



**HAL**  
open science

## Multimodal complex emotions: gesture expressivity and blended facial expressions

Jean-Claude Martin, Radoslaw Niewiadomski, Laurence Devillers, Stéphanie Buisine, Catherine Pelachaud

► **To cite this version:**

Jean-Claude Martin, Radoslaw Niewiadomski, Laurence Devillers, Stéphanie Buisine, Catherine Pelachaud. Multimodal complex emotions: gesture expressivity and blended facial expressions. International Journal of Humanoid Robotics, 2006, pp.3, 1-23. 10.1142/S0219843606000825 . hal-00787584

**HAL Id: hal-00787584**

**<https://hal.science/hal-00787584v1>**

Submitted on 15 Feb 2013

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1

3

5

## MULTIMODAL COMPLEX EMOTIONS: GESTURE EXPRESSIVITY AND BLENDED FACIAL EXPRESSIONS

7

JEAN-CLAUDE MARTIN

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

9

RADOSLAW NIEWIADOMSKI

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

11

13

LAURENCE DEVILLERS

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

15

17

STEPHANIE BUISINE

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

19

21

CATHERINE PELACHAUD

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

23

25

27

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.

29

31

33

35

37

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

## 1 **1. Introduction**

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

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

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

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

40 The remaining parts of this paper are structured as follows. Section 2 summarizes  
41 some of the studies on complex emotions, gesture expressivity, and facial expres-  
42 sions. Section 3 provides two examples of gesture and facial expression behaviors  
43 observed during complex emotions in videos of TV interviews. Section 4 describes

1 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.  
3 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  
5 Greta agent.<sup>1</sup>

## 2. Related Work

7 There has been a lot of psychological research on emotion and nonverbal  
communication in facial expressions,<sup>1</sup> vocal expressions<sup>2-4</sup> and expressive body  
9 movements.<sup>5-8</sup> Yet, these psychological studies were based mostly on acted basic  
emotions: anger, disgust, fear, joy, sadness, surprise. In the area of affective comput-  
11 ing, 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  
13 placed on various points of the whole body in order to recognize four acted basic  
emotions (sadness, joy, anger, fear).<sup>9</sup> Some studies deal with more complex emo-  
15 tions. 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 emo-  
17 tional state.<sup>10</sup> Scherer and his colleagues show in this experiment that some events  
may give rise to several simultaneous emotions. These emotions are referred to as  
19 complex emotions and also as blends of emotions.<sup>1,10,11</sup> They may occur either as a  
quick succession of different emotions, the superposition of emotions, the masking  
21 of one emotion by another one, the suppression of one emotion or the overacting of  
one emotion.

23 In particular, in the visual modalities, these blends produce “multiple simul-  
taneous facial expressions.”<sup>12</sup> Depending on the type of blending, the resulting  
25 facial expressions are not identical. A masked emotion may leak over the displayed  
emotion,<sup>1</sup> while superposition of two emotions will be shown by different facial fea-  
27 tures (one emotion being shown on the upper face while another one on the lower  
face).<sup>1</sup> Perceptual studies have shown that people are able to recognize facial expres-  
29 sion of felt emotion<sup>13,14</sup> as well as fake emotion.<sup>13</sup> Similar studies producing similar  
results have been conducted on ECAs.<sup>15</sup> In a study on a deceiving agent, Rhem and  
31 André found that the users were able to differentiate when the agent was displaying  
expression of felt emotion or expression of fake emotion.<sup>16</sup> Aiming at understand-  
33 ing if facial features or regions play identical roles in emotion recognition, Bassili<sup>17</sup>  
and later on Gouta and Miyamoto,<sup>18</sup> and Constantini *et al.*<sup>19</sup> performed various  
35 perceptual tasks, and Cacioppo *et al.*<sup>20</sup> studied psychological facial activity. They  
found that positive emotions are mainly perceived from the expression of the lower  
37 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 devel-  
39 oped so far for ECAs. The interpolation between facial parameters of given expres-  
sions is commonly used to compute the new expression. MPEG-4 proposes to create  
41 a new expression as a weighted interpolation of any of the six predefined expressions

1 of emotions.<sup>15,21</sup> More complex interpolation schemes have been proposed.<sup>22–24</sup>  
2 Duy Bui<sup>25</sup> introduced a set of fuzzy rules to determine the blended expressions of  
3 the six basic emotions. In this approach, a set of fuzzy rules is attributed to each  
4 pair of emotions. The intensities of muscles contraction for the blended expres-  
5 sion are related to emotions intensities using fuzzy inference. With respect to other  
6 modalities than facial expressions, static postures were recorded by De Silva *et al.*  
7 using a motion capture system during acted emotions (two nuances for each of four  
8 basic emotions).<sup>26</sup> Gunes *et al.* fused the video processing of facial expression and  
9 upper body gestures in order to recognize six acted emotional behaviors (anxiety,  
10 anger, disgust, fear, happiness, uncertainty).<sup>27</sup> A vision-based system that infers  
11 acted mental states (agreeing, concentrating, disagreeing, interested, thinking, and  
12 unsure) from head movement and facial expressions was described by el Kaliouby  
13 *et al.*<sup>28</sup> Choi *et al.* described how video processing of both facial expressions and  
14 gaze are mapped onto combinations of seven emotions.<sup>29</sup> Yet, real-life multimodal  
15 corpora are indeed very few despite the general agreement that it is necessary to  
16 collect audio-visual databases that highlight naturalistic expressions of emotions as  
17 suggested by Douglas-Cowie *et al.*<sup>30</sup>

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

### 29 **3. Complex Emotions: Two Illustrative Examples**

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

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

## 4. Representing, Modeling and Evaluating Expressivity

### 15 4.1. *Representing expressivity*

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

### 37 4.2. *Evaluation of the gesture expressivity model*

38 We have conducted two studies to evaluate our gesture expressivity model which  
 39 is the central part of the copy-synthesis approach described in Sec. 6. These two

1 studies involved a total number of 106 users (80 males, 26 females; aged 17 to 25).  
2 All were first and second year French university students. Each user completed only  
3 one of the two tests. Both tests consisted in observing sets of video clips (two per  
4 trial for the first test, four for the second test) and answering a questionnaire. The  
5 video clips differ only on the gesture expressivity of the agent (same audio and same  
6 gesture type).

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

24 The hypothesis tested in the second study was the following: combining param-  
25 eters in such a way that they reflect a given communicative intent will result in a  
26 more believable overall impression of the agent. Avoiding behavior qualities that  
27 imply an emotional state or a personality trait, we considered the three following  
28 qualities: abrupt, sluggish, and vigorous. Abrupt is characterized by rapid, discon-  
29 tinuous and powerful movements. Sluggish is characterized by slow, effortless and  
30 close to the body but fluid movements. Vigorous is characterized by a lot of large,  
31 fast, fluid and repetitive movements. For each quality, we generated four anima-  
32 tions. One animation corresponds to the neutral, generic animation, two to variants  
33 of the chosen expressive intent (strongly and slightly expressive) and one to an  
34 opposite assignment of expressivity parameters. This test ( $N = 54$ ) was conducted  
35 as a preference ranking task: the user had to order four animations from the most  
36 appropriate to the least appropriate with respect to the expressive intent. For the  
37 abrupt and vigorous qualities, users preferred the coherent performances as we had  
38 hoped [ $F(3/153) = 31.23$ ,  $p < 0.001$  and  $F(3/153) = 104.86$ ,  $p < 0.001$ , respec-  
39 tively]. The relation between our parameterization and users' perception can also  
40 be expressed as a linear correlation, which amounts to +0.655 for the abrupt quality  
41 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.

Modified parameter	Perceived modification							Total	
	Spatial extent	Temporal extent	Fluidity	Power	Repetition	Overall activation	No modification		Do not know
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



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

6 In the first test, we checked if subjects perceived variation of each parameter,  
7 while in the second perceptual test we looked at the interpretation of these varia-  
8 tions. Since our expressivity parameters show some dependency with one another,  
9 we wanted to check that the subject perceived their individual changes and their  
10 combined meaning in two separate perceptual tests. The results confirm that our  
11 general approach for expressivity modeling is worthwhile pursuing. A notable advan-  
12 tage of our implementation is to enable the decomposition of gesture expressivity  
13 and the test of parameters one by one. In the experiment by Wallbott, actors were  
14 instructed to act basic emotions.<sup>8</sup> This experiment revealed that each acted emo-  
15 tion had an impact on all the parameters of expressivity. The first perceptual test  
16 we conducted would have been surely more difficult to control with a human actor  
17 instead of an agent: humans may be able to control their expressivity to a certain  
18 extent but can hardly isolate each parameter. In our animations, the decomposi-  
19 tion of expressivity may have produced artificial behaviors but this step seemed  
20 necessary to evaluate our model and highlight possible ways of improvement. These  
21 results will be used to refine the technical implementation of individual parameters  
22 to achieve higher quality animation and better visibility of changes to the param-  
23 eters. For the second perceptual test, we were careful to avoid introducing labels  
24 related to personality or emotion. While we ultimately want to simulate such traits  
25 and mental states, the link from these high-level concepts to the expressive dimen-  
26 sions is still not clear — the social psychology literature on this problem appears to  
27 be very sparse. This second test mainly showed that we need to integrate gesture  
28 selection and gesture modification when generating an animation. A shortcoming  
29 of the current test was that only a single utterance with a unique gesture selection  
30 was used with varying animations. A wider variety of different utterances and corre-  
31 sponding gesture selections is needed to understand the perception of expressivity.

## 5. Representing and Modeling Blended Facial Expressions

33 In this section, we present a computational model of facial expressions arising from  
34 blends of emotions. Instead of formulating our model at the level of facial muscle  
35 contractions or FAP values, we propose a face partition based model, which not  
36 only computes the complex facial expressions of emotions but also distinguishes  
37 between different types of blending. Blends (e.g. superposition and masking) are  
38 distinguished among each other as they are usually expressed by different facial  
39 areas.<sup>1,52</sup> Expressions may also occur in rapid sequences one after the other. More-  
40 over, the expression of masking a felt emotion by a fake one (i.e. not felt) is different  
41 from the expression corresponding to the superposition of two felt emotions.<sup>1</sup> Thus

1 complex facial expressions can be distinguished depending on the type of emotions,  
 2 their apparition in time (sequence, superposition) as well as if the emotions are  
 3 felt or fake. For the moment, we have considered only two cases of complex facial  
 4 expressions: the superposition of two felt emotions and the masking of a felt emo-  
 5 tion with a fake one. In the following sub-section, we present a general framework  
 6 for our model and describe next details of computational procedures based on fuzzy  
 7 inference.

### 5.1. Blend of emotions

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

27 Let  $E_i$  and  $E_j$  be the emotions occurring in a blend and  $\text{Exp}_{\text{blend}}(E_i, E_j)$   
 28 the resulting complex expression, where *blend* is either masking (*M*) or superpo-  
 29 sition (*S*). The  $\text{Exp}_{\text{blend}}(E_i, E_j)$  is also composed by the combination of  $n$  dif-  
 30 ferent face areas, where each  $F_k^{(E_i, E_j)}$  is equal to one corresponding area from  
 31  $\text{Exp}(E_i)$ ,  $\text{Exp}(E_j)$ ,  $\text{Exp}(N)$ . We note, that for any  $k$  in the interval  $[1, n]$ ,  $F_k^{(E_i, E_j)}$   
 32 cannot contain simultaneously elements of two different expressions; it can be either  
 33  $\text{Exp}(E_i)$ ,  $\text{Exp}(E_j)$ , or  $\text{Exp}(N)$ . That is, a facial area can not show different expres-  
 34 sions at the same time; it can show one expression at a time: this expression can  
 35 come from either emotion or the neutral expression. Combining facial expressions on  
 36 the same facial area can have the artefact to introduce a new expression. For exam-  
 37 ple, if we add the facial actions in the eyebrow region of surprise “raise-eyebrow”  
 38 and of anger “frown” we obtain a new facial action “upper-raised-eyebrow-down”  
 39 that is typically linked to fear. Thus, we opt for the rules that no facial action can  
 40 be added up on a same facial region. This ensures the conformity of our model with  
 41 empirical evidence.<sup>1</sup>

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

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

## 5.2. Masking

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

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

1 and  $E_u$ . Each input variable corresponds to one basic emotion  $E_u$ ,  $u = 1, \dots, 6$ , and  
 each output variable corresponds to one facial region  $F_k$  of the resulting expression.  
 3 In particular, each rule of  $SFR_{\text{felt}}$  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*  
 5 *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  
 $SFR_{\text{felt}}$  by the following rule:

If  $X$  is *SADNESS* then  $F_{\text{brows}}$  is *VISIBLE* and  $F_{\text{upper eyelids}}$  is *VISIBLE* and  
 $\dots$  and  $F_{\text{upper lip}}$  is *NOT-VISIBLE* and  $F_{\text{lower lip}}$  is *NOT-VISIBLE*,

where  $X$  expresses degree of similarity to  $\text{Exp}(\text{SADNESS})$  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  $SFR_{\text{fake}}$  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  $SFR_{\text{fake}}$ :

If  $X$  is *JOY* then  $F_{\text{brows}}$  is *VISIBLE* and  $F_{\text{upper eyelids}}$  is *VISIBLE* and  
 $F_{\text{lower eyelids}}$  is *NOT-VISIBLE* and  $\dots$  and  $F_{\text{upper lip}}$  is *VISIBLE* and  
 $F_{\text{lower lip}}$  is *VISIBLE*.

7 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  
 9 if  $\text{Exp}(E_i)$  and/or  $\text{Exp}(E_j)$  is different from  $\text{Exp}(E_u)$ ,  $u = 1, \dots, 6$ , the  
 model predicts the final expression based on the degree of similarity between  
 11  $\text{Exp}(E_i)$  and/or  $\text{Exp}(E_j)$  and basic expressions. The fake and felt areas of the mask-  
 ing expression are considered separately. Finally, for each  $F_k$ , the results of  $SFR_{\text{felt}}$   
 13 and of  $SFR_{\text{fake}}$  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*.  
 15 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  
 17 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  
 19 (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  
 21 expression. We can notice that the absence of orbicularis oculi activity as indicator  
 of unfelt joy<sup>58</sup>. is visible on both images (c) and (d), the annotated video and the  
 23 corresponding Greta simulation.

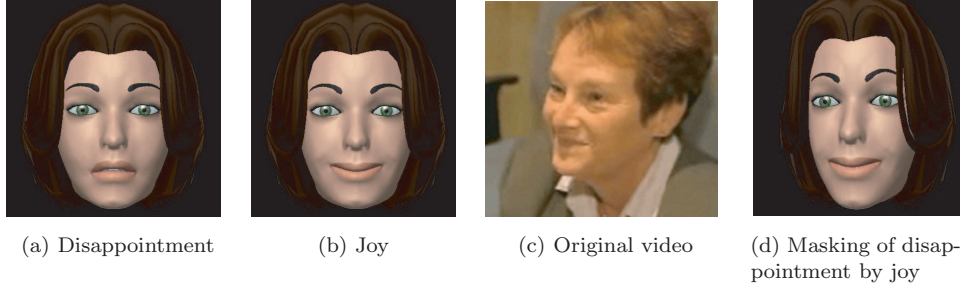


Fig. 1. Disappointment masked by joy.

### 1      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  $\text{Exp}_S(E_i, E_j)$  resulting from the superposition of  $E_i$  and  $E_j$  is equal to the superposition of  $E_j$  and  $E_i$ . That is:  $\text{Exp}_S(E_i, E_j) = \text{Exp}_S(E_j, E_i)$ . Ekman described this case of blending for all pairs of the six basic emotions.<sup>1</sup> 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)<sup>17–19</sup> Let  $Z$  be a set of plausible (according to Ekman) *schemas* for the superposition expression  $\text{Exp}_S$ . By “schema” we intend the particular division of  $n$  face regions  $F_k$ ,  $k = 1, \dots, n$  between any two emotions. At the moment, ten different schemas are considered. The fuzzy inference is used to model the combination of facial expressions  $\text{Exp}(E_i)$  and  $\text{Exp}(E_j)$  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:

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

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

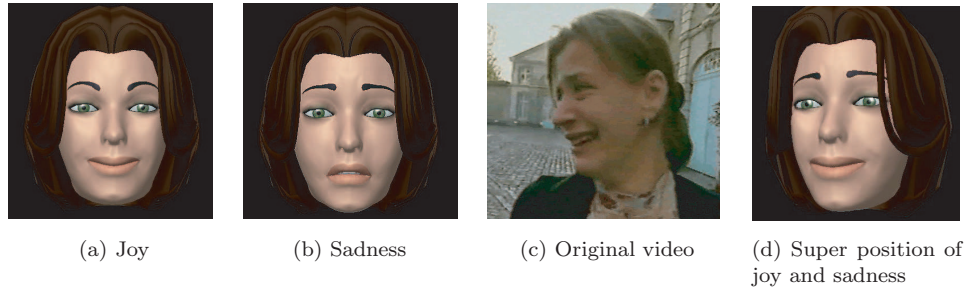


Fig. 2. Superposition of sadness and joy.

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

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

## 6. Copy-Synthesis Approach

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

### 6.1. Annotation

25 Annotation is composed of two steps. Step 1 aims at the automatic annotation of the  
 26 video with data that can be useful either for the manual annotation of the video or  
 27 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

1 achieved following the LDC norms for hesitations, breath, etc. The video is then  
 2 annotated at several temporal levels (whole video, segments of the video, behaviors  
 3 observed at specific moments) and at several levels of abstraction. The global behavior  
 4 observed during the whole video is annotated with communicative act, emotions  
 5 and multimodal cues. The segments are annotated with emotion labels and the

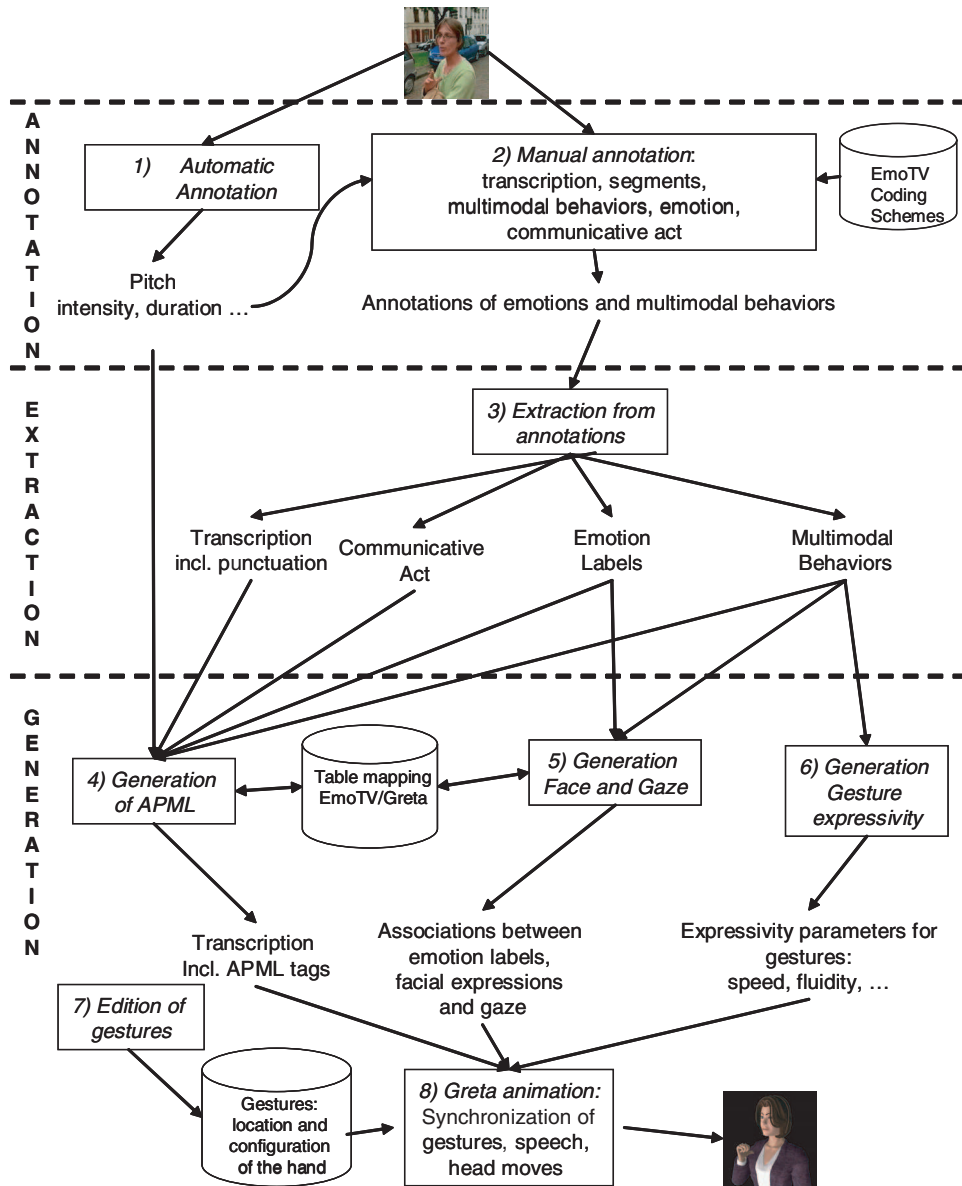


Fig. 3. Copy synthesis approach for studying gesture expressivity during emotions.

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

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

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

29 Whereas the annotations of emotions have been done by three coders and lead  
 30 to computation of agreement,<sup>39</sup> the current protocol used for the validation of the  
 31 annotations of multimodal behaviors is to have a second coder checks the annota-  
 32 tions done by a first coder followed by brainstorming discussions. We are currently  
 33 considering the validation of the annotations by the automatic computation of inter-  
 34 coder agreements from the annotations by several coders.

## 35 **6.2. Extraction from annotations**

36 A module has been designed for extracting from the various annotations the pieces of  
 37 information which have been identified as required for generation (Step 3 in Fig. 3):  
 38 the speech transcription, the communicative act, the emotion labels, the dimensions  
 39 of emotions, the multimodal behaviors (including the number of occurrences and the  
 40 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
	<hr/>			
GESTURE	% fast vs. % slow	47% vs. 3%	33% vs. 13%	83% vs. 0%
EXPRESSIVITY	% 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%

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

### 6.3. Generation

9 Our ECA system, Greta, incorporates communicative conversational and emotional  
 10 qualities.<sup>51</sup> The agent’s behavior is synchronized with her speech and is consistent  
 11 with the meaning of her sentences. To determine speech-accompanying non-verbal  
 12 behaviors, the system relies on a taxonomy of communicative functions proposed by  
 13 Isabella Poggi.<sup>43</sup> A communicative function is defined as a pair (meaning, signal)  
 14 where *meaning* corresponds to the communicative value the agent wants to commu-  
 15 nicate and *signal* to the behavior used to convey this meaning. We have developed a  
 16 language to describe gesture signals in a symbolic form.<sup>49</sup> An arm gesture is described  
 17 by its wrist position, palm orientation, finger direction as well as hand shape. We use  
 18 the HamNoSys system to encode hand shapes.<sup>61</sup> To control the agent’s behavior, we  
 19 are using the APML representation language, where the tags of this language are  
 20 the communicative functions.<sup>62</sup> The system takes as input a text tagged with APML  
 21 labels as well as values for the expressivity dimensions that characterize the manner  
 22 of execution of the agent’s behaviors. The system parses the input text and selects  
 23 which behaviors to perform. Facial expressions and gaze behaviors are synchronized  
 24 with speech defined within APML tags. The system looks for the emphasis word. It  
 25 aligns the stroke of a gesture with this word. Then it computes when the preparation  
 26 phase of the gesture is as well as if a gesture is hold, if it co-articulates to the next  
 27 one, or if it returns to the rest position. The expressivity model controls the spatial

1 and dynamism properties of the gestures. The outputs of the system are animation  
and audio files that drive the animation of the agent.

### 3 6.3.1. *Generation of the APML file*

Step 4 consists of generating the APML file used by the Greta system from the  
5 data extracted from the annotations such as the speech transcription, the pitch,  
the communicative act, and the emotion labels. The transcription is directly used  
7 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  
9 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  
11 closest performative the Greta system knows about. Thus the communicative goal  
“to complain” used for annotating the video #3 is translated to the performative  
13 “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  
15 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.  
17 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  
19 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  
21 is specified as this combination of the two major emotion labels computed from the  
emotional profiles resulting from the annotations (Table 2).

### 23 6.3.2. *Generation of gaze behaviors*

The annotations of facial expressions are used in Step 5 to associate the combined  
25 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  
27 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  
29 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,  
31 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.

### 33 6.3.3. *Generation of expressive parameters for the gestures*

Step 6 aims at generating expressive animation. Five gestures were annotated for  
35 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  
37 (contracted = 100%). These annotations are used to compute the values of the

1 expressive parameters of the expressive agent. For example, in the Greta agent,  
2 the values of the fluidity (FLT) parameter have to be between  $-1$  (jerky) and  $+1$   
3 (smooth). Thus, we computed the value of the FLT parameter for the third segment  
4 of video #3 (Table 2 provides the values of the expressivity parameters for the whole  
5 video) as follows:  $FLT = -1 \times 0.21 + 1 \times 0.79 = 0.58$ . This computation enables us  
6 to set the fluidity of the generated gestures to an average value which represents  
7 the perception of global distribution of smooth versus jerky gestures.

## 7. Conclusions and Future Directions

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

30 In the near future, we aim to perform perceptual tests to evaluate our method-  
31 ology as well as our model of blend of facial expressions. We believe that the results  
32 of the two perceptual tests that we have described in this paper will be used to  
33 improve the copy-synthesis approach and specify other perceptual tests evaluating  
34 if the contextual cues, the emotion and the multimodal behaviors are perceptually  
35 equivalent in the original video and in the simulation of the corresponding behaviors  
36 by the ECA, thus revealing how much such a technique is successful. These percep-  
37 tual tests will also help finding out if differences of quality and of level of details  
38 between the real and the simulated multimodal behaviors have an impact on the  
39 perception of emotion. For example, we currently compute average values for expres-  
40 sivity parameters and we do not specify precisely which gestures are to be performed  
41 by the ECA and with which expressive characteristics. Another application of these

1 tests that we foresee is the possibility to refine our ECA system. Indeed, having  
 2 to reproduce complex real behaviors allows us to refine our behavioral engine; we  
 3 will apply the methodology *learning by imitation*. The corpus will also enable us  
 4 to compute other relations between (i) the multimodal annotations, and (ii) the  
 5 annotation of emotions (labels, intensity and valence), and the global annotations  
 6 such as the modalities in which activity was perceived as relevant to emotion.<sup>39</sup>  
 7 We are considering the use of image processing approaches in order to validate the  
 8 manual annotations. Finally, we intend to extend the part of our model on complex  
 9 facial expressions to include the combination of the expressivity parameters of the  
 10 blended emotions. This will enable us to deal with the masked behaviors observed in  
 11 our corpus and apply the copy-synthesis approach that we have defined for gesture.  
 12 Indeed, in the video #41, a lady masks her disappointment with a tense smile. This  
 13 could be modeled by blending the smile of the faked happiness and the tenseness  
 14 of the felt disappointment.

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

## 19 Acknowledgments

20 This work was partially supported by the Network of Excellence HUMAINE  
 21 (Human-Machine Interaction Network on Emotion) IST-2002-2.3.1.6/Contract no.  
 22 507422 (<http://emotion-research.net/>). The authors are very grateful to Susanne  
 23 Kaiser, Bjoern Hartmann, Maurizio Mancini, and Sarkis Abrilian for their sugges-  
 24 tions and help.

## 25 References

- 26 1. P. Ekman and W. V. Friesen, *Unmasking the Face. A Guide to Recognizing Emotions*  
 27 *from Facial Clues* (Prentice-Hall Inc., Englewood Cliffs, N J, 1975).
- 28 2. M. Schröder, Experimental study of affect burst, *Speech Commun.* **40**(1–2), 99–116  
 29 (2003) [Special Issue following the ISCA Workshop on Speech and Emotion].
- 30 3. R. Cowie, Emotional states expressed in speech, in *ISCA ITRW on Speech and Emo-*  
 31 *tion: Developing a Conceptual Framework for Research* (2000), pp. 224–231.
- 32 4. R. Banse and K. Scherer, Acoustic profiles in vocal emotion expression, *J. Pers. Soc.*  
 33 *Psychol.* **70**(3), 614–636 (1996).
- 34 5. M. DeMeijer, The contribution of general features of body movement to the attribution  
 35 of emotions, *J. Nonverbal Behav.* **13**, 247–268 (1989).
- 36 6. J. Newlove, *Laban for Actors and Dancers* (Routledge, New York, 1993).
- 37 7. R. T. Boone and J. G. Cunningham, Children’s decoding of emotion in expressive  
 38 body movement: The development of cue attunement, *Dev. Psychol.* **34**(5), 1007–  
 39 1016 (1998).
- 40 8. H. G. Wallbott, Bodily expression of emotion, *Eur. J. Soc. Psychol.* **28**, 879–896  
 41 (1998).

- 1     9. A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis and P. F. Driessen, Gesture-based  
3     affective computing on motion capture data, in *1st Int. Conf. Affective Computing and  
Intelligent Interaction (ACII'2005)* (Springer, Beijing, China, 2005), pp. 1–8.
- 5     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.
- 7     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: Emo-  
tion and Brain].
- 9     12. V. P. Richmond and J. C. Croskey, *Non Verbal Behavior in Interpersonal Relations*  
(Allyn & Bacon Inc., 1999).
- 11    13. P. Ekman and W. Friesen, Felt, false, miserable smiles, *J. Nonverbal Behav.* **6**(4),  
238–251 (1982).
- 13    14. M. Wiggers, Judgments of facial expressions of emotion predicted from facial behavior,  
*J. Nonverbal Behav.* **7**(2), 101–116 (1982).
- 15    15. I. S. Pandzic and R. Forchheimer, *MPEG-4 Facial Animation. The Standard, Imple-  
mentation and Applications* (John Wiley & Sons, LTD, 2002).
- 17    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)*  
19    (2005), pp. 937–944.
- 21    17. J. N. Bassili, Emotion recognition: The role of facial movement and the relative impor-  
tance of upper and lower areas of the face, *J. Pers. Soc. Psychol.* **37**(11), 2049–2058  
(1979).
- 23    18. K. Gouta and M. Miyamoto, Emotion recognition, facial components associated with  
various emotions, *Shinrigaku Kenkyu* **71**(3), 211–218 (2000).
- 25    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  
27    (IUI'05)* (San Diego, CA, USA, 2005), pp. 20–27.
- 29    20. J. T. Cacioppo, R. P. Petty, M. E. Losch and H. S. Kim, Electromyographic activ-  
ity over facial muscle regions can differentiate the valence and intensity of affective  
reactions, *J. Pers. Soc. Psychol.* **50**, 260–268 (1986).
- 31    21. J. Ostermann, Animation of synthetic faces in MPEG-4, in *Computer Animation'98*  
(Philadelphia, USA, 1998), pp. 49–51.
- 33    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  
35    (2005) [Special issue on “Language, Speech & Gesture”].
- 37    23. N. Tsapatsoulis, A. Raouzaoui, 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.
- 39    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).
- 41    25. T. Duy Bui, Creating emotions and facial expressions for embodied agents, Depart-  
ment of Computer Science (University of Twente, Enschede, 2004).
- 43    26. P. R. De Silva, A. Kleinsmith and N. Bianchi-Berthouze, Towards unsupervised detec-  
tion of affective body posture nuances, in *1st Int. Conf. Affective Computing and  
45    Intelligent Interaction (ACII'2005)* (Springer, Beijing, China, 2005), pp. 32–40.
- 47    27. H. Gunes and M. Piccardi, Fusing face and body display for bi-modal emotion recogni-  
tion: Single frame analysis and multi-frame post integration, in *1st Int. Conf. Affective  
49    Computing and Intelligent Interaction (ACII'2005)* (Springer, Beijing, China, 2005),  
pp. 102–110.

- 1 28. R. el Kaliouby and P. Robinson, Generalization of a vision-based computational model  
3 of mind-reading, in *1st Int. Conf. Affective Computing and Intelligent Interaction*  
(*ACII'2005*) (Springer, Beijing, China, 2005), pp. 582–590.
- 5 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*  
7 *Interaction (ACII'2005)* (Springer, Beijing, China, 2005), pp. 907–915.
- 9 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).
- 11 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*  
13 *Animation*, Grenoble, France (2004), pp. 347–355.
- 15 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.
- 17 33. I. S. Pandzic, Facial Motion Cloning, *Graph. Models* **65**(6), 385–404 (2003).
- 19 34. M. Kipp, *Gesture Generation by Imitation. From Human Behavior to Computer Character*  
21 *Animation* (Boca Raton, Dissertation.com, Florida, 2004).
- 23 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.
- 25 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.*  
27 *Human-Computer Interaction (HCII'2005)*, LEA, Las Vegas, Nevada, USA (2005).
- 29 37. M. Kipp, Anvil — A generic annotation tool for multimodal dialogue, in *7th Eur.*  
*Conf. Speech Communication and Technology (Eurospeech'2001)*, Aalborg, Denmark  
(2001), pp. 1367–1370.
- 31 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.*  
33 *Conf. Speech Communication and Technology (Interspeech'2005)*, Lisbon, Portugal,  
(2005), pp. 813–816.
- 35 39. L. Devillers, S. Abrilian and J.-C. Martin, Representing real life emotions in audio-  
visual data with non basic emotional patterns and context features, in *1st Int. Conf.*  
37 *Affective Computing and Intelligent Interaction (ACII'2005)* (Springer-Verlag, Berlin,  
Beijing, China, 2005), pp. 519–526.
- 39 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 Eval-*  
41 *uation (LREC'2006)*, Genoa, Italy (2006).
- 43 41. D. McNeill, *Hand and Mind — What Gestures Reveal about Thoughts* (University of  
Chicago Press, IL, 1992).
- 45 42. A. Kendon, *Gesture: Visible Action as Utterance* (Cambridge University Press, 2004).
- 47 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.
- 49 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 45. H. G. Wallbott and K. R. Scherer, Cues and channels in emotion recognition, *J. Pers.*  
*Soc. Psychol.* **51**(4), 690–699 (1986).
- 47 46. P. Gallaher, Individual differences in nonverbal behavior: Dimensions of style, *J. Pers.*  
*Soc. Psychol.* (63), 133–145 (1992).

- 1 47. S. Kopp and I. Wachsmuth, Synthesizing multimodal utterances for conversational  
agents, *J. Comp. Animat. Virtual Worlds* **15**(1), 39–52 (2004).
- 3 48. S. Kopp, P. Tepper and J. Cassell, Towards integrated microplanning of language and  
5 iconic gesture for multimodal output, in *Int. Conf. Multimodal Interfaces (ICMI'04)*,  
Penn State University, State College, PA (2004), pp. 97–104.
- 7 49. B. Hartmann, M. Mancini and C. Pelachaud, Formational parameters and adaptive  
prototype instantiation for MPEG-4 compliant gesture synthesis, in *Computer Ani-  
9 mation (CA'2002)* (IEEE Computer Society, Geneva, Switzerland, 2002), pp. 111–119.
- 11 50. B. Hartmann, M. Mancini and C. Pelachaud, Implementing expressive gesture syn-  
thesis for embodied conversational agents, in *Gesture Workshop (GW'2005)*, Vannes,  
France (2005).
- 13 51. C. Pelachaud, Multimodal expressive embodied conversational agent, in *ACM Multi-  
media, Brave New Topics Session* (ACM, Singapore, 2005), pp. 683–689.
- 15 52. P. Ekman, *The Face Revealed* (Weidenfeld & Nicolson, London, 2003).
- 17 53. D. Keltner, Signs of appeasement: Evidence for the distinct displays of embarrassment,  
amusement, and shame, *J. Pers. Soc. Psychol.* **68**, 441–454 (1995).
- 19 54. D. Matsumoto, More evidence for the universality of a contempt expression, *Motiv.  
Emotion* **16**, 363–368 (1992).
- 21 55. B. Bouchon-Meunier, M. Rifqi and S. Bothorel, Towards general measures of compar-  
ison of objects, *Fuzzy Set. Sys.* **84**, 143–153 (1996).
- 23 56. D. Dubois and H. Prade, *Fuzzy Sets and Systems* (Academic Press, New York, 1980).
- 25 57. P. Ekman and W. V. Friesen, The repertoire of nonverbal behavior's: Categories,  
origins, usage and coding, *Semiotica* **1**, 49–98 (1969).
- 27 58. P. Ekman, Darwin, deception, and facial expression, *Ann. NY Acad. Sci.* **1000**, 205–  
221 (2003).
- 29 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.
- 31 60. P. Ekman and W. V. Friesen, *Manual for the Facial Action Coding System* (Consulting  
Psychology Press, Palo Alto, CA, 1978).
- 33 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).
- 35 62. B. De Carolis, C. Pelachaud, I. Poggi and M. Steedman, APMML, 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.