERFACE'08 project (<http://enterface08.limsi.fr/project/7>) involving CALLAS partners Par8 and FPMs, a device was developed for real-time generation of backchannels when a user is telling a story. The backchannels that were considered in this proof of concept were related to the engagement the agent/robot had in regard to the interaction with the user. Engagement has been defined by C Sidner as "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake". It can also be defined as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction". In this particular study we looked mainly at three components of engagement, namely (dis)interest, (mis)understanding, (dis)like. That is the agent when interacting with the user can display signals (such as smile, head nod and shake) to show how interested or not it is in the conversation, how does it understand or not what is being said and if it likes it or not. The system is able to display various nonverbal backchannels like head nods or smiles. The backchannel signals generated by this system are displayed both by our 3D agent and by an Aibo robot. The interactions with Greta and Aibo were compared. We give hereunder more details about the data we used for this integration work, the architecture of the system and the obtained results.

4.2.1 Data acquisition

In order to model the interaction between the speaker and the listener during a storytelling experiment, we first recorded and annotated a database of human-human interaction: the eINTERFACE08_STEAD database. This database was used for extracting backchannel rules.

We followed the McNeill lab framework (McNeill, 1992): one participant (the speaker), who has previously watched an animated cartoon (Sylvester and Tweety), tells the story to a listener immediately after viewing it. The narration is accompanied by spontaneous communicative signals (filled pauses, gestures, facial expressions, etc.). In contrast, instructions are given to the listener to express his/her interest in the story by giving

nonverbal audio-visual signals in response to the story told by the speaker.

Twenty-two storytelling sessions telling the “Tweety and Sylvester - Canary row” cartoon story were recorded. Thirteen recording sessions were done by a French listener and a French speaker. The last two recordings have exaggerated nonverbal activity (closer to acting than to real-life storytelling). Four recording sessions were done by an Arabic listener and an Arabic speaker. Five recording sessions were done by a speaker and a listener who do not speak or understand each other’s languages; these recordings can be used to study the isolated effect of prosody on the engagement in a storytelling context. The languages used in these sessions were Arabic, Slovak, Turkish, and French.

The videos were annotated (with at least two annotators per session) for describing simple communicative signals of both speaker and listener: smile, head nod, head shake, eyebrows raising or frowning and acoustic prominence. These annotations were done using the ANVIL (Kipp, 2001) annotation program.

The eINTERFACE08_STEAD corpus contents and all the annotations are released under an MIT-like free software license.

4.2.2 System architecture

From the eINTERFACE08_STEAD database, a set of backchannel rules was established. Each rule contains an input signal and a corresponding output with a probability of emission. For example, one rule can be: when the speaker does a *head_nod*, the listener answers with a *head_nod* with a probability of 0.6. The rules can be triggered by more than one single input: for example, a particular backchannel can be generated when the speaker smiles and does a head nod simultaneously. These rules were used to animate Greta while somebody is telling her a story. A Sony Aibo robot was commanded the same way.

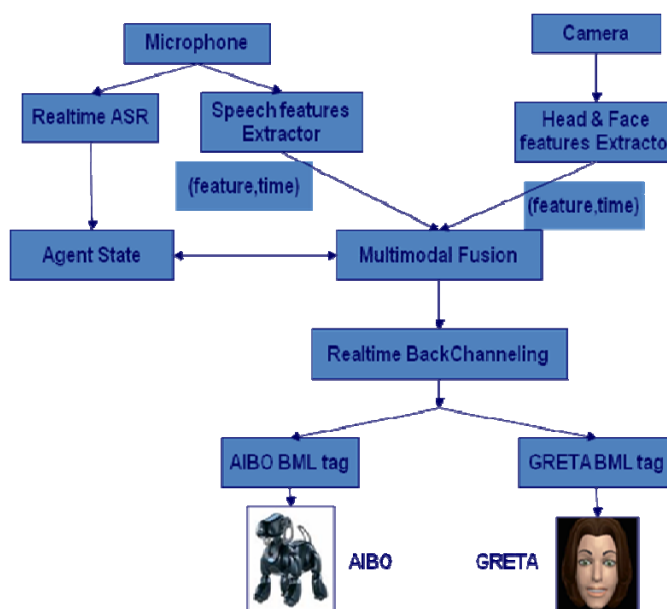


Figure 18: Engagement system architecture

The system requires real-time audio and visual analysis of the speaker’s attitude to detect the input events (audio prominence and head movements). The global architecture of the device, based on the one proposed by Bevacqua et al. (Bevacqua et al., 2008), is represented in Figure 18.

Speech features extraction involved estimation of the boundaries of the utterances through a Voice Activity Detection (VAD) and the evaluation of pitch prominence. This was performed through a statistical modeling of the recent pitch values and detection of outstanding figures. These features were directly used to generate the backchannels.

Real-time Automatic Speech Recognition was also considered to modify the agent state (interested, understanding, liking), which has an influence on the generated backchannels.

From video, characteristic feature points were extracted and tracked. Smiles were detected by measuring the space formed by the lips. Head nods (shakes) were estimated by computing the average vertical (horizontal) displacements of the head features between following frames. Head activity was also measured through the displacements of the feature points.

Multi-modal fusion was performed to produce backchannels suited to all the received information. The goal was to generate backchannels based on the combination of features (for example, prominence + head_nod). The fused information was sent to the BackChanneling module, which mapped the inputs to corresponding outputs thanks to the established rules. Past backchannels were taken into account to avoid generating an unrealistic sequence of backchannels.

Finally, the backchannel commands were transcribed in BML to interact with Greta or Aibo. Aibo did not produce “human-like” movements like smiles but attitudes adapted to a pet, like moving the tail, etc. Greta and AIBO were commanded simultaneously with the same instructions, only the interpretation of the backchannel to generate was different.

4.2.3 Project outcomes

Since the goal of this eINTERFACE project was to compare backchannel provided by two types of embodiments (a virtual character and a robot) rather than to evaluate the multimodal backchannel rules implemented in each of these systems, we decided to have users tell a story to both Greta and Aibo at the same time. They had then to report on how they perceived the agent/robot’s level of engagement in the interaction.

An instruction form was provided to the subject before the session. Then users watched the cartoon sequence, and were asked to tell the story to both Aibo and Greta. Finally, users had to answer a questionnaire. The questionnaire was designed to compare both systems with respect to the realization of feedback (general comparison between the two listeners, evaluation of feedback quality, perception of feedback signals and general comments). The experiment involved 10 participants and their questionnaire answers are summarized in Table 1. Aibo was judged more interested in the story and liking it more than Greta, but Greta was estimated as better understanding the story.

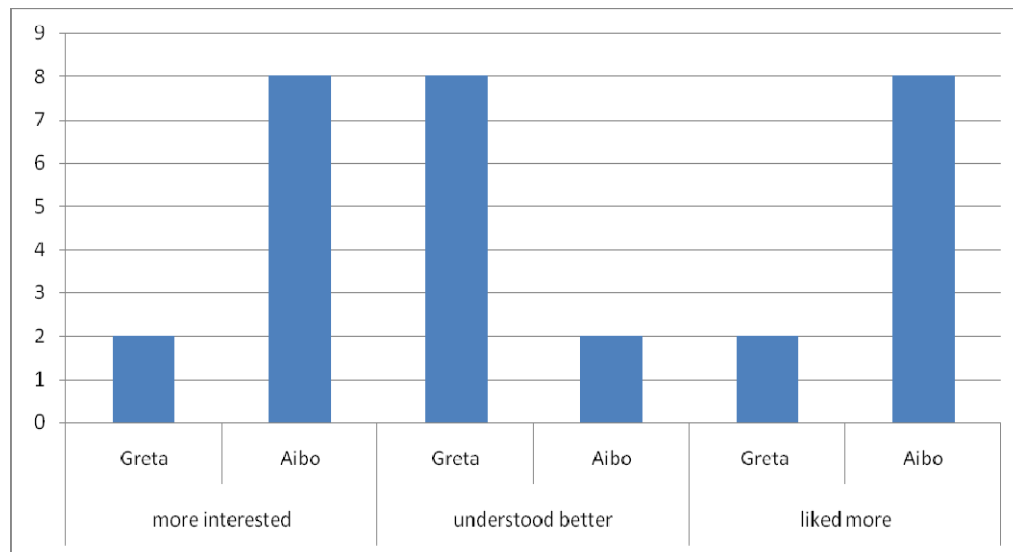


Table 1: Comparing the feedback given by the Great and Aibo

5. Affective and cognitive theory of mind for attentive agents

5.1 Background

Having a virtual companion that stays for a long period of time with a user and that learns and knows about the preferences and wishes of its owner continuously, necessitates a cognitive model for emotional intelligence. Such model should be capable to detect users' affective state and its current focus of attention in real-time (Gratch et al., 2007). Designing a cognitive model for virtual agents that consider users' real-time input is challenging. Not only that any human acts in a different way, but also the recognition rates of users' affect and attention are still not on a very reliable level and errors must be taken into account in such a cognitive model. Most current embodied conversational agents stand out with their capability of affective display (i.e. the display of feelings and emotions through facial expressions, gaze, hand gestures or voice). What they miss is emotional intelligence, that allows them to observe, to estimate and to manage their and the others emotions - an affective theory of mind (ToM). A model for theory of mind is necessary to give such agents cognitive capabilities about themselves and the others. Because most systems lack in responsive capabilities towards users, they are conspicuous by apathy (*ignoring how user feels*). We would like to see agents that are able to act and react with emotional intelligence, for instance showing empathy (*recognizing how user feels*), emotional contagion (*mirroring and feeling user's emotion*), sympathy (*recognizing how user feels and trying to help*) or pity (*recognizing that user needs help*). Deciding if virtual agents with the ability of empathic reasoning respond parallel or reactive is not trivial (McQuiggan et al., 2008). Mimicry (e.g. parallel empathy) is the capability to display the user's emotion in a similar manner to the user's current emotional expression. In contrast, reactive behavior aims to understand the user's affective state and tries to alter or enhance it. While a lot of work has been done in creation of the affective output for virtual characters, less work was done in combining the recognition of user's affective and attentive state with the affective display of current systems for embodied conversational agents.

5.2 State of art

Emotional sensitive virtual agents that interact with real humans need certain capabilities. Basically, they need skills that are similar and as demanding as in real human-human interaction. Such a virtual agent must be able to sense the signals from its environment, it must be able to interpret and understand these signals. Further, the virtual character must be able to react to it and to display an appropriate output. McQuiggan and Lester developed the framework CARE (McQuiggan and Lester, 2007), a data-driven affective architecture for learning empathy by observing human-human social interactions. It is used to generate empathic behavior (semantic, gestural, and postural) for virtual characters, which is derived from interactions performed by human controlled avatars within a virtual world. In a later study they investigate the users' awareness of parallel and reactive empathy performed by virtual characters in the CARE framework (McQuiggan et al., 2008). As users are able to percept empathic behavior from virtual characters, it still might be unclear, when such reaction is appropriate. Ochs and colleagues implemented an empathic dialog agent in a mail system to figure out, when affective feedback is at the proper place (Ochs et al., 2008). This application was meant for face to face communication and thus utilizes an embodied conversational agent for a human-computer interaction. Gratch and colleagues (Gratch et al., 2007) describe a system for rapport in human-machine dialogs. They detect speech and head orientation from the user to create continuing dialogs with their system. They do not focus on a system for affective feedback, but on right timings of feedback. Prendinger et al.

showed how data from the autonomic nervous system (ANS) can be used to derive users' emotional state in real-time for interaction with virtual characters (Prendinger et al., 2004). In a latter work these methods were used to display empathic behavior with an ECA (Becker et al., 2005). Boukricha (Boukricha, 2008) will extend this work to enable an ECA with a model of theory of mind for parallel and empathic behavior.

5.3 Theory of Mind

Theory of mind is the cognitive ability to understand what others intend to do or think. It enables us to interpret the counterpart's behavior. Furthermore, it allows us to assume or predict what our counterpart intends, desires, and believes. Such characteristic is essential for a virtual agent with emotional intelligence. As we are particularly interested in the interaction between real and virtual world (human-agent), a model that reflects the real and virtual world in an agent's mind is necessary. The attentive capability of our agent model will include both worlds. In contrast to the affective cognitive model, that will be limited to the real world and the user only, as our empathic listener is currently alone in its virtual world.

Children develop a theory of mind with 3-5 years. Typical tests for humans to detect the capability of theory of mind are the appearance-reality or false-belief task. A cognitive model that passes the latter task was implemented by Bringsjord (Bringsjord et al., 2008). As our virtual agents do not have to understand false-beliefs, we will simplify our theory of mind and split the cognitive ability of our virtual agent into two parts: (1) an affective theory of mind for mirroring users' emotions and (2) an cognitive theory of mind for being aware of users' emotions, which will allow us to react on users' emotions (Mehrabian and Epstein, 1972).

Affective Theory of Mind. Mimicry is the capability to response to another person's current emotional expression. This behavioral pattern is innate and the expression of emotional feedback is involuntarily. Such feedback behavior does not need a high level of cognitive capabilities (Hoffman, 2000).

Cognitive Theory of Mind. The process of role-taking is a more complex process that allows to understand how a user feels (Higgins, 1981), for instance by showing empathy (*recognizing how user feels*), sympathy (*recognizing how user feels and trying to help*) or pity (*recognizing that user needs help*) (Hogan, 1969), (Weinstein, 1969), (Ickes, 1997).

To react with pity or sympathy to the users' emotional state, our system needs a higher level of cognitive processing. The virtual agent must understand in what emotional state the user is to react in an appropriate way. So it is necessary for us, when the input components detect e.g. sadness from the user to define that the emotional model of our virtual character moves to something appropriate to 'pity for'. This allows us to let the virtual character display the correct emotion.

5.4 Approach for affective interaction

Our objective is the creation of an empathic listening agent that responds to the user's emotive and attentive state. To realize such an agent, we have been experimenting with several metaphors, such as that of a virtual pet or that of a virtual butler. While the virtual pet is not able to verbally respond to the user's state, the virtual butler gives both verbal as well as nonverbal feedback. Imagine the user has a rather bad day and is talking to the agent with a depressed voice. In the case of parallel empathy, the agent would simply mimicry the user's emotive state and show depression as well. In contrast to that, reactive empathy requires the agent to decide which emotion to display as a response to the user's emotive state. Here, the agent's emotions do not necessarily coincide with the user's emotions. That is the agent might, for example, decide to cheer the user up by showing encouragement. The agent is also able to provide simple verbal feedback, such as "I know how you feel!" or "That is really awful!". The user's eye gaze is analyzed in order to detect his or her interest in maintaining the conversation. Furthermore, the agent follows the user's eye gaze in order to show

attentive awareness.

We plan to combine components for affective and attentive sensory input with a cognitive model for our empathic listener that allows acting or reacting to the users' feelings. Another important part will be appropriate feedback. Although the feedback as a listener is limited in a way, timing and understanding the user still is crucial. Our approach will detect users' emotions from voice via the tool EmoVoice (Vogt et al., 2008) and users' focus of attention via an eye tracker. Further we will use a realistic virtual character with highly expressive facial emotions.

The agent architecture allows us to derive users' affective states with components that use for instance machine learning to recognize discrete emotional values (i.e. joy, sadness, ...) dependent on the training of the classifier. And parallel, we can use components that map their sensor data directly to the PAD (pleasure-arousal-dominance) model. This architecture provides currently components to sense emotional states from the user using a microphone and eye tracker, a component to process affective states for mimicry or role taking, and a component to display affect with virtual characters.

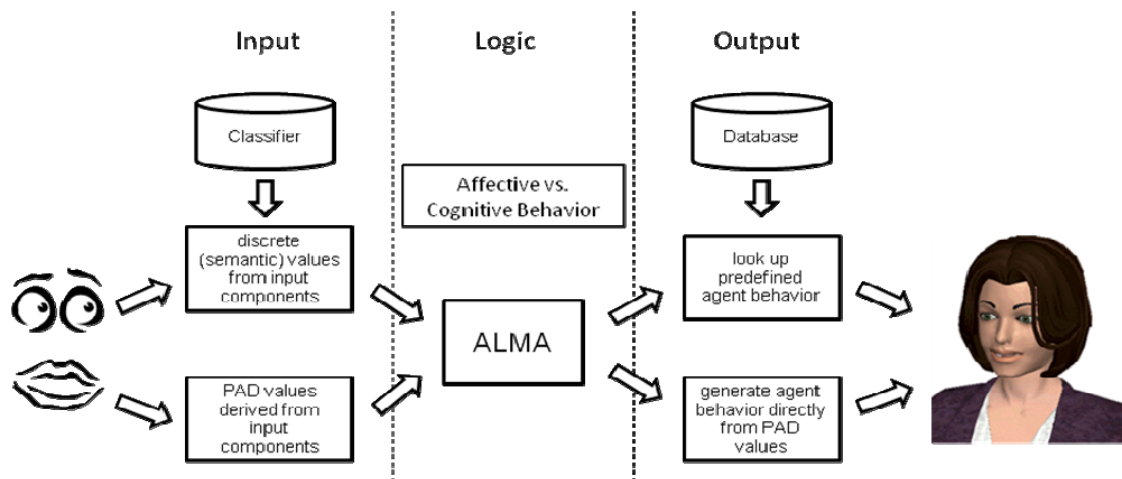


Figure 19. Agent framework for processing discrete and direct signal values of sensors for detecting user's emotional state and reacting with affective or cognitive behavior.

5.5 Mimicry vs. role-taking

5.5.1 Model for mimicry

For mimicry (empathy or emotional contagion) our approach (see Figure 19) does not need a complex model to understand and interpret users' emotions. It is sufficient to map users' recognized emotions directly to the affective display of a virtual character. We created an emotional model to represent the user's emotion in our approach. We simply take the current state in the model and display it to mirror the user's emotional state. This model allows setting the PAD values directly in the emotional model. Thus, our input components do not have to detect emotions, but have to map the signal to the PAD values directly. The same mapping between PAD values and the facial animation system is needed to display emotional expression with the virtual character. Mehrabian describes how to map variables in general to the PAD model (Mehrabian, 1995). The challenge with the sensor components will be to find an appropriate mapping between sensor data and the PAD value.

5.5.2 Model for role-taking

Role-taking (sympathy or pity) is a more complex process within our approach. Our system must understand the user's current situation. For instance, if the user feels sad, how should the system react? The semantic result of the emotion recognition (i.e. EmoVoice recognizes emotions dependent on how it was trained before; if you train the emotional categories joy and anger, EmoVoice will deliver either joy or anger as result dependent on users input.) must be understood and a reaction, dependent on the agents attitude towards the current user, ought to be generated. While mapping the user's input data directly to the PAD model is analogous to the model of mimicry, the interpretation of these values for generating emotional output behavior becomes challenging. With a model for role-taking it is not possible to use the emotion model for input and output, as the output behavior might differ from what is recognized from the user.

One approach to get a cognitive model of emotions for a human-computer interaction would be to create two emotion models, one for the user and one for the virtual character. Dependent on the emotional state of the user the emotion model of the virtual character should be adjusted. For instance, if the user feels sad and the input components recognize sadness, the emotion model of the user is set to 'sad', while the emotion model of the virtual character is set to 'pity', because a rule of the cognitive model for the virtual character is set to 'I like the user'. The PAD space is structured in eight octants (i.e. exuberant vs. bored, dependent vs. disdainful, relaxed vs. anxious, and docile vs. hostile) (Mehrabian, 1996). The PAD values for sadness can be found in the 'hostile' octant, while the PAD values for pity can be found in the 'docile' octant, the opposite octant in the PAD space. The rule for adjusting the emotional model of the virtual character in this case would be to set the opposite PAD value in regard to the user's current state.

References

1. I. Albrecht, M. Schröder, J. Haber, H. Seidel. Mixed feelings: expression of non-basic emotions in a muscle-based talking head. *Virtual Reality* 8 (4), pages 201-212, 2005.
2. J. Allwood, J.Nivre, and E.Ahlsén. On the semantics and pragmatics of linguistic feedback. *Semantics*, 9(1), 1993.
3. B.P. Arnason and A. Porsteinsson. The CADIA BML Realizer. <http://cadia.ru.is/projects/bmlr/>, 2008.
4. S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias. Non-verbal feedback on user interest based on gaze direction and head pose. In: 2nd International Workshop on Semantic Media Adaptation and Personalization (SMAP 2007), London, United Kingdom, December, 2007.
5. C. Becker, H. Prendinger, M. Ishizuka, and I. Wachsmuth. Empathy for Max. In *Proceedings of the 2005 International Conference on Active Media Technology (AMT 2005)*, pages 541-545, 2005.
6. E. Bevacqua, D. Heylen, M. Tellier, and C. Pelachaud. Facial feedback signals for ECAs. In: *AISB'07 Annual convention, workshop "Mindful Environments"*, pages 147 - 153, Newcastle, UK, April 2007.
7. E. Bevacqua, M. Mancini, and C. Pelachaud. A listening agent exhibiting variable behavior. In: H. Prendinger, J.C. Lester, and M. Ishizuka, editors, *Proceedings of 8th International Conference on Intelligent Virtual Agents*, volume 5208 of *Lecture Notes in Computer Science*, Springer, pages 262-269, Tokyo, Japan, 2008.
8. H. Boukricha. A first approach for simulating affective theory of mind through mimicry and role-taking. In *The Third International Conference on Cognitive Science, Symposium: Emotional Computer Systems and Interfaces*, 2008.
9. S. Brave, C. Nass, K. Hutchinson. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*. 62(2): 161-178, 2005.
10. S. Bringsjord, A. Shilliday, M. Clark, D. Werner, J. Taylor, A. Bringsjord, and E. Charpentie. Toward logic-based cognitively robust synthetic characters in digital environments. In *Proceedings of the First Conference on Artificial General Intelligence (AGI-08)*, 2008.
11. T.D. Bui. Creating emotions and facial expressions for embodied agents. Ph.D. thesis, University of Twente, Department of Computer Science, 2004.
12. B. De Carolis, C. Pelachaud, I. Poggi, and M. Steedman. APMML, a mark-up language for believable behavior generation. In: H. Prendinger and M. Ishizuka, editors, *Lifelike Characters. Tools, Affective Functions and Applications*. Springer, 2004.
13. J. Cassell. BEAT: The behavior expression animation toolkit. In: *SIGGRAPH '01: Proceedings of the 28th annual conference on Computer Graphics and Interactive Techniques*, ACM Press, pages 477-486, 2001.
14. J. Cassell and T. Bickmore. Embodiment in conversational interfaces: Rea. In: *Conference on Human Factors in Computing Systems*, Pittsburgh, PA, 1999.
15. A. Cerekovic, H.-H. Huang, G. Zoric, K. Tarasenko, V. Levacic, I.S. Pandzic, Y.I. Nakano, and T. Nishida. Towards an embodied conversational agent talking in Croatian. In: *The 9th International Conference on Telecommunications (ConTEL 2007)*, Zagreb, Croatia, 2007.
16. R. Cowie, E. Douglas-Cowie, B. Appolloni, J. Taylor, A. Romano, W. Fellenz, 1999. What a neural net needs to know about emotion words. In: N. Mastorakis (ed.), *Computational Intelligence and Applications*. World Scientific & Engineering Society Press, pages 109-114, 1999.
17. E. Douglas-Cowie, N. Campbell, P. Roach. Emotional speech: Towards a new generation of databases. *Speech Communication* 40(1-2), pages 33-60, 2003.
18. C. Darwin. *The expression of the emotions in man and animals*, 1872/1998.
19. G.B.A. Duchenne. *The mechanism of human facial expression* (R. A. Cuthbertson, Ed. & Trans.). Cambridge, England: Cambridge University Press, 1876/1999.
20. A. Egges, S. Kshirsagar, N. Magnenat-Thalmann. Imparting individuality to virtual humans. In: *First International Workshop on Virtual Reality Rehabilitation (Mental Health, Neurological, Physical, Vocational)*, Lausanne, Switzerland, pages 201-208, 2002.
21. N. Eisenberg, R.A. Fabes, P.A. Miller, J. Fultz, R. Shell, R.M. Mathy, R.R. Reno. Relation of sympathy and personal distress to prosocial behavior: a multimethod study, *Journal of personality and social psychology*, vol. 57, no 1, pages 55-66, 1989.

22. P. Ekman. Universals and cultural differences in facial expression of emotion. In J. R. Cole (Ed.), Nebraska symposium on motivation. Lincoln, NE: University of Nebraska Press, pages 207-283, 1972.
23. P. Ekman. Unmasking the Face. A guide to recognizing emotions from facial clues. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1975.
24. P. Ekman. Telling lies: Clues to deceit in the marketplace, politics, and marriage. W.W. Norton & Company, 1985.
25. P. Ekman. Basic emotions. In: T. Dalgleish and M. Power (eds.). Handbook of Cognition and Emotion. Sussex, U.K.: John, 1999.
26. P. Ekman. Darwin, masking, and facial expression. Ann. N.Y. Acad. Sci. 1000, pages 205-221, 2003a.
27. P. Ekman. The Face Revealed. Weidenfeld & Nicolson, London, 2003b.
28. P. Ekman, W. Friesen. The repertoire of nonverbal behavior's: Categories, origins, usage and coding. Semiotica 1, pages 49-98, 1969.
29. P. Ekman, W.V. Friesen. Pictures of facial affect. Consulting psychologists Press, Palo Alto, CA, 1976.
30. P. Ekman, W. Friesen. Felt, false, miserable smiles. Journal of Nonverbal Behavior 6 (4), pages 238-251, 1982.
31. The EULER project. <http://tcts.fpms.ac.be/synthesis/euler/home.html>.
32. J.M. Fernández-Dols. Facial expression and emotion: A situationist view. In: P. Philippot, R.S. Feldman, & E.J. Coats (eds.) The social context of nonverbal behavior. New York: Cambridge University Press, pages 242-261, 1999.
33. The Festival speech synthesis system. <http://www.cstr.ed.ac.uk/projects/festival/>.
34. M.G. Frank, P. Ekman, W.V. Friesen. Behavioral Markers and Recognizability of the Smile of Enjoyment. P. Ekman, E.L. Rosenberg, (eds.), What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), Oxford University Press, 1995.
35. W.V. Friesen, P. Ekman. EMFACS-7. Unpublished manual, 1984.
36. P. Goldie. Compassion: a natural, moral emotion, final draft for "Die Moralitaet der Gefuehle", Special Issue of Deutsche Zeitschrift für Philosophie 4, S. A. Doering and V. Mayer (eds.), Berlin: Akademie, 4, pages 199-211, 2002.
37. P. Gosselin, G. Kirouac, F.Y. Dore. Components and Recognition of Facial Expression in the Communication of Emotion by Actors. In: P. Ekman, E.L. Rosenberg, E.L. (eds.). What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), Oxford University Press, pages 243-267, 1995.
38. J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating rapport with virtual agents. In Intelligent Virtual Agents (IVA 2007), pages 125-138, 2007.
39. J. Haidt, D. Keltner. Culture and Facial Expression: Open-ended Methods Find More Expressions and a Gradient of Recognition, Cognition and Emotion, Volume 13, Number 3, 1, pages 225-266, 1999.
40. J. A. Harrigan, D.M. O'Connell. How do you look when feeling anxious? Facial displays of anxiety. Personality and Individual Differences, Volume 21, Number 2, pages 205-212, 1996.
41. B. Hartmann, M. Mancini, and C. Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In: The 6th International Workshop on Gesture in Human-Computer Interaction and Simulation, VALORIA, University of Bretagne Sud, France, 2005.
42. D. Heylen, S. Kopp, S. Marsella, C. Pelachaud, H. Vilhjálmsón. Why Conversational Agents do what they do? Functional Representations for Generating Conversational Agent Behavior. In: The First Functional Markup Language Workshop, The Seventh International Conference on Autonomous Agents and Multiagent Systems Estoril, Portugal, 2008.
43. D. Heylen, E. Bevacqua, M. Tellier, and C. Pelachaud. Searching for prototypical facial feedback signals. In: proceedings of 7th International Conference on Intelligent Virtual Agents IVA 2007, Paris, France, pages 147-153, 2007.
44. T. E. Higgins. Role taking and social judgment: alternative developmental perspectives and processes. In Flavell and Ross, editors, Social Cognitive Development, pages 119-153. Cambridge University Press, 1981.
45. M.L. Hoffman. Empathy and Moral Development: Implications for Caring and Justice. Cambridge University Press, 2000.
46. R. Hogan. Development of an empathy scale. Journal of consulting and clinical psychology, 33(3):307-316, 1969.

47. H.-H. Huang, T. Masuda, A. Cerekovic, K. Tarasenko, I.S. Pandzic, Y.I. Nakano, and T. Nishida. Toward a universal platform for integrating embodied conversational agent components. In B. Gabrys, R.J. Howlett, and L.C. Jain, (eds.), *Proceedings of 10th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, volume 4252 of *Lecture Notes in Computer Science*, Springer, pages 220-226, 2006.
48. W. Ickes. *Empathic Accuracy*. The Guilford Press, 1997.
49. C.E. Izard. *Human emotion*, New York: Plenum Press, 1977.
50. A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, B. Wrede. The ICSI Meeting Project: Resources and research. In: *NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal, May 2004.
51. A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters. The ICSI Meeting Corpus. In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong-Kong, April 2003.
52. S. Kaiser, T. Wehrle, T. Facial expressions as indicators of appraisal processes. In: K. R. Scherer, A. Schorr & T. Jonshstone (eds). *Appraisal processes in emotion: theory, methods, research*. New York and Oxford: Oxford University Press, 2001.
53. D. Keltner. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. In P. Ekman & E. L. Rosenberg (eds.), *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Series in affective science, New York: Oxford University Press, pages 133-160, 1995.
54. D. Keltner, B.N. Buswell. Evidence for the distinctness of embarrassment, shame, and guilt: A study of recalled antecedents and facial expressions of emotion. *Cognition and Emotion*, 1996.
55. D. Keltner, B.N. Buswell. Embarrassment: Its distinct form and appeasement functions. *Psychological Bulletin*, 122, pages 250-270, 1997.
56. D. Keltner, J. Haidt. Approaching awe, a moral, spiritual, and aesthetic emotion. *Cognition and Emotion*, 17, pages 297-314, 2003.
57. D. Keltner, M.N. Shiota. New displays and new emotions: A commentary on Rozin and Cohen. *Emotion*, 3, pages 86-91, 2003.
58. P.G. Kenny, T.D. Parsons, J. Gratch, and A.A. Rizzo. Evaluation of Justina: A Virtual Patient with PTSD. In: H. Prendinger, J.C. Lester, and M. Ishizuka, editors, *Proceedings of 8th International Conference on Intelligent Virtual Agents*, Tokyo, Japan, volume 5208 of *Lecture Notes in Computer Science*, Springer, pages 394 - 408, 2008.
59. M. Kipp. ANVIL A Generic Annotation Tool for Multimodal Dialogue. In: *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367-1370, 2001.
60. M. Kipp. Creativity Meets Automation: Combining Nonverbal Action Authoring with Rules and Machine Learning. In *Proceedings of 6th conference on Intelligent Virtual Agents*, vol. 4133 of *Lecture Notes in Computer Science*, Springer, pages 230-242, 2006.
61. S. Kopp, B. Jung, N. Lessmann, and I. Wachsmuth. Max - a multimodal assistant in virtual reality construction. *KI*, 17(4), pages 11-18, 2003.
62. S. Kopp, B. Krenn, S. Marsella, A.N. Marshall, C. Pelachaud, H. Pirker, K.R. Thorisson, H. Vilhjálmsón, N. Badler, and L. Johnson. *Behavior Markup Language*. Website, 2007.
63. M.-J. Lafontaine, T. Todoroff. The world starts every second. Artistic Installation held at the “Musée des Beaux-Arts”, Angers, France, December 2007.
64. J.L. Lakin, V.A. Jefferis, C.M. Cheng, and T.L. Chartrand. Chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Nonverbal Behavior*, 27(3), pages 145-162, 2003.
65. E. Lasarcyk and J. Trouvain. Imitating conversational laughter with an articulatory speech synthesis. In: *Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter*, pages 43-48, Saarbrücken, Germany, August 2007.
66. K. Laskowski, S. Burger. On the correlation between perceptual and contextual aspects of laughter in meetings. In: *Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter*, pages 55-60, Saarbrücken, Germany, August 2007.
67. J. Lee, S. Marsella. Nonverbal behavior generator for embodied conversational agents. In: J. Gratch, M. Young, R. Aylett, D. Ballin, and P. Olivier, editors, *Proceedings of 6th International Conference on Intelligent Virtual Agents*, Marina Del Rey, CA, USA, volume 4133 of *Lecture Notes in Computer Science*, Springer, pages 243-255, 2006.

68. R.M. Maatman, J. Gratch, and S. Marsella. Natural behavior of a listening agent. In T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, and T. Rist, editors, *Proceedings of 5th International Working Conference on Intelligent Virtual Agents*, Kos, Greece, volume 3661 of *Lecture Notes in Computer Science*, Springer, pages 25-36, 2005.
69. M. Mancini, R. Bresin and C. Pelachaud. A Virtual Head Driven by Music Expressivity. In: *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.
70. M. Mancini and C. Pelachaud. Dynamic behavior qualifiers for conversational agents. In: C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, editors, *Proceedings of 7th International Conference on Intelligent Virtual Agents*, volume 4722 of *Lecture Notes in Computer Science*, Springer, pages 112-124, Paris, France, 2007.
71. M. Mancini and C. Pelachaud. Distinctiveness in multimodal behaviors. In: L. Padgham, D.C. Parkes, J. Muller, and S. Parsons, editors, *Proceedings of Conference on Autonomous Agents and Multi-Agent Systems (AAMAS08)*, 2008.
72. A.S.R. Manstead, A.H. Fischer, E.B. Jakobs. The Social and Emotional Functions of Facial Displays. In: P. Philippot, R.S. Feldman, E.J. Coats (eds.), *The Social Context of Nonverbal Behavior (Studies in Emotion and Social Interaction)*, Cambridge University Press, pp. 287-316, 2005.
73. The MARY text-to-speech system. <http://mary.dfki.de/>.
74. S.W. McQuiggan and J. C. Lester. Modeling and evaluating empathy in embodied companion agents. *International Journal of Human-Computer Studies*, 65(4):348-360, April 2007.
75. S.W. McQuiggan, J. L. Robison, R. Phillips, and J. C. Lester. Modelling parallel and reactive empathy in virtual agents: An inductive approach. In *Proc. Of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, pages 167-174, 2008.
76. A. Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*, 121(3):339-361, 1995.
77. A. Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261-292, 1996.
78. A. Mehrabian and N. Epstein. A measure of emotional empathy. *Journal of Personality*, 40(4):525-543, 1972.
79. L. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In: *Proceedings of the 7th International Conference on Multimodal Interfaces*, ACM New York, NY, USA, pages 18-24, 2005.
80. R. Niewiadomski and C. Pelachaud. Model of facial expressions management for an embodied conversational agent. In: *Proceedings of the 2nd Conference on Affective Computing and Intelligent Interaction*, volume 4738 of *Lecture Notes in Computer Science*, Springer, pages 12-23, 2007a.
81. R. Niewiadomski and C. Pelachaud. Fuzzy similarity of facial expressions of embodied agents. In: C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, D. Pelé (eds.), *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, Springer, pages 86-98, 2007b.
82. M. Ochs, C. Pelachaud, and D. Sadek. An empathic virtual dialog agent to improve human-machine interaction. In *7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, pages 89-96, 2008.
83. J. Ostermann. Face animation in MPEG-4. In: I. Pandzic and R. Forchheimer, editors, *MPEG-4 Facial Animation - The Standard Implementation and Applications*, Wiley, England, pages 17-55, 2002.
84. X. Pan, M. Gillies, T.M. Sezgin and C. Loscos, Expressing Complex Mental States Through Facial Expressions, In *Proceedings of Second Affective Computing and Intelligent Interaction Conference ACII07*, *Lecture Notes in Computer Science Volume 4738/2007*, Springer, pages 745-746, 2007.
85. C. Peters. A perceptually-based theory of mind model for agent interaction initiation. In: *International Journal of Humanoid Robotics (IJHR)*, special issue *Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids*, pages 321-340, 2006.
86. C. Peters, S. Asteriadis, K. Karpouzis, and E. deSevin. Towards a real-time gaze-based shared attention for a virtual agent. In: *Workshop on Affective Interaction in Natural Environments (AFFINE)*, Crete, Greece, 2008.
87. C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi. A model of attention and interest using gaze behavior. In: T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, and T. Rist, editors, *Proceedings of 5th International Working Conference on Intelligent Virtual Agents*, Kos, Greece, volume 3661 of *Lecture Notes in Computer Science*, Springer-Verlag, pages 229-240, 2005.
88. R. Plutchik. *Emotions: A psychoevolutionary Synthesis*. Harper & Row, New York, 1980.
89. I. Poggi. *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*. Weidler, 2007.

90. F.E. Pollick, H. Paterson, A. Bruderlin, A.J. Sanford. Perceiving affect from arm movement. *Cognition*, 82, pages 51-61, 2001.
91. H. Prendinger, T. Eichner, E. André and M. Ishizuka. Gaze-based infotainment agents. *Advances in Computer Entertainment Technology*, pages 87-90, 2007.
92. H. Prendinger, H. Dohi, H. Wang, S. Mayer, and M. Ishizuka. Empathic embodied interfaces: Addressing users' affective state. In: *Proceedings Tutorial and Research Workshop on Affective Dialogue Systems, LNAI 3068*, pages 53-64, 2004.
93. P. Qvarfordt and S. Zhai. Conversing with the user based on eye-gaze patterns. In: *CHI'05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221-230, New York, NY, USA, 2005.
94. M. Rehm, A. André. Informing the design of embodied conversational agents by analysing multimodal politeness behaviors in human-human communication. In: *Workshop on Conversational Informatics for Supporting Social Intelligence and Interaction*, 2005.
95. J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum, and B. Swartout. Towards a new generation of virtual humans for interactive experiences. *IEEE Intelligent Systems*, pages 32-38, July/August 2002.
96. Z. Ruttkay. Constraint-based facial animation. *Int. Journal of Constraints*, Vol. 6, pages 85-113, 2001.
97. Z. Ruttkay, H. Noot, P. ten Hagen. Emotion disc and emotion squares: Tools to explore the facial expression face. *Computer Graphics Forum* 22 (1), pages 49-53, 2003.
98. B. Scassellati. Mechanisms of shared attention for a humanoid robot. In: *Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium*, AAAI, 1996.
99. E. Schegloff and H. Sacks. Opening up closings. Technical report, 1969.
100. K.R. Scherer, K. R. What does facial expression express? In: K. Strongman (ed.), *International review of studies on emotion*. Chichester, UK: Wiley. 2, pages 139-165, 1992.
101. K.R. Scherer, H. Ellgring. Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, 7, 1, pages 113-130, 2007.
102. L. Schilbach, A.M. Wohlschlaeger, N.C. Kraemer, A. Newen, N. Jon Shah, G.R. Fink and K. Vogeley. Being with virtual others: neural correlates of social interaction. *Neuropsychologia*, 44, pages 718-730, 2006.
103. M.N. Shiota, B. Campos, D. Keltner. The faces of positive emotion: prototype displays of awe, amusement, and pride. *Ann N Y Acad Sci*. 1000, pages 296-299, 2003.
104. C. Sidner, C. Kidd, C. Lee, and N. Lesh. Where to look: A study of human-robot interaction. In *Intelligent User Interfaces Conference*, ACM Press, pages 78-84, 2004.
105. D. Simon, K.D. Craig, F. Gosselin, P. Belin, P. Rainville. Recognition and discrimination of prototypical dynamic expressions of pain and emotions. *Pain* vol. 135, no 1-2, Elsevier, Amsterdam, pages 55-64, 2008.
106. S. Sundaram and S. Narayanan. Automatic acoustic synthesis of human-like laughter. *Journal of the Acoustical Society of America*, 121(1):527-535, January 2007.
107. M. Thiebaux, S. Marsella, A.N. Marshall, and M. Kallmann. SmartBody: behavior realization for embodied conversational agents. In L. Padgham, D.C. Parkes, J. Muller, and S. Parsons, editors, *Proceedings of Conference on Autonomous Agents and Multi-Agent Systems (AAMAS08)*, pages 151-158, 2008.
108. K.R. Thorisson, T. List, C. Pennock, and J. Dipirro. Whiteboards: Scheduling blackboards for semantic routing of messages & streams. In: *AAAI-05 Workshop on Modular Construction of Human-Like Intelligence*, pages 8-15, 2005.
109. N. Tsapatsoulis, A. Raouzaoui, S. Kollias, R. Crowie, E. Douglas-Cowie. Emotion recognition and synthesis based on MPEG-4 FAPs. In: Pandzic, I., Forchheimer, R. (eds.), *MPEG-4 Facial Animation - The standard, implementations, applications*. John Wiley & Sons, pages 141-168, 2002.
110. H.H. Vilhjálmsón, N. Cantelmo, J. Cassell, N.E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A.N. Marshall, C. Pelachaud, Z. Ruttkay, K.R. Thórisson, H. van Welbergen, and R.J. vander Werf. The Behavior Markup Language: Recent developments and challenges. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, editors, *Proceedings of 7th International Conference on Intelligent Virtual Agents*, Paris, France, volume 4722 of *Lecture Notes in Computer Science*, Springer, pages 99-111, 2007.
111. R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung. Gaze-2: conveying eye contact in group video conferencing using eye-controlled camera direction. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 521-528, 2003.
112. T. Vogt, E. André, and N. Bee. EmoVoice – a framework for online recognition of emotions from voice. In *Proceedings of Workshop on Perception and Interactive Technologies*, 2008.

- 113. H. G. Wallbott. Bodily expression of emotion. *European Journal of Social Psychology*, 28, pages 879-896, 1998.
- 114. N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 23, pages 1177-1207, 2000.
- 115. E. A. Weinstein. The development of interpersonal competence. In D. A. Goslin, editor, *Handbook of Socialization Theory and Research*. Rand McNally & Co, 1969.
- 116. C. Whissell. The dictionary of affect in language. In: R. Plutchik, H. Kellerman, H. (eds.), *Emotion: Theory, Research, and Experience*, volume 4: *The Measurement of Emotions*, Academic Press. Inc., SanDiego, pages 113-131, 1989.