Computational Model of Believable Conversational Agents

Catherine Pelachaud
LINC - Paragraphe
IUT of Montreuil - University of Paris 8
c.pelachaud@iut.univ-paris8.fr

Massimo Bilvi
Department of Computer and System Science
University of Rome "La Sapienza"
bilvi@dis.uniroma1.it

November 22, 2002

1 Abstract

In this chapter we present the problems and issues involved in the creation of Embodied Conversational Agents (ECAs). These agents may have a humanoid aspect and may be embedded in a user interface with the capacity to interact with the user; that is they are able to perceive and understand what the user is saying, but also to answer verbally and nonverbally to the user. ECAs are expected to interact with users as in human-human conversation. They should smile, raise their brows, nods, and even gesticulate, not in a random manner but in co-occurrence with their speech. Results from research in human-human communication are applied to human-ECA communication, or ECA-ECA communication. The creation of such agents requires several steps ranging from the creation of the geometry of the body and facial models to the modeling of their mind, emotion, and personality, but also to the computation of the facial expression, body gesture, gaze that accompany their speech. In this chapter we will present our work toward the computation of nonverbal behaviors accompanying speech.

2 Introduction

We convey our thoughts through our (conscious or unconscious) choice of words, facial expressions, body postures, gestures... Faces are an important mean of communication and may have several communicative functions. They are used to control the flow of conversation; that is they help in regulating the exchange of speaking turns, keeping the floor or asking for it. Actions such as smiling, raising the eyebrows, and wrinkling the nose often co-occur with a verbal message. Some facial expressions accompany the flow of speech and

are synchronized at the verbal level, punctuating accented phonemic segments and pauses. Other facial expressions may substitute for a word or string of words, or emphasize what is being said. They can also express attitude toward one own speech (such as irony) or toward the interlocutor (like showing submission). They are the primary channel to express emotion. Facial expressions do not occur randomly, but rather are synchronized to one's own speech, or to the speech of others [9, 16, 29].
Faces exhibit not only expressions of emotions but also a large variety of communicative functions that are essential to a conversation. To control the agent 'Greta' we are using the taxonomy of communicative behavior as proposed by Poggi [26]. This taxonomy is based on the type of information a behavior displayed by a speaker communicates to conversants, each class may be composed of several functions:

Information on the Speaker's belief : this cluster includes expressions that provide information on different types of speaker's beliefs:

- certainty function: the speaker may be certain or uncertain of what she is saying; she may respectively frown or raise her eyebrow to punctuate her attitude.
- belief-relation functions: the speaker may contrast several elements in her speech by raising her eyebrows.
- adjectival functions: the speaker may mimic the property of abstract('great idea') or object ('small box') by squeezing or opening wide the eyes.
- deictic functions: the speaker may gaze at a point in space to direct the conversant's attention.

Information on the Speaker's intention : this cluster gathers expression used to underline the particular intention of a speaker:

- performative function: the speaker may request (say by an order, a suggest or an implore), ask (interrogate), inform (by warning). In previous work we have exposed the link existing between performative and facial expressions [27].
- topic-comment function: the topic is the background information the speaker is taking for granted as being shared with the conversant, and the comment is the new information the speaker considers relevant and worth to communicate. The speaker may mark this new information by raised eyebrows and/or head nod.
- turn-taking function: this function refers to how people negotiate speaking turns in a conversation. Gaze plays a large role in the negotiation.

Information on the Speaker's affective : this cluster represents the expression of an emotion felt or referred to by the speaker.

Information on the Speaker's meta-cognitive : the expressions correspond to a particular thinking activity of the speaker (breaking the gaze while remembering a fact or planning what to say).

A communicative function is made of two components: a signal and a meaning. Signal may be a facial expression, a gaze/ head direction, or a head movement; while meaning corresponds to the communicative value of a signal. We have decided to cluster communicative functions not from the signals involved in the expression (e.g. raising eyebrows) but from their meanings. Indeed the same expression may change meaning depending on its place and time of occurrence in the conversation. Raising eyebrows signal surprise but also emphasis of what is being said; they signal question mark, specially in the case of non-syntactically questions but they are also part of the expression used when suggesting something to someone. A smile may be a sign of happiness but it may also be used as a greeting or a back-channel signs. Moreover not everybody uses the same expression to carry a given function. Some people mark accented words with, for example, eye flashes, other will raise their eyebrows, or nod their head. We believed one has to consider this variety of behaviors in the creation of believable ECAs.
The work presented in this chapter is part of a larger system developed within a European project, MagiCster[1]. The project aim at building a new type of human-computer interface, based on a Conversational Embodied Agent. It wishes to make this Agent 'believable and expressive': that is, able to communicate complex information through the combination and the tight synchronization of verbal and nonverbal signals. As application, the agent is embedded in user interface where it may dialog with a user or with other agent(s).
In the remaining of this chapter we describe how, given a text to be output by the agent (this text may have been generated by a dialog system [25]) and a set of communicative function, to compute the corresponding animation of the agent. We present our system architecture as well as each of its components in the next sections. In section 9 we describe in detail how we solve conflict at the facial expression levels while in sections 11 and 11.2 we define a description language for facial expressions.

3 Representation Language, APML

To ensure portability of the facial model, it is compliant with MPEG-4 standards. To ensure independence between the specification of the facial expressions and the facial models (that is we wish to be able to define facial expressions to be applied to any type of facial models) we define a set of tags using XML format.
To ensure the portability of the system as well as to ensure independence between the specification of the facial expressions and the facial models (that is we wish to be able to define facial expressions to be applied to any type of facial models) we are using an XML language, called Affective Presentation Markup Language (APML) [10]. The types of the tags represents the communicative functions as defined above. XML offers also a synchronisation scheme between the verbal and the nonverbal channels as it delineated the action of signals over text spans. An example of annotated text is:

[1]IST project IST-1999-29078, partners: University of Edinburgh, Division of Informatics; DFKI, Intelligent User Interfaces Department; Swedish Institute of Computer Science; University of Bari, Dipartimento di Informatica; University of Rome, Dipartimento di Informatica e Sistemistica; AvartarME

<APML>
<turn-allocation type="take turn">
<performative type="greet">
Good Morning, Angela.
</turn-allocation>
<affective type="happy">
It is so <topic-comment type="comment">**wonderful**</topic-comment> **to see you again.**
</affective>
<certainty type="certain"> **I was sure we would do so, one day!** </certainty>
</APML>

Figure 1: Example of XML input

4 Architecture

Our system takes as input a text marked with tags denoting the communicative functions. The tags are part of the APML representation language. The system interprets the input text by instantiating the communicative function into their corresponding facial expressions. The output of the system is a facial animation file and an audio file. Figure 2 illustrates the detailed architecture of our system, the *Greta* agent system, composed of several modules whose main functions are:

- **APML Parser**: XML parser that validates the input format as specified by the APML language.
- **Expr2Signal Converter**: given a communicative function and its meaning, this module returns the list of facial signals to activate for the realization of the facial expression.
- **TTS Festival**: manages the speech synthesis and give us the information needed for the synchronisation of the facial expressions to the speech (i.e. list of phonemes and phonemes duration).
- **Conflicts Resolver**: resolves the conflicts that may happened when more than one facial signals should be activated on the same facial parts (example: the co-occurring signals should be "eyebrow raising" and "frown" on the eyebrow region).
- **Face Generator**: converts the facial signals into MPEG-4 Facial Animation Parameters (FAPs) needed to animate the 3D facial model.
- **Viseme Generator**: converts each phoneme, given by Festival, into a set of FAP values needed for the lips animation.

- **MPEG4 FAP Decoder**: is an MPEG-4 compliant Facial Animation Engine.

5 APML Parser

The input to the agent engine in an XML string which contains the text to be pronounced by the agent enriched with XML-tags indicating the communicative functions that are attached to the text. The APML parser takes such an input and validates it with the DTD (Document Type Definition). The elements of the DTD correspond to the communicative functions described in section 2. The next step is to pass the text to be said (specified in bold in figure 1) to the speech synthesiser Festival [4] while the information contained in the markers is stored in a structure that will be used subsequently.

6 Speech Synthesizer - Festival

In the current version of the system we are using Festival as speech synthesizer [4]. Festival returns a list of couples $(phoneme, duration)$ for each phrase of APML tagged text. These information are then used to compute the lip movement and to synchronise the facial expression with speech.

7 Synchronisation of the Facial Expressions

Facial expressions and speech are tightly synchronised. In our system the synchronisation is implemented at the word level, that is, the timing of the facial expressions is connected to the text embedded between the markers. The XML parser returns a tree structure from which we calculate, using the list of the phonemes returned by Festival, the timings of each individual expression. The leaves of the tree correspond to the text while the intermediary nodes correspond to facial expressions except for the root that corresponds to the APML marker (see Figure 3).

7.1 Temporal course of an expression

Knowing the starting time and duration of an expression, the next step is to calculate the course of the expression intensity. The intensity of the expression is viewed as the amplitude of the facial movements, variable in the time, that compose the expression.
Each expression is characterised by three temporal parameters [12]:

- **onset**: is the time that, starting from the neutral face, the expression takes to reach its maximal intensity.
- **apex**: is the time during which the expression maintains its maximal intensity.

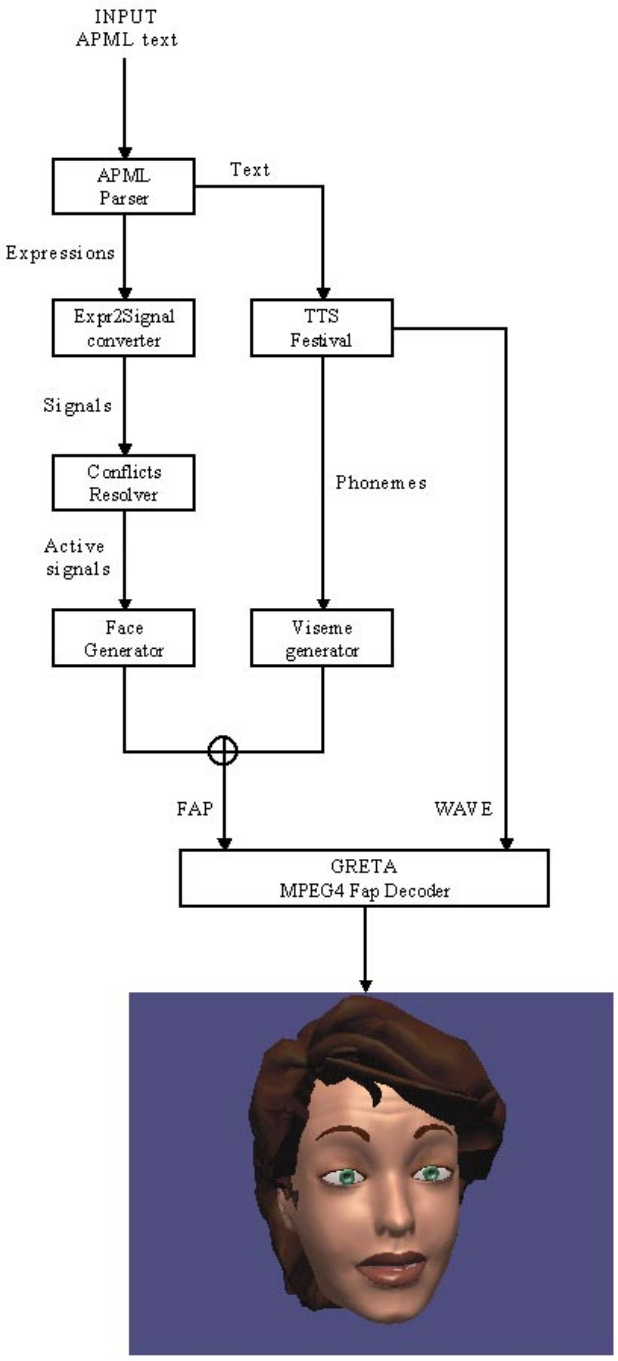

Figure 2: Agent Architecture

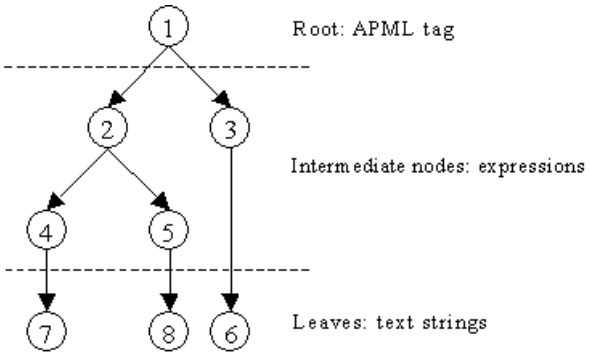

Figure 3: Tree structure from XML input

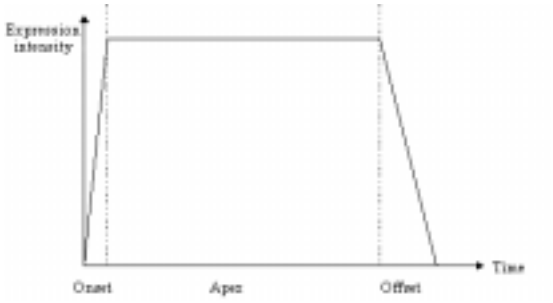

Figure 4: Temporal course of the expression " surprise" with its respective parameters *onset*, *apex* and *offset*

- **offset**: is the time that, starting from the maximal intensity, the expression takes to return to the neutral expression.

Such parameters are different from expression to expression. For example the "sadness" expression is characterised by a long *offset* (the expression takes more time to disappear), while the "surprise" expression has a short *onset*.
The values used for these parameters, have been taken from researches based on the analysis of facial expressions [14, 30, 3].
It has been showed experimentally that the amplitude of a facial movement is much more complex [14] than a simple decomposition in three linear parameters but for sake of simplicity and for lack of data, we use such trapezoidal functions to represent the temporal aspects of facial expressions.

Figure 5: 'Satisfaction' expression

8 Instantiation of the APML Tags - Expr2Signal Converter

The APML tags correspond to the meaning of a given communicative function. Thus, the next step is to convert the markers of an input text into their corresponding facial signals. The conversion is done by looking up the definition of each tag into the library that contained the pairs of the type (meaning, signals) .
Let us consider the following example:

<affective type="satisfaction">
I was sure we will arrive to an agreement.
</affective>

This text contains one communicative function represented by the marker *affective* which value is *satisfaction* as specified by the field *type*. The list of signals for this communicative function is:

$$affective(satisfaction) = \{raised\ eyebrows,\ smile,\ head\ nod\}$$

Figure 5 illustrates the corresponding expression.
Now, let us consider the following example:

<certainty type="certain">
I was sure we will arrive to an agreement.
</certainty>

Here, the communicative functions is given by the marker *certainty* with *certain* as a value. The list of signals for this function is:

$$certainty(certain) = \{frown\}$$

Figure 6: 'Certain' expression

Figure 6 illustrates the expression of *certain*.
In these two examples we have seen two "different" communicative functions that activate "different" signals on the same facial part (eyebrow).
Let us consider the following example:

<affective type="satisfaction">
<certainty type="certain">
I was sure we will arrive to an agreement.
</certainty>
</affective>

We have two communicative functions that activate in the same time interval two different signals (*frown* and *raised eyebrow*) on the same facial region (*eyebrow*). So we have a conflict that must solve before visualising the animation. When a conflict at the level of facial signals is detected, the system calls up a special module for the resolution of conflicts, the "conlict resolver" in figure 2 (described in details in the section 9). Such a module determines which signal, between those that should be active on the same facial region, must prevail on the others. If we go back to our previous example, *Conflicts Resolver* returns:

$$\text{resolve\_conflict}(affective(satisfaction), certainty(certain)) = \{frown\ ,\ smile,\ head\ nod\}$$

The resulting expression is shown in Figure 7. As we can see the *Conflicts Resolver* has decided that the signal *frown* prevails over the signal *raised eyebrows*.

9 Conflicts resolver

Few attempts have been made to combine co-occurring expressions. Often additive rules are applied [6, 24] that is all signals corresponding to the co-occurring communicative functions

Figure 7: Expression of 'satisfaction', 'certain' and combination of both expressions after conflict resolution.

are added to each other. Lately, Cassell et al [8] have proposed a hierarchical distinctions of the signal: only the signal with the highest priority rule will be displayed. These last two methods do not allow combination of several communicative functions to create a complex expression. Our proposal is to apply *Belief Networks* (BN) to the management of this problem. Our BN includes the following types of nodes (see figure 8):

communicative functions nodes correspond to the communicative functions: performative, certainty, belief-relation, emotion, topic-comment, turn-taking, meta-cognitive.

facial parts nodes are the eyes, eyebrows, mouth shape, head movement and head direction. For example, the values we count for the eyebrows are: raised, frown, oblique, and neutral. The values we consider for the mouth are: lip tense, lip corner up, lip corner down, and neutral.

performative dimensions : Performatives may be described along a small set of dimensions which are 'power-relationship', 'in whose interest is the requested action', 'degree of certainty', 'type of social encounter', 'affective state' [27]. We have singled out two dimensions among the five ones that are relevant in the characterisation of performatives [27]: 'power relationship' and 'in which interest is the requested action', that are called, respectively, in the BN 'dominance' (whose values are submissive, neutral, dominant) and 'orientation' (whose values are self-oriented, neutral, other-oriented). These dimensions allow us to differentiate performatives not as for their meaning (which requires strictly five dimensions) but as for the facial parts that are used to express the performative, and in which conflict may arise (see figure 9). Indeed, a common feature of the performatives whose value along the orientation dimension is 'other-oriented' is a 'head nod': performatives of this category are, for example, 'praise', 'approve', 'confirm', 'agree'. On the other hand, 'Submissive' and 'self-oriented' performatives (e.g. 'implore') show inner eyebrow raising, while 'self-oriented', and 'dominant/neutral' performatives (such as 'order', 'criticise', 'disagree', 'refuse') have a frown in common. In our BN, the two dimensions are represented as

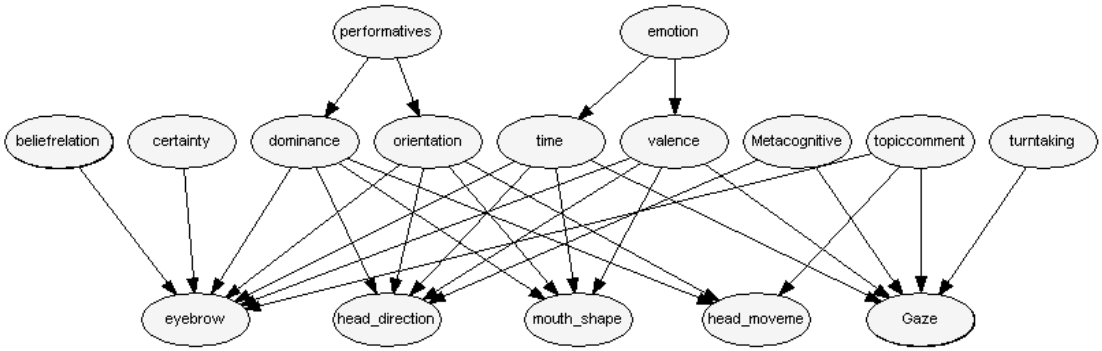

Figure 8: Belief Network linking facial communicative functions and facial signals

intermediary nodes (thus simplifying the construction of the BN), which are linked to the leaf (signal) nodes. For example the performative 'implore' is characterised as being 'submissive' and in 'self-oriented', 'advice' as being 'neutral' and 'other-oriented', 'order' as being 'dominant' and 'self-oriented'. On one hand this allows us to study how common features of performatives prevail in the final facial expressions; on the other hand, it also helps us in reducing the number of entry nodes of our BN.

emotion dimensions : Using the same reasoning as for the performatives, we define emotion along few dimensions. These dimensions are 'valence' (positive or negative) and 'time' (past, current and future) [20]. Valence is commonly used to differentiate emotions. Examples of positive emotions are 'joy', 'happy-for', 'satisfaction', 'like') while examples of negative emotions are 'anger', 'sadness', 'fear', 'dislike', 'reproach'. The dimension 'time' refers to the time at which the event that triggers the emotion is happening [20]. 'Fear' or 'distress' refer to an event that might happen in the future, while 'sadness' or 'resentment' are due to events happened in the 'past'. 'Disgust' is due to an event happening at the 'current' time. Furthermore this representation allows one to characterise emotions based on their facial expressions. 'Tense lips' are common to the negative emotions (envy, jealousy, anger, fear) while a 'frown' will characterise negative emotions happened at the 'current time' (for example anger). 'Positive' emotions are often distinguished by a 'smile' (e.g. 'joy', 'happy-for', 'satisfaction', 'gratitude').

When a conflict is encountered, the BN initialised the concerned communicative function at 100. The BN delivers the probabilities that each signal (involved in both communicative functions) has to be selected to form the new expression. The emotion 'satisfaction' and the certainty 'certain' are initialised at 100 by the BN. Knowing which emotion has been selected, the values of the intermediary nodes 'valence' and 'time' are computed (the values are shown in the figure). The value of the eyebrows for resolution of the signal conflicts is then output by the BN: 'frown' receives the higher probability. Thus, the expression

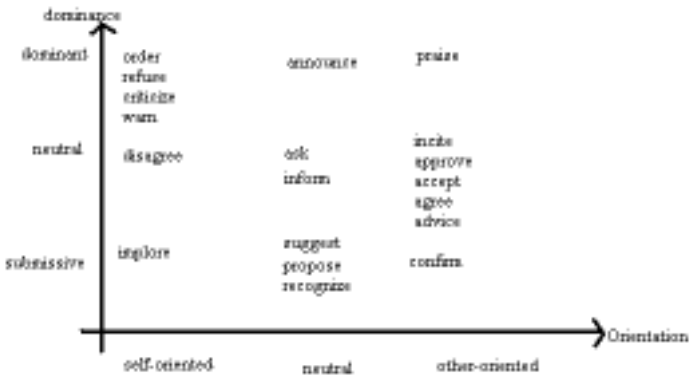

Figure 9: Cluster of performatives along the dimensions 'dominance' and 'orientation'

resulting from the combination of the affective function 'satisfaction' ('raised eyebrow' + 'smile' + 'head nod') with the certainty function 'certain' (frown) will simply be 'frown' + 'smile' + 'head nod'; that is, it cuts off the 'satisfaction' signal at the eyebrow level (See Figure 7). This method allows us to combine expressions at a finer level and to resolve the possible conflicts at the signal level.

10 Generation of the facial animation

After resolving the potential conflicts between the facial signals, we proceed with the animation generation for the agent. Lip shapes are computed based on a computation model described in [23]. The animation is obtained by conversing each facial signal in their corresponding facial parameters. Our facial model is compliant with MPEG-4 standard [11, 21]. The facial model is the core of an MPEG-4 decoder and is based on the specifications for "Simple Facial Animation Object Profile"[23]. Two sets of parameters describe and animate the 3D facial model: facial animation parameter set (FAPs) and facial definition parameter (FDP). The FDPs define the shape of the model while FAPs define the facial actions. FAPs correspond to the displacements of facial features. When the model has been characterized with FDP, the animation is obtained by specifying for each frame the values of FAPs. So we represent each signal as a set of FAPs. For instance: a raising eyebrow that marks uncertainty is generated by the FAPs 31, 32, 33 for the left eyebrow and the FAPs 34, 35, 35 for the right eyebrow. A facial expression is characterized not only by the muscular contraction that gives rise to it, but also by an intensity factor and a duration. The intensity factor is rendered by specifying a given intensity for every FAP. The temporal factor is modeled by three parameters: onset, apex and offset [12] (as explained in section 7.1). Thus, in our system, every facial signal is characterized by a set of FAPs to define its corresponding facial expression as well as by an onset and offset. Moreover, our model includes wrinkles

and folds to ensure more realism.

11 The Facial Display Definition Language

Humans are very good at showing a large spectrum of facial expressions; but at the same time, humans may display facial expressions varying by very subtle differences, but whose differences are still perceivable. We have developed a language to describe facial expressions as (meaning, signal) pairs. These expressions are stored in a library. Defining facial expressions using keyword such as 'happiness, raised eyebrow, surprise' does not capture these slight variations. In our language, an expression may be defined at a high level (a facial expression is a combination of other facial expressions already pre-defined) or at a low level (a facial expression is a combination of facial parameters). The low level facial parameters correspond to the MPEG-4 Facial Animation Parameters (FAPs) [23]. The language allows one to create a large variety of facial expressions for any communicative functions as well as the subtleties that distinguish facial expressions. It allows also us to create a "facial display dictionary" which can easily be expanded. When a text marked with communicative function tags is given in input, the 'Greta' system looks in the library to which signals corresponds each meaning specified by the APML tag; These tags gets then instantiated by the corresponding signals values.
Paradiso et al [22] have established an algebra to create facial expressions. The authors have elaborated operators that combine and manipulate facial expressions. Our language has the only purpose to create facial expressions that are associated to a given communicative functions.
In the next sections we describe the language we have developed to define and to store facial expressions.

11.1 Facial Basis

In our system we distinguish "facial basis" (FB) from "facial display" (FD). An FB involves one facial part such as the eyebrow, mouth, jaw, eyelid and so on. FB includes also facial movements such as nodding, shaking, turning the head and movement of the eyes. Each FB is defined as a set of MPEG-4 compliant FAP parameters:

$$FB = \{fap3 = v_1, \ldots\ldots\ldots, fap69 = v_k\};$$

where v_1,. . . ,v_k specify the FAPs intensity value. An FB can also be defined as a combination of FB's by using the '+' operator in this way:

$$FB' = FB_1 + FB_2;$$

where FB_1 and FB_2 can be:

- Previously defined FB's

Figure 10: The combination of "raise_left" FB (left) and "raise_right" FB (centre) produces "raise_eyebrows" FB (right)

- an FB of the form: $\{fap3 = v_1, \ldots\ldots\ldots, fap69 = v_k\}$

Let us consider the *raising eyebrows* movement. We can define this movement as a combination of the *left* and *right* raising eyebrow. Thus, in our language, we have:

$raise\_eyebrows = raise\_left + raise\_right;$

where *raise_left* and *raise_right* are defined, respectively, as:

$raise\_left = \{fap31 = 50, fap33 = 100, fap35 = 50\};$ *and* $raise\_right = \{fap32 = 50, fap34 = 100, fap36 = 50\};$

Figure 10 illustrates the resulting *raise_eyebrows* FB.
We can also increase or decrease the intensity of a single facial basis by using the operator '*':

$$FB' = FB * c = \{fap3 = v_1 * c, \ldots\ldots\ldots, fap69 = v_k * c\};$$

Where FB is a "facial basis" and 'c' a constant. The operator '*' multiplies each of the FAPS constituting the FB by the constant 'c'. For example if we want a eyebrows raising with greater intensity (Figure 11):

$large\_eyebrows\_raising = raise\_eyebrows * 2;$

11.2 Facial Displays

A facial display (FD) corresponds to a facial expression. Every FD is made up of one or more FB's:

$$FD = FB_1 + FB_2 + FB_3 + \ldots\ldots + FB_n;$$

We can define the 'surprise' facial display in this way:

Figure 11: The "raise_eyebrows" FB (left) and the "large_eyebrows_raising" FB (right)

Figure 12: The combination of "surprise" FD (left) and "sadness" FD (centre) produces the "worried" facial display (right)

$surprise = raise\_eyebrows + raise\_lids + open\_mouth$;

We can also define an FD as a linear combination of two or more (already) defined facial displays using the '+' and '*' operators. For example we can define the "worried" facial display as a combination of "surprise" (slightly decreased) and "sadness" facial displays (Figure 12):

$worried = (surprise * 0.7) + sadness$;

12 State of the art

In the construction of embodied agents capable of expressive and communicative behaviors, an important step is to reproduce affective and conversational facial expressions on synthetic faces [2, 6, 5, 17, 18, 28, 15]. For example, REA, the real estate agent [5], is an interactive agent able to converse with a user in real-time. REA exhibits refined interactional behaviors such as gestures for feedback or turn-taking functions. Cassell and Stone [7] designed a multi-modal manager whose role is to supervise the distribution of behaviors across the several channels (verbal, head, hand, face, body and gaze). BEAT [8]

is a toolkit to synchronize verbal and nonverbal behaviors. Cosmo [17] is a pedagogical agent particularly keen on space deixis and on emotional behavior: a mapping between pedagogical speech acts and emotional behavior is created by applying Elliott's theory [13]. Ball and Breese [2] apply bayesian networks to link emotions and personality to (verbal and non-verbal) behaviors of their agents. André et al. [1] developed a rule-based system implementing dialogs between lifelike characters with different personality traits (extroversion and agreeableness). Marsella et al. [19] developed an interactive drama generator, in which the behaviors of the characters are consistent with their emotional state and individuality.

13 Conclusion

In this chapter we have presented our work toward the creation of ECAs. We have integrated in our system some aspects of non-verbal communication. The set of communicative functions we are considering are clustered depending on the type of information they provide: information on the speaker's belief, intention, affect and also on the speaker's cognitive state. To each of these function corresponds a signal in the form of facial expression, gaze behavior, head movement. Working at the level of communicative function rather than at the signal level allows us to concentrate on the type of information a face would communicate as well as to be independent of the way a communicative functions get instantiated as a signal.
A language has been established to define these signals. In this current work we are concentrating only on "prototype" communicative functions in the sense that we have defined a correspondence between the meaning and the signal associated to a communicative function without any information regarding the speaker's identity. Identity is the aggregation of several components such as culture, gender, age, profession, physical state, personality. These aspects intervene in the selection of appropriate signals to display the information to convey and their expressivity. Indeed culture could vary the allowed amount of gaze toward our interlocutor, the display or not of a given emotion; age is a determinant for the selection of gesture; a young child do not have a large variety of communicative gesture; gender may affect the amount of gaze toward our conversation partner... Thus, we need to define a formalism that would integrate identity aspects into the creation of ECAs.

14 Acknowledgement

We are grateful to Isabella Poggi and Fiorella de Rosis for their valuable help. We greatly thank Elisabetta Bevacqua for developing the lip shape model for speech.

References

[1] E. Andre, T. Rist, S. van Mulken, M. Klesen, and S. Baldes. The automated design of believable dialogues for animated presentation teams. In S. Prevost J. Cassell,

J. Sullivan and E. Churchill, editors, *Embodied Conversational Characters*. MITpress, Cambridge, MA, 2000.

[2] G. Ball and J. Breese. Emotion and personality in a conversational agent. In S. Prevost J. Cassell, J. Sullivan and E. Churchill, editors, *Embodied Conversational Characters*. MITpress, Cambridge, MA, 2000.

[3] M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, 36(2):253–263, 1999.

[4] A.W. Black, P. Taylor, R. Caley, and R. Clark. Festival. http://www.cstr.ed.ac.uk/projects/festival/.

[5] J. Cassell, J. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in conversational interfaces: Rea. In *CHI'99*, pages 520–527, Pittsburgh, PA, 1999.

[6] J. Cassell, C. Pelachaud, N.I. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Computer Graphics Proceedings, Annual Conference Series*, pages 413–420. ACM SIGGRAPH, 1994.

[7] J. Cassell and M. Stone. Living hand and mouth. Psychological theories about speech and gestures in interactive dialogue systems. In *AAAI99 Fall Symposium on Psychological Models of Communication in Collaborative Systems*, 1999.

[8] J. Cassell, H. Vilhjálmsson, and T. Bickmore. BEAT : the Behavior Expression Animation Toolkit. In *Computer Graphics Proceedings, Annual Conference Series*. ACM SIGGRAPH, 2001.

[9] W.S. Condon and W.D. Osgton. Speech and body motion synchrony of the speaker-hearer. In D.H. Horton and J.J. Jenkins, editors, *The Perception of Language*, pages 150–184. Academic Press, 1971.

[10] N. DeCarolis, V. Carofiglio, and C. Pelachaud. From discourse plans to believable behavior generation. In *International Natural Language Generation Conference*, New-York, 1-3 July 2002.

[11] P. Doenges, T.K Capin, F. Lavagetto, J. Ostermann, I.S. Pandzic, and E. Petajan. MPEG-4: Audio/video and synthetic graphics/audio for real-time, interactive media delivery, signal processing. *Image Communications Journal*, 9(4):433–463, 1997.

[12] P. Ekman. About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, *Human ethology: Claims and limits of a new discipline: contributions to the Colloquium*, pages 169–248. Cambridge University Press, Cambridge, England; New-York, 1979.

[13] C. Elliott. *An Affective Reasoner: A process model of emotions in a multiagent system.* PhD thesis, Northwestern University, The Institute for the Learning Sciences, 1992. Technical Report No. 32.

[14] I.A. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. *Proceedings of Computer Vision and Pattern Recognition (CVPR 94)*, pages 76–83, 1994.

[15] W.L. Johnson, J.W. Rickel, and J.C. Lester. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *To appear in International Journal of Artificial Intelligence in Education*, 2000.

[16] A. Kendon. Movement coordination in social interaction: Some examples described. In S. Weitz, editor, *Nonverbal Communication*. Oxford University Press, 1974.

[17] J.C. Lester, S.G. Stuart, C.B. Callaway, J.L. Voerman, and P.J. Fitzgerald. Deictic and emotive communication in animated pedagogical agents. In S. Prevost J. Cassell, J. Sullivan and E. Churchill, editors, *Embodied Conversational Characters*. MITpress, Cambridge, MA, 2000.

[18] M. Lundeberg and J. Beskow. Developing a 3D-agent for the August dialogue system. In *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing*, Santa Cruz, USA, 1999.

[19] S. Marsella, W.L. Johnson, and K. LaBore. Interactive pedagogical drama. In *Proceedings of the 4th International Conference on Autonomous Agents*, pages 301–308, Barcelona, Spain, June 2000.

[20] A. Ortony. On making believable emotional agents believable. In R. Trappl and P. Petta, editors, *Emotions in humans and artifacts*. MIT Press, Cambridge, MA, in press.

[21] J. Ostermann. Animation of synthetic faces in MPEG-4. In *Computer Animation'98*, pages 49–51, Philadelphia, USA, June 1998.

[22] A. Paradiso and M. L'Abbate. A model for the generation and combination of emotional expressions. In *Multimodal Communication and Context in Embodied Agents, Proceedings of the AA'01 workshop*, Montreal, Canada, May 2001.

[23] C. Pelachaud. Visual text-to-speech. In Igor S. Pandzic and Robert Forchheimer, editors, *MPEG4 Facial Animation - The standard, implementations and applications*. John Wiley & Sons, to appear.

[24] C. Pelachaud, N.I. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, January-March 1996.

[25] C. Pelachaud, V. Carofiglio, B. De Carolis, and F. de Rosis. Embodied contextual agent in information delivering application. In *First International Joint Conference on Autonomous Agents & Multi-Agent Systems (AAMAS)*, Bologna, Italy, July 2002.

[26] I. Poggi. Mind markers. In N. Trigo M. Rector, I. Poggi, editor, *Gestures. Meaning and use.* University Fernando Pessoa Press, Oporto, Portugal, 2002.

[27] I. Poggi and C. Pelachaud. Facial performative in a conversational system. In S. Prevost J. Cassell, J. Sullivan and E. Churchill, editors, *Embodied Conversational Characters*. MITpress, Cambridge, MA, 2000.

[28] I. Poggi, C. Pelachaud, and F. de Rosis. Eye communication in a conversational 3D synthetic agent. *AI Communications*, 13(3):169–181, 2000.

[29] A.E. Scheflen. The significance of posture in communication systems. *Psychiatry*, 27, 1964.

[30] Y. Yacoob and L. Davis. *Computer Vision and Pattern Recognition Conference*, chapter Computing spatio-temporal representations of human faces, pages 70–75. IEEE Computer Society, 1994.