MULTIMODAL COMPLEX EMOTIONS: GESTURE EXPRESSIVITY AND BLENDED FACIAL EXPRESSIONS

JEAN-CLAUDE MARTIN

LIMSI-CNRS, BP 133, Orsay, 91403, France
martin@limsi.fr

RADOSLAW NIEWIADOMSKI

Department of Mathematics and Computer Science,
University of Perugia, Italy
radek@dipmat.unipg.it

LAURENCE DEVILLERS

LIMSI-CNRS, BP 133, Orsay, 91403, France
devil@limsi.fr

STEPHANIE BUISINE

LCPI-ENSAM, 151 boulevard de l'Hôpital,
Paris, 75013, France
stephanie.buisine@paris.ensam.fr

CATHERINE PELACHAUD

LINC, IUT of Montreuil, Université Paris VIII,
140 rue Nouvelle France, Montreuil, 93100, France
c.pelachaud@iut.univ-paris8.fr

One of the challenges of designing virtual humans is the definition of appropriate models of the relation between realistic emotions and the coordination of behaviors in several modalities. In this paper, we present the annotation, representation and modeling of multimodal visual behaviors occurring during complex emotions. We illustrate our work using a corpus of TV interviews. This corpus has been annotated at several levels of information: communicative acts, emotion labels, and multimodal signs. We have defined a copy-synthesis approach to drive an Embodied Conversational Agent from these different levels of information. The second part of our paper focuses on a model of complex (superposition and masking of) emotions in facial expressions of the agent. We explain how the complementary aspects of our work on corpus and computational model is used to specify complex emotional behaviors.

Keywords: Emotion; multimodality; Embodied Conversational Agent; corpus.

1. Introduction

One of the challenges of designing virtual humans is the definition of appropriate models of the relation between realistic emotions and the coordination of behaviors in several modalities. Studies of the non-verbal behaviors occurring during emotions have focused on mono-modal and acted basic emotions during experimental in-lab situations. Yet, in order to design Embodied Conversational Agents (ECAs) with human-like qualities, other sources of knowledge on multimodal behaviors occurring during real-life complex emotions deserve consideration, such as audiovisual corpora of spontaneous behaviors. This raises several questions: How does one collect data on spontaneous emotions? How does one represent and classify such complex emotions? Which dimensions of multimodal behaviors are perceptually related to these emotions and require representation?

Our aim is not only to reproduce multimodal behaviors with an ECA but also to study the coordination between modalities during emotional behaviors, in particular in the case of complex emotions. In order to design ECAs with such human-like qualities, one preliminary step is to identify the levels of representation of emotional behavior. For example, regarding the analysis of videos of real-life behaviors, before achieving the long-term goal of fully automatic processing of emotion from low levels (e.g. image processing, motion capture) to related behaviors in different modalities, a manual annotation phase might help to identify the representation levels that are relevant for the perception of complex emotions. Similarly to the copy-synthesis approaches that have been developed for speech, the replay by an ECA of these manually annotated behaviors can be useful for the validation of the model relating emotions and multimodal behaviors.

Since the externalization of nonverbal behaviors plays an important role in the perception of emotions, our approach is to model what is visible; that is we consider the signals and how they are displayed and perceived. We do not model the processes that were made to arrive to the display of such and such signals; we simply model the externalization part. We are interested in understanding and modeling how a given emotion would be both perceived and expressed quantitatively and qualitatively.

In this paper, we propose a model for the representation of non-verbal visual behaviors occurring during complex emotions. It makes a distinction between two types of complex emotions: superposition of emotions and masking of emotions. The first part of the model aims at representing gesture expressive behaviors during superposition of emotions and is grounded in a video corpus. The second part of the model aims at representing facial behaviors during superposition of emotions and masking of emotions. It is grounded in the literature of facial expressions during complex emotions.

The remaining parts of this paper are structured as follows. Section 2 summarizes some of the studies on complex emotions, gesture expressivity, and facial expressions. Section 3 provides two examples of gesture and facial expression behaviors observed during complex emotions in videos of TV interviews. Section 4 describes the part of the model that we propose for representing gesture expressivity. Section 5 describes the part of the model focusing on facial expressions of complex emotions. Section 6 explains how this model has been used for the annotation of expressive behaviors observed in videos, and for the specification of expressive gestures in the Greta agent.[1]

2. Related Work

There has been a lot of psychological research on emotion and nonverbal communication in facial expressions,[1] vocal expressions[2–4] and expressive body movements.[5–8] Yet, these psychological studies were based mostly on acted basic emotions: anger, disgust, fear, joy, sadness, surprise. In the area of affective computing, recent studies are also limited with respect to the number of modalities or the spontaneity of the emotion. Cameras are used by Kapur *et al.* to capture markers placed on various points of the whole body in order to recognize four acted basic emotions (sadness, joy, anger, fear).[9] Some studies deal with more complex emotions. In the "Lost Luggage" experiment, passengers at an airport were informed that their luggage has been lost, and the participants were asked to rate their emotional state.[10] Scherer and his colleagues show in this experiment that some events may give rise to several simultaneous emotions. These emotions are referred to as complex emotions and also as blends of emotions.[1,10,11] They may occur either as a quick succession of different emotions, the superposition of emotions, the masking of one emotion by another one, the suppression of one emotion or the overacting of one emotion.

In particular, in the visual modalities, these blends produce "multiple simultaneous facial expressions."[12] Depending on the type of blending, the resulting facial expressions are not identical. A masked emotion may leak over the displayed emotion,[1] while superposition of two emotions will be shown by different facial features (one emotion being shown on the upper face while another one on the lower face).[1] Perceptual studies have shown that people are able to recognize facial expression of felt emotion[13,14] as well as fake emotion.[13] Similar studies producing similar results have been conducted on ECAs.[15] In a study on a deceiving agent, Rhem and André found that the users were able to differentiate when the agent was displaying expression of felt emotion or expression of fake emotion.[16] Aiming at understanding if facial features or regions play identical roles in emotion recognition, Bassili[17] and later on Gouta and Miyamoto,[18] and Constantini *et al.*[19] performed various perceptual tasks, and Cacioppo *et al.*[20] studied psychological facial activity. They found that positive emotions are mainly perceived from the expression of the lower face (e.g. smile) while negative emotion from the upper face (e.g. frown).

Very few models of facial expressions for such complex emotions have been developed so far for ECAs. The interpolation between facial parameters of given expressions is commonly used to compute the new expression. MPEG-4 proposes to create a new expression as a weighted interpolation of any of the six predefined expressions of emotions.[15,21] More complex interpolation schemes have been proposed.[22–24] Duy Bui[25] introduced a set of fuzzy rules to determine the blended expressions of the six basic emotions. In this approach, a set of fuzzy rules is attributed to each pair of emotions. The intensities of muscles contraction for the blended expression are related to emotions intensities using fuzzy inference. With respect to other modalities than facial expressions, static postures were recorded by De Silva *et al.* using a motion capture system during acted emotions (two nuances for each of four basic emotions).[26] Gunes *et al.* fused the video processing of facial expression and upper body gestures in order to recognize six acted emotional behaviors (anxiety, anger, disgust, fear, happiness, uncertainty).[27] A vision-based system that infers acted mental states (agreeing, concentrating, disagreeing, interested, thinking, and unsure) from head movement and facial expressions was described by el Kaliouby *et al.*[28] Choi *et al.* described how video processing of both facial expressions and gaze are mapped onto combinations of seven emotions.[29] Yet, real-life multimodal corpora are indeed very few despite the general agreement that it is necessary to collect audio-visual databases that highlight naturalistic expressions of emotions as suggested by Douglas-Cowie *et al.*[30]

Regarding the design of ECAs, the majority of the works in this research area use either motion capture data,[31,32] or videos.[23,33] Some studies do not use any corpus but propose sophisticated models of mixed emotional expressions. For example, an algorithm for generating facial expressions for a continuum of pure and mixed emotions of varying intensity is described by Albrecht *et al.*[22] Results from the literature in psychology are useful for the specification of ECAs, but provide few details, nor do they study variations about the contextual factors of multimodal emotional behavior. Very few researchers have been using context specific multimodal corpora for the specification of an ECA.[34] Cassell *et al.*[35] described how the multimodal behaviors of subjects describing a house were annotated and used for informing the generation grammar of the REA agent.

3. Complex Emotions: Two Illustrative Examples

In this section, we briefly describe two illustrative examples of multimodal behaviors observed during complex emotions in videos of TV interviews from the EmoTV corpus.[36] In video #3, a woman is reacting to a recent trial in which her father was kept in jail. As revealed by the manual annotation of this video by three coders, her behavior is perceived as a complex combination of despair, anger, sadness and disappointment. Furthermore, this emotional behavior is perceived in speech and in several visual modalities (head, eyes, torso, shoulders and gestures). In another video (video #41), a woman is pretending to be positive after negative election results. Such a video has been annotated as a combination of negative labels (disappointment, sadness, anger) and positive labels (pleased, serenity). The annotation of multimodal behaviors reveals that her lips show a smile but with lips pressed. This example illustrates the combinations of facial features during complex emotions.

Several levels of annotation are coded in EmoTV using the Anvil tool[37]: some information regards the whole video (called the "global level"); while some other information is related to emotional segments (the "local" level); at the lowest level, there is detailed time-based annotation of multimodal behaviors including movement expressivity. Several emotional segments are identified by the annotators as being perceptually consistent. The annotation scheme enables the coders to select two verbal labels describing the emotion for a single emotional segment. Three annotators created this segmentation and labeled each segment with one or two labels.[36] The three annotations are combined into a single soft vector.[38,39] In video #3, three emotional segments have been identified by the coders and annotated with the following vectors: segment 1 (100% anger), segment 2 (67% anger, 11% despair, 11% disappointment, 11% sadness), segment 3 (56% despair, 33% anger, 11% sadness). A perceptive test on this video with 40 coders validated these annotations.[40]

4. Representing, Modeling and Evaluating Expressivity

4.1. *Representing expressivity*

Several taxonomies of communicative gestures have been proposed highlighting the link between gesture signals and its meaning.[41–43] The type of the gesture, its position in the utterance, its shape but also its manner of execution provide information about the speaker's mental and emotional state. Facial expressions are recognized for their power of expressing emotional state. Many studies have characterized facial expressions for emotion categories[1] and for appraisal dimensions.[44] While there is a less direct link between gesture shapes and emotions, several studies have shown that gesture manners are good indicators of emotional state.[8,45,46] Gesture manners are also linked to personality traits (nervousness), physiological characteristics (graciousness), physical state (tiredness), etc. Most of computational models of ECA behavior have dealt with gesture selection and gesture synchronization with speech.[47–49] We propose a model of gesture manner, called gesture expressivity, that acts on the production of communicative gestures. Our model of expressivity is based on studies of nonverbal behavior.[8,45,46] We describe expressivity as a set of six dimensions.[50] Each dimension acts on a characteristic of communicative gestures. *Spatial Extent* describes how large the gesture is in space. *Temporal Extent* describes how fast the gesture is executed. *Power* describes how strong the performance of the gesture is. *Fluidity* describes how two consecutive gestures are co-articulated one merging with the other. *Repetition* describes how often a gesture is repeated. *Overall activity* describes how many behavior activities there are over a time span. This model has been implemented in the Greta ECA.[51]

4.2. *Evaluation of the gesture expressivity model*

We have conducted two studies to evaluate our gesture expressivity model which is the central part of the copy-synthesis approach described in Sec. 6. These two studies involved a total number of 106 users (80 males, 26 females; aged 17 to 25). All were first and second year French university students. Each user completed only one of the two tests. Both tests consisted in observing sets of video clips (two per trial for the first test, four for the second test) and answering a questionnaire. The video clips differ only on the gesture expressivity of the agent (same audio and same gesture type).

The goal of the first study was to test the following hypothesis: the chosen implementation for mapping single dimensions of expressivity onto animation parameters is appropriate — a change in a single dimension can be recognized and correctly attributed by users. In this test, users ($N = 52$) were asked to identify a single dimension in forced-choice comparisons between pairs of animations. Table 1 presents the distribution of users' answers for each parameter. Gray cells indicate when they met our expectations: this diagonal totals 320 answers, which corresponds to 43.1% of accurate identifications of parameters. The chi-square test shows that this distribution cannot be attributed to chance [$\chi^2(35) = 844.16,\ p < 0.001$]. Recognition was best for the dimensions Spatial Extent and Temporal Extent. Modifications of Fluidity and Power were judged incorrectly more often, but the correct classification still had the highest number of responses. The parameter Repetition was frequently interpreted as Power. Overall Activation was not well recognized. Overall, we take the results of the first test as indication that the mapping from dimensions of expressivity to gesture animation parameters is appropriate for the Spatial Extent and Temporal Extent dimensions while it needs refinement for the other parameters.

The hypothesis tested in the second study was the following: combining parameters in such a way that they reflect a given communicative intent will result in a more believable overall impression of the agent. Avoiding behavior qualities that imply an emotional state or a personality trait, we considered the three following qualities: abrupt, sluggish, and vigorous. Abrupt is characterized by rapid, discontinuous and powerful movements. Sluggish is characterized by slow, effortless and close to the body but fluid movements. Vigorous is characterized by a lot of large, fast, fluid and repetitive movements. For each quality, we generated four animations. One animation corresponds to the neutral, generic animation, two to variants of the chosen expressive intent (strongly and slightly expressive) and one to an opposite assignment of expressivity parameters. This test ($N = 54$) was conducted as a preference ranking task: the user had to order four animations from the most appropriate to the least appropriate with respect to the expressive intent. For the abrupt and vigorous qualities, users preferred the coherent performances as we had hoped [$F(3/153) = 31.23,\ p < 0.001$ and $F(3/153) = 104.86,\ p < 0.001$, respectively]. The relation between our parameterization and users' perception can also be expressed as a linear correlation, which amounts to +0.655 for the abrupt quality and +0.684 for the vigorous quality. Conversely for the sluggish quality, the effect of input stimuli was not significant [$F(3/153) = 0.71$, N.S.]: the overall rating of

Table 1. Distribution of users' answers as a function of the modified parameter.

	Perceived modification								
	Spatial extent	Temporal extent	Fluidity	Power	Repetition	Overall activation	No modification	Do not know	Total
Modified parameter									
Spatial Extent	77	2	5	5	3	3	3	8	106
Temporal Extent	3	104	7	13	7	1	1	5	141
Fluidity	2	4	42	10	23	2	34	7	124
Power	7	8	23	42	9	6	27	8	130
Repetition	18	12	17	20	35	5	10	8	125
Overall Activation	7	7	7	17	6	20	41	11	116
Total	114	137	101	107	83	37	116	47	742

stimuli was random and the linear correlation was almost null (+0.047). This may be attributable partly to the inadequacy between the specific gestures that accompanied the text and the way a sluggish person would behave. This finding raises the need for integrating gesture selection and gesture modification to best express an intended meaning.

In the first test, we checked if subjects perceived variation of each parameter, while in the second perceptual test we looked at the interpretation of these variations. Since our expressivity parameters show some dependency with one another, we wanted to check that the subject perceived their individual changes and their combined meaning in two separate perceptual tests. The results confirm that our general approach for expressivity modeling is worthwhile pursuing. A notable advantage of our implementation is to enable the decomposition of gesture expressivity and the test of parameters one by one. In the experiment by Wallbott, actors were instructed to act basic emotions.[8] This experiment revealed that each acted emotion had an impact on all the parameters of expressivity. The first perceptual test we conducted would have been surely more difficult to control with a human actor instead of an agent: humans may be able to control their expressivity to a certain extent but can hardly isolate each parameter. In our animations, the decomposition of expressivity may have produced artificial behaviors but this step seemed necessary to evaluate our model and highlight possible ways of improvement. These results will be used to refine the technical implementation of individual parameters to achieve higher quality animation and better visibility of changes to the parameters. For the second perceptual test, we were careful to avoid introducing labels related to personality or emotion. While we ultimately want to simulate such traits and mental states, the link from these high-level concepts to the expressive dimensions is still not clear — the social psychology literature on this problem appears to be very sparse. This second test mainly showed that we need to integrate gesture selection and gesture modification when generating an animation. A shortcoming of the current test was that only a single utterance with a unique gesture selection was used with varying animations. A wider variety of different utterances and corresponding gesture selections is needed to understand the perception of expressivity.

5. Representing and Modeling Blended Facial Expressions

In this section, we present a computational model of facial expressions arising from blends of emotions. Instead of formulating our model at the level of facial muscle contractions or FAP values, we propose a face partition based model, which not only computes the complex facial expressions of emotions but also distinguishes between different types of blending. Blends (e.g. superposition and masking) are distinguished among each other as they are usually expressed by different facial areas.[1,52] Expressions may also occur in rapid sequences one after the other. Moreover, the expression of masking a felt emotion by a fake one (i.e. not felt) is different from the expression corresponding to the superposition of two felt emotions.[1] Thus complex facial expressions can be distinguished depending on the type of emotions, their apparition in time (sequence, superposition) as well as if the emotions are felt or fake. For the moment, we have considered only two cases of complex facial expressions: the superposition of two felt emotions and the masking of a felt emotion with a fake one. In the following sub-section, we present a general framework for our model and describe next details of computational procedures based on fuzzy inference.

5.1. *Blend of emotions*

The analysis of the video corpus has revealed the evidence of disparity between different types of complex expressions.[38] Different situations such as "superposed," "masked" or "sequential" were recognized by annotators. In our model, we have defined for each type of blend a set of fuzzy rules *SFR*. In Ekman's research on blend of emotions, his analysis is restricted to a small number of so-called basic emotions: anger, disgust, fear, joy, sadness and surprise. Our model is based on the set of rules he has established for the blending of these six emotions. However, there exist many more expressions of emotions,[23] some of which are considered to have a universal aspect as well.[53,54] Emotions like disappointment, despair or pride appear in the annotation of our video corpus. To overcome this restriction, we introduced the notion of similarity between expressions. We compute similarity between expressions of any given emotion and basic emotion using fuzzy similarity.[55] Let $\mathrm{Exp}(E_i)$ be the expression of an emotion E_i and $\mathrm{Exp}(N)$ be the neutral expression. Let us suppose that any facial expression is divided into n areas $F_k,\ k = 1, \ldots, n$. Each F_k represents a unique facial part like brows or eyes. For any emotion E_i, $\mathrm{Exp}(E_i)$ is composed by n different facial areas. Thus, $\mathrm{Exp}(E_i) = \{F_k^{(Ei)},\ k = 1, \ldots, n\}$. In our model we are currently considering seven areas, namely brows, upper eyelids, lower eyelids, cheeks, nose, upper lip and lower lip.

Let E_i and E_j be the emotions occurring in a blend and $\mathrm{Exp}_{\mathrm{blend}}(E_i, E_j)$ the resulting complex expression, where *blend* is either masking (M) or superposition (S). The $\mathrm{Exp}_{\mathrm{blend}}(E_i, E_j)$ is also composed by the combination of n different face areas, where each $F_k^{(Ei,Ej)}$ is equal to one corresponding area from $\mathrm{Exp}(E_i),\ \mathrm{Exp}(E_j),\ \mathrm{Exp}(N)$. We note, that for any k in the interval $[1, n]$, $F_k^{(Ei,Ej)}$ cannot contain simultaneously elements of two different expressions; it can be either $\mathrm{Exp}(E_i),\ \mathrm{Exp}(E_j)$, or $\mathrm{Exp}(N)$. That is, a facial area can not show different expressions at the same time; it can show one expression at a time: this expression can come from either emotion or the neutral expression. Combining facial expressions on the same facial area can have the artefact to introduce a new expression. For example, if we add the facial actions in the eyebrow region of surprise "raise-eyebrow" and of anger "frown" we obtain a new facial action "upper-raised-eyebrow-down" that is typically linked to fear. Thus, we opt for the rules that no facial action can be added up on a same facial region. This ensures the conformity of our model with empirical evidence.[1]

Let E_u be one of the basic emotions and let E_i be an input emotion. We aim to compute which basic emotion is the most similar E_i expression-wise. Thus, the fuzzy similarity between E_i and E_u needs to be established. Each emotion E_u is associated to a set of fuzzy intervals in which all plausible expressions for this emotion are defined. That is, for each numerical parameter of an expression of E_u there is a fuzzy interval that specifies a range of plausible values. The value of fuzzy similarity for each significant parameter of $\mathrm{Exp}(E_i)$ and $\mathrm{Exp}(E_u)$ is then established. Finally, all values are combined linearly. At the moment the M-measure of resemblance on FAP values of each expression is used to establish similarity values.[55,56]

Our algorithm works as follows: for each input expression $\mathrm{Exp}(E_i)$ we first define its similarity with the six basic expressions $\mathrm{Exp}(E_u),\ u = 1, \ldots, 6$. The best value, that is, the highest value of similarity, defines the basic emotion whose expression is the most similar to the input one. According to the degree of similarity, the final expression $\mathrm{Exp}_{\mathrm{blend}}(E_i, E_j)$is chosen based on rules of the adequate *SFR* set. Each type of the blend $\{S, M\}$ uses different set of fuzzy rules (SFR_S in case of superposition or SFR_{fake} and $SFR_{\mathrm{felt:}}$ in case of masking; see also Secs. 5.2 and 5.3). These rules describe the principles of composition of facial expressions depending on the blending type. The final expression $\mathrm{Exp}_{\mathrm{blend}}(E_i,\ E_j)$ is obtained by combining *face areas of* $\mathrm{Exp}(E_i),\ \mathrm{Exp}(E_j)$ *and/or* $\mathrm{Exp}(N)$.

5.2. *Masking*

Masking occurs when a felt emotion should not be displayed for some reason; it is preferred to display a different emotional expression. It may be due to some socio-cultural norms, often called *display rules*.[57] Masking can be seen as an asymmetric emotion-communicative function in the sense that given two emotions E_i and E_j, the masking of E_i by E_j leads to a different facial expression than the masking of E_j by E_i.[1] Often humans are not able to control all their facial muscles. Ekman claims that the features of the upper face of any expression are usually more difficult to control.[1] Moreover, felt emotions may be characterized by specific facial features: e.g. sadness brows[1] or *orbicularis oculi* activity in case of joy.[58] Such reliable features lack in fake emotions as they are difficult to do voluntarily.[58] Ekman describes, for any of the so-called basic emotions, which features are missing in fake expressions, in particular in the case of masking. On the other hand, people are not able to inhibit felt emotions completely. Based on Darwin's work, Ekman proposed the *inhibition hypothesis*: elements of facial expressions that are hardly done voluntarily, are also hardly inhibited.[58] Finally, Ekman provides a description of which part of the felt expression leaks during masking.[1]

We call $\mathrm{Exp}_M(E_i, E_j)$ the expression resulting from the masking of a felt emotion E_i by a fake emotion E_j. Two independent sets of fuzzy rules, SFR_{fake} and SFR_{felt}, are defined in the case of masking. The first one — SFR_{fake} — describes the features of the fake expression, while SFR_{felt} — of the felt expression. All rules are of the certainty type.[59] The value of fulfilment of a rule is a degree of similarity between E_i

and E_u. Each input variable corresponds to one basic emotion $E_u,\ u = 1, \ldots, 6$, and each output variable corresponds to one facial region F_k of the resulting expression. In particular, each rule of SFR_{felt} describes leakage of the felt emotion E_u during the masking. Each rule is defined as: *the more the input expression of E_i is similar to the expression of E_u, the more certain the face areas corresponding to the reliable features of E_u should be used in the final expression.*

For example, in the case of the rule for the felt sadness the following information is applied: "*the more the **input expression** is (similar to) **sadness**, the more certain the input **brows** and **upper_eyelids** should be visible.*" It is described in SFR_{felt} by the following rule:

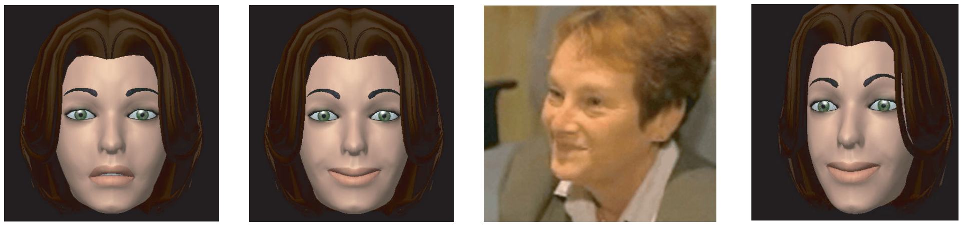
$$\text{If } X \textit{ is } SADNESS \text{ then } F_{\text{brows}} \text{ is } VISIBLE \text{ and } F_{\text{upper eyelids}} \text{ is } VISIBLE \text{ and } \\ \ldots \text{ and } \mathrm{F}_{\text{upper lip}} \text{ is NOT\_VISIBLE and } \mathrm{F}_{\text{lower lip}} \text{ is NOT\_VISIBLE},$$

where X expresses degree of similarity to $\text{Exp}(SADNESS)$ and F_k are face areas of the input expression E_j. According to the inhibition hypothesis, if there is a face area in the masking expression that is not used by the felt emotion, it does not mean that it has to be used by the fake emotion. Each rule of SFR_{fake} describes the reliable features which will certainly not appear in the fake expression of E_i. For example, in the case of the fake joy the following rule is applied: "*the more the **input expression** is (similar to) **joy**, the more certain the area of **lower_eyelids** should **not** be visible.*" It corresponds to the following rule of SFR_{fake}:

$$\text{If } X \text{ is } JOY \text{ then } F_{\text{brows}} \text{ is } VISIBLE \text{ and } F_{\text{upper eyelids}} \text{ is } VISIBLE \text{ and } \\ \mathbf{F}_{\text{lower eyelids}} \text{ is } \textbf{NOT\_VISIBLE} \text{ and } \ldots \text{ and } F_{\text{upper lip}} \text{ is } VISIBLE \text{ and } \\ F_{\text{lower lip}} \text{ is } VISIBLE.$$

The system takes as input two emotion labels: the felt E_i and fake E_j. If the expressions of both emotions are not one of the basic ones (that is if $\text{Exp}(E_i)$ *and/or* $\text{Exp}(E_j)$ is different from $\text{Exp}(E_u)$, $u = 1, \ldots, 6$, the model predicts the final expression based on the degree of similarity between $\text{Exp}(E_i)$ *and/or* $\text{Exp}(E_j)$ and basic expressions. The fake and felt areas of the masking expression are considered separately. Finally, for each F_k,the results of SFR_{felt} and of SFR_{fake} are composed to obtain $\text{Exp}_M(E_i, E_j)$ expression. The conflicts that may rise on some facial areas are resolved according to the *inhibition hypothesis.* In the case in which neither the felt nor the fake emotion can be shown in a given region of the face, the neutral expression is used instead. The final expression is composed of facial regions of the felt emotion, the fake and the neutral ones.

Figure 1 shows the agent displaying the masked expression of disappointment (computed as similar to sadness) and fake joy. The images (a) and (b) display the expressions of disappointment and joy, respectively. Image (d) shows the masking expression. We can notice that the absence of orbicularis oculi activity as indicator of unfelt joy[58]. is visible on both images (c) and (d), the annotated video and the corresponding Greta simulation.



(a) Disappointment (b) Joy (c) Original video (d) Masking of disappointment by joy

Fig. 1. Disappointment masked by joy.

5.3. *Superposition*

Superposition occurs when two different emotions are felt and shown simultaneously. Contrary to the masking case, it does not have the property of asymmetry. The expression $\mathrm{Exp}_S(E_i, E_j)$resulting from the superposition of E_i and E_j is equal to the superposition of E_j and E_i. That is: $\mathrm{Exp}_S(E_i,\, E_j) = \mathrm{Exp}_S(E_j,\, E_i)$. Ekman described this case of blending for all pairs of the six basic emotions.[1] No constructive rules to build the superposition were introduced and only the resulting expressions are described. The superposition of two emotions is usually expressed by combining the upper part of one expression with the lower part of the other one. However, not all combinations of the upper and the lower faces are plausible. As mentioned in Sec. 2, negative emotions are mainly recognized by their upper face (e.g. frown of anger) while positive emotion by their lower face (e.g. smile of happiness) [17–19] Let Z be a set of plausible (according to Ekman) *schemas* for the superposition expression Exp_S. By "schema" we intend the particular division of n face regions $F_k,\ k = 1, \ldots, n$ between any two emotions. At the moment, ten different schemas are considered. The fuzzy inference is used to model the combination of facial expressions $\mathrm{Exp}(E_i)$ and $\mathrm{Exp}(E_j)$ of two emotions E_i and E_j. Each fuzzy rule associates *a pair of basic emotions to an element of* Z. Each rule is defined as: *the more the input expression of* E_i *is (similar to) the expression of* E_u *and the more the input expression of* E_j *is (similar to) the expression of* E_w*, the more certain the upper/lower face areas of* E_i *and lower/upper face areas of* E_j *should be used.*

For example, the superposition of an emotion similar to sadness (X) and of an emotion similar to joy (Y) is described in SFR_S by the following rule:

$$\text{If } X \text{ is } SADNESS \text{ and } Y \text{ is } JOY \text{ then } S_1 \text{ is } FALSE \text{ and } S_2 \text{ is } FALSE \text{ and } S_3 \text{ is } FALSE \text{ and } S_4 \text{ is } FALSE \text{ and } \mathbf{S_5 \text{ is TRUE}} \text{ and } S_6 \text{ is } FALSE \text{ and } S_7 \text{ is } FALSE \text{ and } S_8 \text{ is } FALSE \text{ and } S_9 \text{ is } FALSE \text{ and } S_{10} \text{ is } FALSE$$

where S_i are schemas from a set Z. In particular S_5 corresponds to the schema in which the face areas F_{brows} and $F_{\mathrm{uppereyelids}}$ belong to X while the other face areas

(a) Joy (b) Sadness (c) Original video (d) Super position of joy and sadness

Fig. 2. Superposition of sadness and joy.

belong to Y. The meaning of this rule is: *the more one of the* ***input expressions*** *is* (*similar to*) ***sadness*** *and the other* ***input expression*** *is* (*similar to*) ***joy***, *the more certain is that the final expression contains* **brows**, ***and upper eyelids*** *of the first expression and the* ***mouth*** *area rest of the second.*

The inputs to our system consist of two emotion labels E_i and E_j. The model predicts the final expression based on the degrees of similarity between $\text{Exp}(E_i)$ (*resp.* $\text{Exp}(E_j)$) and $\text{Exp}(E_u)$, $u = 1, \ldots, 6$. The values of fuzzy similarity between adequate pairs of expressions serve to classify an input pair according to plausible schemas for superposition Z. The inhibition hypothesis is not applied in the superposition case. As consequences the neutral expression is not used in the computation of the final expression. Figure 2 shows an example of superposition expression computed by our model. Images (a) and (b) show, respectively, the expressions of joy and of sadness. Image (d) shows the superposition of both expressions as a composition of face areas of both input expressions. In that image, the upper face expresses sadness, and the lower face joy. However, the expression of joy is expressed by $F_{\text{lowereyelids}}$, which contains orbicularis oculi muscle contraction, sign of felt joy. We can note that this muscular contraction was not shown in the Masking condition (Fig. 1). Image (c) shows a video frame annotated with superposition of joy and sadness. Image (d) shows the corresponding Greta simulation.

6. Copy-Synthesis Approach

Our copy-synthesis approach (Fig. 3) is composed of three main steps, namely, annotation of the data, extraction of parameters, and generation of the synthetic agent.

6.1. *Annotation*

Annotation is composed of two steps. Step 1 aims at the automatic annotation of the video with data that can be useful either for the manual annotation of the video or the specification of the agent's behavior: pitch, intensity, etc. Step 2 involves manual annotations of the video. The word-by-word transcription including punctuation is achieved following the LDC norms for hesitations, breath, etc. The video is then annotated at several temporal levels (whole video, segments of the video, behaviors observed at specific moments) and at several levels of abstraction. The global behavior observed during the whole video is annotated with communicative act, emotions and multimodal cues. The segments are annotated with emotion labels and the

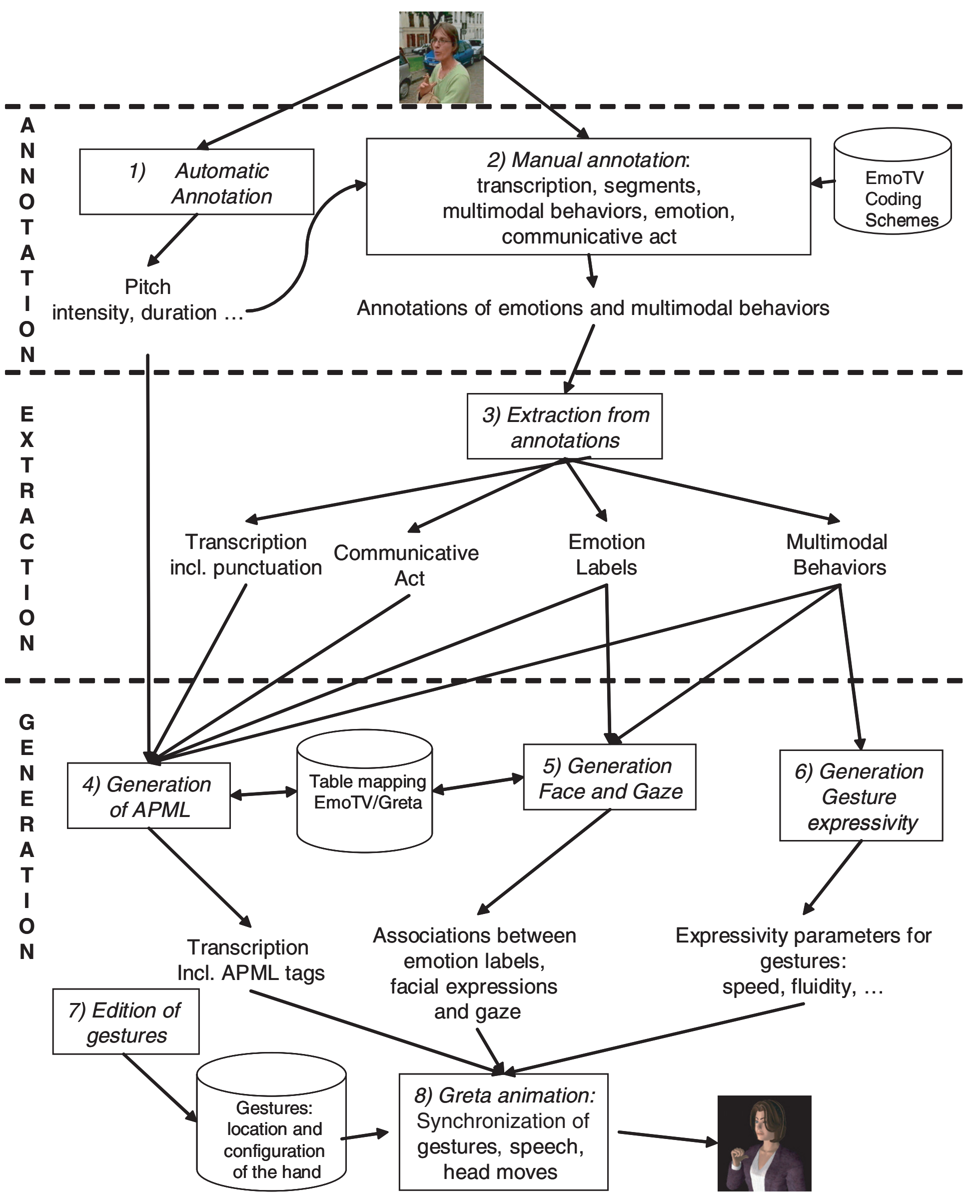


Fig. 3. Copy synthesis approach for studying gesture expressivity during emotions.

modalities perceived as relevant with regards to emotion. We have grounded this coding scheme in requirements collected from the parameters known as perceptually relevant for the study of emotional behavior, and the features of our emotionally rich TV interviews. This section describes how each modality is annotated in order to enable subsequent computation of the relevant parameters of emotional behavior.

Each track is annotated one after the other while playing the audiovisual clip (e.g. the annotator starts by annotating the first track for the whole video and then proceeds to the next track). Movement expressivity is annotated for torso, head, shoulders, and hand gestures. The annotators were instructed to use their own perception for annotating these expressive dimensions. The head pose track contains pose attributes adapted from the FACS coding scheme.[60] Facial expressions are coded using combinations of Action Units.

As for gesture annotation, we have kept some of the attributes used in research on gestures. Thus, our coding scheme enables the annotation of the structural description ("phases") of gestures as their temporal patterns might be related to emotion[34,41]: preparation (bringing arm and hand into stroke position), stroke (the most energetic part of the gesture), sequence of strokes (a number of successive strokes), hold (a phase of stillness just before or just after the stroke), and retract (movement back to rest position). We have selected the following set of gesture functions ("phrases") as they were revealed to be observed in our corpus: manipulator (contact with body or object), beat (synchronized with the emphasis of the speech), deictic (arm or hand is used to point at an existing or imaginary object), illustrator (represents attributes, actions, relationships about objects and characters), emblem (movement with a precise, culturally defined meaning).[34,41] Currently, the hand shape is not annotated since it is not considered as a main feature of emotional behavior in our survey of experimental studies nor in our videos.

Whereas the annotations of emotions have been done by three coders and lead to computation of agreement,[39] the current protocol used for the validation of the annotations of multimodal behaviors is to have a second coder checks the annotations done by a first coder followed by brainstorming discussions. We are currently considering the validation of the annotations by the automatic computation of intercoder agreements from the annotations by several coders.

6.2. *Extraction from annotations*

A module has been designed for extracting from the various annotations the pieces of information which have been identified as required for generation (Step 3 in Fig. 3): the speech transcription, the communicative act, the emotion labels, the dimensions of emotions, the multimodal behaviors (including the number of occurrences and the duration of each multimodal behavior within each segment). The data extracted are

Table 2. Illustrative multimodal emotional profiles extracted from the annotations of three videos (global profile of the whole videos).

	Videos	Video #3	Video #36	Video #30
	Duration	37 s	7 s	10 s
EMOTION	Emotion labels	Anger (55%) Despair(45%)	Anger (62%) Disapoint. (25%) Sadness (13%)	Exaltation (50%) Joy (25%) Pride (25%)
	Intensity (1: min – 5: max)	5	4.6	4
	Valence (1: neg – 5: pos)	1	1.6	4
GESTURE EXPRESSIVITY	% fast vs. % slow	47% vs. 3%	33% vs. 13%	83% vs. 0%
	% hard vs. % soft	17% vs. 17%	20% vs. 0%	0% vs. 27%
	% jerky vs. % smooth	19% vs. 8%	6% vs. 0%	5% vs. 50%
	% expanded vs. % contracted	0% vs. 38%	13% vs. 20%	0% vs. 33%

used to compute a model of multimodal expressive behavior along three dimensions: emotion, activation of head/torso/hand, and gesture expressivity. Table 2 illustrates such results. The percentages indicated in Table 2 are percentages of time and are computed by considering the duration of a given annotation (e.g. Anger) over the whole duration of annotated segments. As explained below, the role of these descriptive profiles is to drive the specifications of the emotional behavior to be replayed by the ECA.

6.3. *Generation*

Our ECA system, Greta, incorporates communicative conversational and emotional qualities.[51] The agent's behavior is synchronized with her speech and is consistent with the meaning of her sentences. To determine speech-accompanying non-verbal behaviors, the system relies on a taxonomy of communicative functions proposed by Isabella Poggi.[43] A communicative function is defined as a pair (meaning, signal) where *meaning* corresponds to the communicative value the agent wants to communicate and *signal* to the behavior used to convey this meaning. We have developed a language to describe gesture signals in a symbolic form.[49] An arm gesture is described by its wrist position, palm orientation, finger direction as well as hand shape. We use the HamNoSys system to encode hand shapes.[61] To control the agent's behavior, we are using the APML representation language, where the tags of this language are the communicative functions.[62] The system takes as input a text tagged with APML labels as well as values for the expressivity dimensions that characterize the manner of execution of the agent's behaviors. The system parses the input text and selects which behaviors to perform. Facial expressions and gaze behaviors are synchronized with speech defined within APML tags. The system looks for the emphasis word. It aligns the stroke of a gesture with this word. Then it computes when the preparation phase of the gesture is as well as if a gesture is hold, if it co-articulates to the next one, or if it returns to the rest position. The expressivity model controls the spatial and dynamism properties of the gestures. The outputs of the system are animation and audio files that drive the animation of the agent.

6.3.1. *Generation of the APML file*

Step 4 consists of generating the APML file used by the Greta system from the data extracted from the annotations such as the speech transcription, the pitch, the communicative act, and the emotion labels. The transcription is directly used in the APML file since it corresponds to the text that the Greta agent has to produce. It is enhanced with several tags. The pitch enables to validate/correct the annotation of prosodic curves adapted from the ToBI model and used by APML. We have also defined a table connecting the annotated communicative act with the closest performative the Greta system knows about. Thus the communicative goal "to complain" used for annotating the video #3 is translated to the performative "to criticize" which corresponds to a specification of the global behavior of the agent (gaze at listener + frown + mouth_criticize). In the videos we studied, the emotional behaviors are complex and are often annotated with several emotional labels. These annotations made by three or more annotators are grouped into an emotional vector. The third segment of video #3 has been annotated with the following vector: 56% of despair, 33% of anger and 11% of sadness. The two most represented categories are grouped into a label "superposition(Despair, Anger)" that is sent to the blend computation module (see Sec. 5). The value of the affect attribute of the rheme tag is specified as this combination of the two major emotion labels computed from the emotional profiles resulting from the annotations (Table 2).

6.3.2. *Generation of gaze behaviors*

The annotations of facial expressions are used in Step 5 to associate the combined emotion label to the annotated gaze behaviors. The durations of the annotation of the gaze are used to specify in the agent the durations of gaze towards the right and left, and the maximum duration of gaze towards the camera. In the third segment of video #3, which has a total duration of 13 seconds, 41 annotations were done for the gaze: towards left (12% of the duration of the segment), towards right (45%). In order to simplify the specification of the behavior to be replayed by the ECA, the gazes which were not directed towards left or right were grouped into a single class of gazes towards the camera for 43% of the segment's duration.

6.3.3. *Generation of expressive parameters for the gestures*

Step 6 aims at generating expressive animation. Five gestures were annotated for the third segment. Gesture quality was annotated as follows: fluidity (79% of the gesture annotations were perceived as being smooth, and 21% as being jerky), power (soft = 10%, hard = 21%, normal = 69%), speed (fast = 100%), spatial extent (contracted = 100%). These annotations are used to compute the values of the expressive parameters of the expressive agent. For example, in the Greta agent, the values of the fluidity (FLT) parameter have to be between -1 (jerky) and $+1$ (smooth). Thus, we computed the value of the FLT parameter for the third segment of video #3 (Table 2 provides the values of the expressivity parameters for the whole video) as follows: $\mathrm{FLT} = -1 \times 0.21 + 1 \times 0.79 = 0.58$. This computation enables us to set the fluidity of the generated gestures to an average value which represents the perception of global distribution of smooth versus jerky gestures.

7. Conclusions and Future Directions

We have presented a model of multimodal complex emotions involving gesture expressivity and blended facial expressions. We have described a methodology based on the manual annotation of a video corpus to create expressive ECAs via an analytical approach; we have proposed a representation scheme and a computational model for such an agent. We explained how the multi-level annotation of TV interviews is compatible with the multi-level specifications of our ECA. Our approach is at an exploratory stage and does not currently include the computation of statistics over a large number of videos. Yet, it did enable us to identify the relevant levels of representation for studying the complex relation between emotions and multimodal behaviors in non-acted and non-basic emotions. Whereas the first part of the model focuses on gesture expressivity, the second part of the model addresses how such complex emotions can impact on the display of superposed or masked facial expressions. Currently, we do not use all the annotations provided by the EmoTV corpus. The manual annotations of intensity are not considered yet: we only differentiate between major and minor labels. These annotations of intensity could be involved in the computation of the vector of emotion labels which is used for generating the emotional behavior of the ECA. The context annotations include other information related to "appraisal" dimensions such as the time of the event, the implication of the person, etc. which might be interesting to consider in the model of the agent. Other levels might also be relevant (head movements) so as to generate different behaviors with different levels of fidelity.

In the near future, we aim to perform perceptual tests to evaluate our methodology as well as our model of blend of facial expressions. We believe that the results of the two perceptual tests that we have described in this paper will be used to improve the copy-synthesis approach and specify other perceptual tests evaluating if the contextual cues, the emotion and the multimodal behaviors are perceptually equivalent in the original video and in the simulation of the corresponding behaviors by the ECA, thus revealing how much such a technique is successful. These perceptual tests will also help finding out if differences of quality and of level of details between the real and the simulated multimodal behaviors have an impact on the perception of emotion. For example, we currently compute average values for expressivity parameters and we do not specify precisely which gestures are to be performed by the ECA and with which expressive characteristics. Another application of these tests that we foresee is the possibility to refine our ECA system. Indeed, having to reproduce complex real behaviors allows us to refine our behavioral engine; we will apply the methodology *learning by imitation*. The corpus will also enable us to compute other relations between (i) the multimodal annotations, and (ii) the annotation of emotions (labels, intensity and valence), and the global annotations such as the modalities in which activity was perceived as relevant to emotion.[39] We are considering the use of image processing approaches in order to validate the manual annotations. Finally, we intend to extend the part of our model on complex facial expressions to include the combination of the expressivity parameters of the blended emotions. This will enable us to deal with the masked behaviors observed in our corpus and apply the copy-synthesis approach that we have defined for gesture. Indeed, in the video #41, a lady masks her disappointment with a tense smile. This could be modeled by blending the smile of the faked happiness and the tenseness of the felt disappointment.

Complex emotions are common in everyday conversation. Display rules, lies, and social context often lead to the combination of emotions as those observed in our corpus. We believe that the methodology that we have described might be useful with other real-life situations than TV interviews.

Acknowledgments

This work was partially supported by the Network of Excellence HUMAINE (Human-Machine Interaction Network on Emotion) IST-2002-2.3.1.6/Contract no. 507422 (http://emotion-research.net/). The authors are very grateful to Susanne Kaiser, Bjoern Hartmann, Maurizio Mancini, and Sarkis Abrilian for their suggestions and help.

References

1. P. Ekman and W. V. Friesen, *Unmasking the Face. A Guide to Recognizing Emotions from Facial Clues* (Prentice-Hall Inc., Englewood Cliffs, N J, 1975).
2. M. Schröder, Experimental study of affect burst, *Speech Commun.* **40**(1–2), 99–116 (2003) [Special Issue following the ISCA Workshop on Speech and Emotion].
3. R. Cowie, Emotional states expressed in speech, in *ISCA ITRW on Speech and Emotion: Developing a Conceptual Framework for Research* (2000), pp. 224–231.
4. R. Banse and K. Scherer, Acoustic profiles in vocal emotion expression, *J. Pers. Soc. Psychol.* **70**(3), 614–636 (1996).
5. M. DeMeijer, The contribution of general features of body movement to the attribution of emotions, *J. Nonverbal Behav.* **13**, 247–268 (1989).
6. J. Newlove, *Laban for Actors and Dancers* (Routledge, New York, 1993).
7. R. T. Boone and J. G. Cunningham, Children's decoding of emotion in expressive body movement: The development of cue attunement, *Dev. Psychol.* **34**(5), 1007–1016 (1998).
8. H. G. Wallbott, Bodily expression of emotion, *Eur. J. Soc. Psychol.* **28**, 879–896 (1998).

9. A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis and P. F. Driessen, Gesture-based affective computing on motion capture data, in *1st Int. Conf. Affective Computing and Intelligent Interaction (ACII'2005)* (Springer, Beijing, China, 2005), pp. 1–8.
10. K. R. Scherer, Analyzing emotion blends, in *Proc. 10th Conf. Int. Soc. for Research on Emotions* (Fischer, A., Würzburg, Germany, 1998), pp. 142–148.
11. L. Devillers, L. Vidrascu and L. Lamel, Emotion detection in real-life spoken dialogs recorded in call center, *J. Neural Network.* **18**(4), 407–422 (2005) [Special issue: Emotion and Brain].
12. V. P. Richmond and J. C. Croskey, *Non Verbal Behavior in Interpersonal Relations* (Allyn & Bacon Inc., 1999).
13. P. Ekman and W. Friesen, Felt, false, miserable smiles, *J. Nonverbal Behav.* **6**(4), 238–251 (1982).
14. M. Wiggers, Jugments of facial expressions of emotion predicted from facial behavior, *J. Nonverbal Behav.* **7**(2), 101–116 (1982).
15. I. S. Pandzic and R. Forchheimer, *MPEG-4 Facial Animation. The Standard, Implementation and Applications* (John Wiley & Sons, LTD, 2002).
16. M. Rehm and E. André, Catch me if you can — Exploring lying agents in social settings, in *Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS'2005)* (2005), pp. 937–944.
17. J. N. Bassili, Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face, *J. Pers. Soc. Psychol.* **37**(11), 2049–2058 (1979).
18. K. Gouta and M. Miyamoto, Emotion recognition, facial components associated with various emotions, *Shinrigaku Kenkyu* **71**(3), 211–218 (2000).
19. E. Constantini, F. Pianesi and M. Prete, Recognizing emotions in human and synthetic faces: The role of the upper and lower parts of the face, in *Intelligent User Interfaces (IUI'05)* (San Diego, CA, USA, 2005), pp. 20–27.
20. J. T. Cacioppo, R. P. Petty, M. E. Losch and H. S. Kim, Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions, *J. Pers. Soc. Psychol.* **50**, 260–268 (1986).
21. J. Ostermann, Animation of synthetic faces in MPEG-4, in *Computer Animation'98* (Philadelphia, USA, 1998), pp. 49–51.
22. I. Albrecht, M. Schröder, J. Haber and H.-P. Seidel, Mixed feelings: Expression of non-basic emotions in a muscle-based talking head, *J. Virtual Reality* **8**(4), 201–212 (2005) [Special issue on "Language, Speech & Gesture"].
23. N. Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Cowie and E. Douglas-Cowie, Emotion recognition and synthesis based on MPEG-4 FAPs, in MPEG-4 Facial Animation, eds. I. S. Pandzic and R. Forchheimer (John Wiley & Sons, UK, 2002), pp. 141–167.
24. Z. Ruttkay, H. Noot and P. ten Hagen, Emotion disc and emotion squares: Tools to explore the facial expression face, *Comp. Graph. Forum* **22**(1), 49–53 (2003).
25. T. Duy Bui, Creating emotions and facial expressions for embodied agents, Department of Computer Science (University of Twente, Enschede, 2004).
26. P. R. De Silva, A. Kleinsmith and N. Bianchi-Berthouze, Towards unsupervised detection of affective body posture nuances, in *1st Int. Conf. Affective Computing and Intelligent Interaction (ACII'2005)* (Springer, Beijing, China, 2005), pp. 32–40.
27. H. Gunes and M. Piccardi, Fusing face and body display for bi-modal emotion recognition: Single frame analysis and multi-frame post integration, in *1st Int. Conf. Affective Computing and Intelligent Interaction (ACII'2005)* (Springer, Beijing, China, 2005), pp. 102–110.

28. R. el Kaliouby and P. Robinson, Generalization of a vision-based computational model of mind-reading, in *1st Int. Conf. Affective Computing and Intelligent Interaction (ACII'2005)* (Springer, Beijing, China, 2005), pp. 582–590.
29. S. M. Choi and Y. G. Kim, An affective user interface based on facial expression recognition and eye-gaze tracking, in *1st Int. Conf. Affective Computing and Intelligent Interaction (ACII'2005)* (Springer, Beijing, China, 2005), pp. 907–915.
30. E. Douglas-Cowie, N. Campbell, R. Cowie and P. Roach, Emotional speech: Towards a new generation of databases, *Speech Commun.* **40**, 33–60 (2003).
31. Y. Cao, P. Faloutsos, E. Kohler and F. Pighin, Real-time speech motion synthesis from recorded motions, in *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Grenoble, France (2004), pp. 347–355.
32. A. Egges, S. Kshirsagar and N. Magnenat-Thalmann, Imparting individuality to virtual humans, in *Int. Workshop Virtual Reality Rehabilitation,* Lausanne, Switzerland (2002), pp. 201–108.
33. I. S. Pandzic, Facial Motion Cloning, *Graph. Models* **65**(6), 385–404 (2003).
34. M. Kipp, *Gesture Generation by Imitation. From Human Behavior to Computer Character Animation* (Boca Raton, Dissertation.com, Florida, 2004).
35. J. Cassell, M. Stone and Y. Hao, Coordination and context-dependence in the generation of embodied conversation, in *1st Int. Natural Language Generation Conf. (INLG'2000)*, Mitzpe Ramon, Israel (2000), pp. 171–178.
36. S. Abrilian, L. Devillers, S. Buisine and J.-C. Martin, EmoTV1: Annotation of real-life emotions for the specification of multimodal affective interfaces, in *11th Int. Conf. Human-Computer Interaction (HCII'2005),* LEA, Las Vegas, Nevada, USA (2005).
37. M. Kipp, Anvil — A generic annotation tool for multimodal dialogue, in *7th Eur. Conf. Speech Communication and Technology (Eurospeech'2001),* Aalborg, Danemark (2001), pp. 1367–1370.
38. E. Douglas-Cowie, L. Devillers, J.-C. Martin, R. Cowie, S. Savvidou, S. Abrilian and C. Cox, Multimodal databases of everyday emotion: Facing up to complexity, in *9th Eur. Conf. Speech Communication and Technology (Interspeech'2005),* Lisbon, Portugal, (2005), pp. 813–816.
39. L. Devillers, S. Abrilian and J.-C. Martin, Representing real life emotions in audiovisual data with non basic emotional patterns and context features, in *1st Int. Conf. Affective Computing and Intelligent Interaction (ACII'2005)* (Spinger-Verlag, Berlin, Beijing, China, 2005), pp. 519–526.
40. S. Abrilian, L. Devillers and J.-C. Martin, Annotation of emotions in real-life video interviews: Variability between coders, in *5th Int. Conf. Language Resources and Evaluation (LREC'2006)*, Genoa, Italy (2006).
41. D. McNeill, *Hand and Mind — What Gestures Reveal about Thoughts* (University of Chicago Press, IL, 1992).
42. A. Kendon, *Gesture: Visible Action as Utterance* (Cambridge University Press, 2004).
43. I. Poggi, Mind markers, in *Gestures. Meaning and Use*, eds. M. Rector, I. Poggi and N. Trigo (Edicoes Universidade Fernando Pessoa, Oporto, 2003), pp. 119–132.
44. S. Kaiser and T. Wehrle, Facial expressions as indicators of appraisal processes, in *Appraisal Theories of Emotions: Theories, Methods, Research*, eds. K. R. Scherer, A. Schorr and T. Johnstone (Oxford University Press, New York, 2001), pp. 285–300.
45. H. G. Wallbott and K. R. Scherer, Cues and channels in emotion recognition, *J. Pers. Soc. Psychol.* **51**(4), 690–699 (1986).
46. P. Gallaher, Individual differences in nonverbal behavior: Dimensions of style, *J. Pers. Soc. Psychol.* (63), 133–145 (1992).

47. S. Kopp and I. Wachsmuth, Synthesizing multimodal utterances for conversational agents, *J. Comp. Animat. Virtual Worlds* **15**(1), 39–52 (2004).
48. S. Kopp, P. Tepper and J. Cassell, Towards integrated microplanning of language and iconic gesture for multimodal output, in *Int. Conf. Multimodal Interfaces (ICMI'04),* Penn State University, State College, PA (2004), pp. 97–104.
49. B. Hartmann, M. Mancini and C. Pelachaud, Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis, in *Computer Animation (CA'2002)* (IEEE Computer Society, Geneva, Switzerland, 2002), pp. 111–119.
50. B. Hartmann, M. Mancini and C. Pelachaud, Implementing expressive gesture synthesis for embodied conversational agents, in *Gesture Workshop (GW'2005)*, Vannes, France (2005).
51. C. Pelachaud, Multimodal expressive embodied conversational agent, in *ACM Multimedia, Brave New Topics Session* (ACM, Singapore, 2005), pp. 683–689.
52. P. Ekman, *The Face Revealed* (Weidenfeld & Nicolson, London, 2003).
53. D. Keltner, Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame, *J. Pers. Soc. Psychol.* **68**, 441–454 (1995).
54. D. Matsumoto, More evidence for the universality of a contempt expression, *Motiv. Emotion* **16**, 363–368 (1992).
55. B. Bouchon-Meunier, M. Rifqi and S. Bothorel, Towards general measures of comparison of objects, *Fuzzy Set. Sys.* **84**, 143–153 (1996).
56. D. Dubois and H. Prade, *Fuzzy Sets and Systems* (Academic Press, New York, 1980).
57. P. Ekman and W. V. Friesen, The repertoire of nonverbal behavior's: Categories, origins, usage and coding, *Semiotica* **1**, 49–98 (1969).
58. P. Ekman, Darwin, deception, and facial expression, *Ann. NY Acad. Sci.* **1000**, 205–221 (2003).
59. E. Hüllermeier, D. Dubois and H. Prade, Fuzzy rules in case-based reasoning, in *Conf. AFIA-99 Raisonnement à Partir de Cas*, Paris, France (1999), pp. 45–54.
60. P. Ekman and W. V. Friesen, *Manual for the Facial Action Coding System* (Consulting Psychology Press, Palo Alto, CA, 1978).
61. S. Prillwitz, R. Leven, H. Zienert, T. Hanke and J. Henning, Hamburg notation system for sign languages: An introductory guide, in *International Studies on Sign Language and Communication of the Deaf* (Signum Press, Hamburg, Germany, 1989).
62. B. De Carolis, C. Pelachaud, I. Poggi and M. Steedman, APML, a markup language for believable behavior generation, in Life-Like Characters. *Tools, Affective Functions and Applications*, eds. H. Prendinger and M. Ishizuka (Springer, 2004), pp. 65–85.

Jean-Claude Martin received his Ph.D. degree in computer science in 1995 from the Ecole Nationale Supérieure des Télécommunications. Since 1999, he has been as Associate Professor at LIMSI, CNRS. His research focuses on the study of cooperation between modalities both in human communication and HCI.

Radoslaw Niewiadomski received his M.S. degree in computer science from the Adam Mickiewicz University, Poznan, Poland in 2001. He is a Ph.D. student at the University of Perugia, Italy.

Laurence Devillers received her Ph.D. degree in computer science from the University of Paris-Sud, France, in 1992. Since 1995, she has been an Associate Professor at the University of Paris-Sud, and a member of the LIMSI-CNRS Spoken Language Processing research group.

Stéphanie Buisine received her Ph.D. degree in Cognitive Psychology and Ergonomics from the University of Paris 5 in 2005. She currently holds the position of Associate Researcher in a higher-engineering institute in Paris (Ecole Nationale Supérieure d'Arts et Métiers).

Catherine Pelachaud has been a Professor at the University of Paris 8, IUT of Montreuil since 2002. Her research interests include representation language for agents, embodied conversational agents, nonverbal communication (face, gaze, and gesture), expressive gesture and multimodal interfaces.