Towards Modelling Embodied Conversational Agent Character Profiles Using Appraisal Theory Predictions in Expression Synthesis

L. Malatesta, A. Raouzaiou, K. Karpouzis and S. Kollias

Image, Video and Multimedia Systems Laboratory, National Technical University of Athens,
9, Heroon Politechniou str., Zografou 15780, Greece
{lori, araouz}@image.ece.ntua.gr
kkarpou@softlab.ece.ntua.gr
stefanos@cs.ntua.gr

Abstract. Appraisal theories in psychology study facial expressions in order to deduct information regarding the underlying emotion elicitation processes. Scherer's component process model provides predictions regarding particular face muscle deformations that are attributed as reactions to the cognitive appraisal stimuli in the study of emotion episodes. In the current work, MPEG-4 facial animation parameters are used in order to evaluate these theoretical predictions for intermediate and final expressions of a given emotion episode. We manipulate parameters such as intensity and temporal evolution of synthesized facial expressions. In emotion episodes originating from identical stimuli, by varying the cognitive appraisals of the stimuli and mapping them to different expression intensities and timings, various behavioural patterns can be generated and thus different agent character profiles can be defined. The results of the synthesis process are consequently applied to Embodied Conversational Agents (ECAs), aiming to render their interaction with humans, or other ECAs, more affective.

1.Introduction

Affective computing dictates the importance of creating interfaces which are not solely limited to the synthetic representation of the face and the human body, but which also expresses feelings through facial expressions, gestures and the body pose. The most significant challenge is the compatibility of an embodied conversational agent with MPEG-4 standard and its use in various applications. The use of affective agents can be applied in many sectors – culture, gaming, e-learning, while their compatibility with the MPEG-4 standard makes it possible for them to interact with synthetic objects and to be seamlessly integrated in different scenes.

Scherer's appraisal theory, the component process model, investigates the link between the elicitation of an emotion and the response patterning in facial expression [1],[2]. It

predicts intermediate expressions based on the results of stimulus appraisal checks and postulates a cumulative effect of these checks and their corresponding intermediate expressions on the final expression.

Our current work aims to investigate the use of component process model predictions in facial expressions. By manipulating the involved parameters we aim to model facial expressions of different synthetic character profiles as reactions to identical stimuli.

At this point we focus on MPEG-4 synthesis of facial expressions based on appraisal theory predictions for a given cognitive evaluation of stimulus (from collected psychological data) without being concerned about the stimuli itself. The long term plan is to design scenarios -situated in virtual environments- that can cater for the desired stimuli and interactions.

2.Agent's facial expression and the component process model

The processes of emotion elicitation and emotion expression constitute central issues in rendering a virtual agent more affective. Emotion theory offers a variety of models each aspiring to capture the emotion expression process. One would expect that the choice of the modelling approach would be irrelevant of the task at hand and would aim to capture global patterns. Contrary to this intuition, relevant research has shown that the choice of the modelling approach is strongly correlated to the task the agent will be asked to carry out. For example the dimensional approach [3] to emotion modelling is more fitting for the case of emotion recognition i.e. anger detection. It remains a challenge to identify the emotion model for an agent that will not be dependant of specific action examples.

By studying the requirements for a naturalistic interaction with a virtual character, it is made clear that there exist crucial details on the process of facial expression evolution such as which muscles of the face participate in the expression, for how long and with what intensity. We chose to investigate how Scherer's component process model's predictions can actually provide this required information. This emotion model gives predictions for intermediate expressions as well as a prediction for the final emotion expression based on cognitive appraisal checks, preformed on various specifically defined components. In our current work we are interested in evaluating this theoretical model and in investigating ways in which appraisal check results and the accompanying predictions can become behaviour metrics for virtual agents in dynamic environments.

According to cognitive theories of emotion, emotions are closely related to the situation that is being experienced (or, indeed, imagined) by the agent. More specifically, emotions are connected to mental representations that emphasize key elements of a situation and are identified as being either positive or negative. These representations have generally been called appraisals. An appraisal can be thought of as a model which is selective and valenced – i.e., highlights key elements of a situation and their values for good or ill [4]. Early examples of this approach can be found in [5], [6]. Appraisals are not necessary conscious, thus the evaluation processes can occur also by an unconscious

way as demonstrated by an important corpus study in cognitive neuroscience, with different methods of subliminal presentations of stimuli or by clinical neuropsychology (e.g. [7]).

Scherer has developed an appraisal model of emotion in which emotions are conceptualized as the outcome of a fixed sequence of checks [1], [8]. According to Scherer's view, emotion serves an important function as "…an evolved phylogenetically continuous mechanism that allows increasingly flexible adaptation to environmental contingencies by decoupling stimulus and response and thus creating a latency time for response optimization" [1].

Cognitive appraisal is modelled through a sequence of Stimulus Evaluation checks (SECs), which represent the smallest set of criteria necessary to account for the differentiation of main groups of emotional states. These checks are not necessarily binary and are subjective (i.e. they depend on both the appraising individual's perception of and inference about the specific characteristics of the event [1]).

The individual stimulus evaluation checks can be grouped together in terms of what are called Appraisal Objectives. Four appraisal objectives are defined:

1) Relevance Detection: comprising Novelty Check, Intrinsic Pleasantness Check, and Goal Relevance Check;

2) Implication Assessment: comprising Causal Attribution Check, Discrepancy from Expectation Check, Goal/Need Conduciveness Check, and Urgency Check;

3) Coping Potential Determination: comprising Control Check, Power Check, and Adjustment Check (can the event be controlled, if so by how much power do I have to exert control, and if not can I adjust?);

4) Normative Significance Evaluation: comprising Internal Standards Check, and External Standards Check. A major assumption of Scherer's Stimulus Evaluation Checks Theory is that the sequence of the checks and of their groups is fixed. However, this does not rule out parallel processing as, in theory, all of the stimulus evaluation checks are processed almost simultaneously.

Using this model of emotion, all representations of emotional states are explained in terms of cognitive appraisals of the antecedent situation. Results from these appraisals account for the differentiated nature of emotional responses, both within and between individuals. Appraisals also make appropriate emotional responses likely, and conflict between automatic, unconscious appraisals and more consciously deliberated ones may explain some of the more irrational aspects of emotions [5].

3.MPEG-4 based representation and the Facial Action Coding System

In the framework of MPEG-4 standard [9], a set of parameters has been specified for Face and Body Animation (FBA) by defining specific Face and Body nodes in the scene graph. MPEG-4 specifies 84 feature points on the neutral face, which provide spatial reference for facial animation parameters' definition. The facial animation parameter set consists of

two high-level parameters, visemes and expressions. A viseme is a basic unit of speech in the visual domain that corresponds to a phoneme (which is the basic unit of speech in the acoustic domain). It describes the particular facial and oral movements that occur alongside the voicing of phonemes.

Most of the techniques for facial animation are based on a well-known system for describing "all visually distinguishable facial movements" called the Facial Action Coding System (FACS). This coding system is anatomically oriented and based on the definition of "Action Units" (AU) of a face that cause facial movements. The system tries to distinguish the visually distinguishable facial movements using the knowledge of facial anatomy. An Action Unit could combine the movement of two muscles or work in the reverse way, i.e., one muscle movement could be expressed in more than one action units. MPEG-4's facial animation parameters are strongly related to action units [2]. A description of archetypal expressions by means of muscle movements and action units has been the starting point for describing archetypal expressions using facial animation parameters.

In particular, the set of Facial Definition Parameters (FDP) was designed in the MPEG-4 framework to allow the definition of a facial shape and texture, eliminating the need for specifying the topology of the underlying geometry. The set of Facial Animation Parameters (FAP) was designed to allow the animation of faces reproducing expressions, emotions and speech pronunciation. Viseme definition has been included in the standard for synchronizing movements of the mouth related to phonemes with facial animation. By monitoring facial gestures corresponding to FDP and/or FAP movements over time, it is possible to derive cues about user's expressions and emotions. Various results have been presented regarding classification of archetypal expressions of faces, mainly based on features or points mainly extracted from the mouth and eyes areas of the faces. These results indicate that facial expressions, possibly combined with gestures and speech, when the latter is available, provide cues that can be used to perceive a person's emotional state.

4.Facial expression synthesis based on Appraisal Theory predictions

Using the predictions of Scherer's component process model for the intermediate expressions we chose hot anger and fear [10] and generated videos animating the transition between the predicted expressions using the GretaPlayer MPEG-4 decoder. These predictions determine which action units participate in intermediate expressions. Each intermediate expression is the outcome of an appraisal (a stimulus evaluation check as it is formally called). The final expression is based on the cumulative effect of these intermediate expressions. The animation process in MPEG-4 was based on the mapping of Ekman's Action Units [11] to MPEG-4 Facial Animation Parameters (FAPs) [3]. A detailed list of the stimulus evaluation checks and their predictions both in action units and in facial animation parameters can be seen in Table1 for the case of fear and Table 2 for the case of hot anger. Our focus on these two emotions aims to be the beginning of an

attempt to model the effects of stimulus evaluations checks on facial expressions in general, taking advantage of the flexibility and the expressivity the GretaPlayer engine has to offer.

Table 1. Emotion: Fear

Stimulus Evaluation Check/ Result	Predicted participating action units	Corresponding MPEG-4 Facial Animation Parameters
novelty/ sudden	1: inner brow raiser + 2: outer brow raiser + 5: upper lid raiser	<31,32> + <35,36> + <19,20>
intrinsic pleasantness/ unpleasant	9: Nose wrinkeler + 10: Upper lip raiser + 15: Lip corner depressor + 35: Nostril compressor	(no MPEG-4 equivalents for au9) <4, 61, 62, 63> + <-12, -13> (no MPEG-4 equivalents for au35)
expectation/ discrepant	4: Brow lowerer + 7: Lid tightener	<31 to 38> + <21, 22>
goal attainment/ obstructive	17: chin raiser + 23: lip tightener	<18> + <5,6,7>
power, control/ low	20: Lip stretcher + 26: Jaw drops + 27: Mouth stretches	<5, 6, 7, 10, 11, -12, -13, 55, 56, 57, 59, 56, 56, 58, 60> (no MPEG-4 equivalents for au26-27)

Table 2. Emotion: Hot Anger

Stimulus Evaluation Check/ Result	Predicted participating action units	Corresponding MPEG-4 Facial Animation Parameters
novelty/ sudden	1: inner brow raiser + 2: outer brow raiser + 5: upper lid raiser	<31,32 33,34> + <35,36> + <19,20>
goal attainment/ obstructive	4: Brow lowerer + 7: Lid tightener	<31 to 38> + <21,22>
Control Potential/ Control high, power high	4: Brow lowerer + 7: Lid tightener + 10: Upper lip raiser + 17: Chin raiser + 24: Lip pressor	<31 to 38> + <21,22> + <4,61,62> + <18> + <4,5,8,9,10,11>

Until recently, most of our work in facial expression analysis had to do with static images of the apex of a facial expression, since no videos with satisfactory resolution were available so as to allow for the tracking of the evolution of a facial animation

parameter (a specific point in the face) in successive frames. Nevertheless, because we are dealing with synthesis and not analysis of expressions at this point, it is possible to simulate the movement of the specific points in the face as an expression evolves, if some intermediate frames are available. The component process model's predicted intermediate and final expressions can play the role of these in between frames in a continuous sequence. Now, in contrast to a display of static images of the intermediate predictions (which is what has been put forward till now in corresponding psychological research), in the process of video synthesis we have to deal with several novel unanswered issues for which no theoretical predictions are available. The temporal evolution of the expressions, the track each facial animation parameter follows, the intensity of the deformation of the predicted participating points in the face as well as the way all these are combined to give the final expression are questions that can only be answered through example synthesis and rating tests. Thus, we decided to investigate various methods of transition between the intermediate expressions and evaluate them through a rating test.

The component process model predicts a *cumulative effect* of intermediate predictions on the final expression of an emotion. In our attempt of a hands-on investigation of this effect we have identified two major ways of treating the evolution of an expression between the intermediate expression predictions provided by the appraisal checks, an additive animation and a sequential one. They are methods based on principles of computer graphics that require further empirical testing on the naturalness of their outcome. In this preliminary research both approaches were tested in depth for the emotions of hot anger and fear. We choose to present a representative subset of our results on a frame to frame level with additive fear animation in Figure 1 and sequential hot anger animation in Figure 2.

4.1 *Additive Animation*

In the case of additive animation, as seen in the fear example (Figure 1), each intermediate expression is derived by the addition of the predicted participating action units of the current expression (table 1) to the action units of the previous intermediate expression. Since we are talking about MPEG-4 compliant synthesis the animation is attained by using the equivalent facial animation parameters. Table 1 includes the appraisal dimensions (stimulus evaluation checks), the predicted participating action units and the corresponding facial animation parameters. It is beyond the scope of this work to explain in detail the AU to FAP mapping which is analysed in depth in [3].

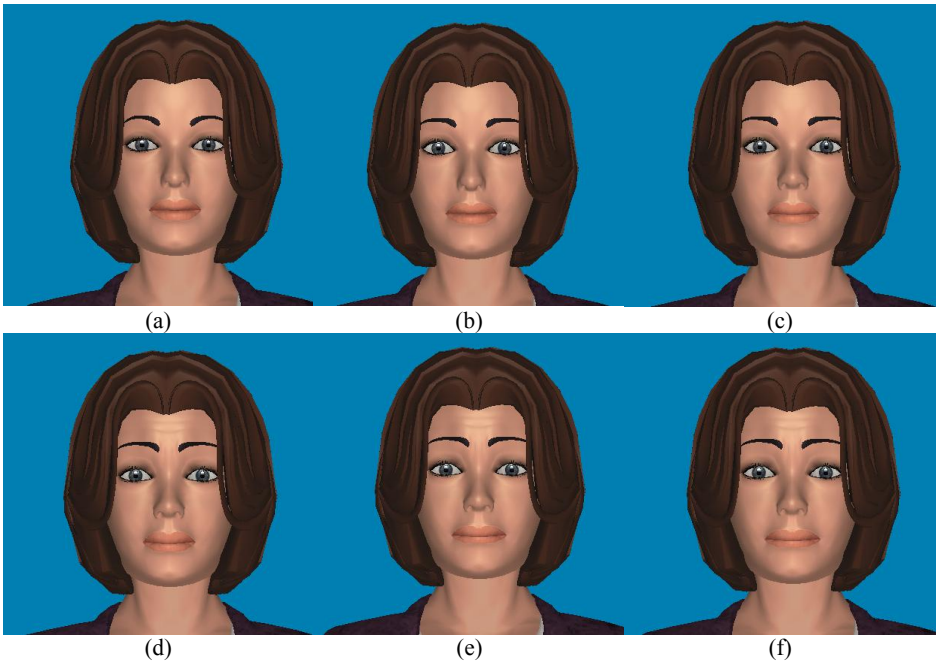

Fig. 1. Intermediate predictions of facial expressions according to Scherer's component process model for the case of fear-(a) neutral, (b) novelty-sudden, (c) unpleasant, (d) discrepant, (e) goal obstractive, (f) low control-final expression –fear. Each expression is derived from the "addition" of the previous expression's AUs and those of the current one.

This additive animation approach was found to be problematic in the cases when subsequent expressions are constituted of conflicting animations. For example in the case of hot anger the "novelty high" intermediate expression, according to the component process model predictions ([5]) includes eyebrows being raised. The next intermediate prediction is "goal obstructive" and predicts lowered eyebrows. This conflict renders the animation problematic since the parts of different intermediate expressions some times counterbalance themselves.

4.2 *Sequential Animation*

In the case of sequential animation – as shown in the hot anger example (Figure 2), all intermediate expressions are animated in sequence. This could have be realized either by interposing the neutral expression between the predictions or by "tweening" from one expression to the other keeping the common deformations as the common denominator. Tweening is a key process in all types of animation, including computer animation. It is short for "in-betweening" and it's the process of generating intermediate frames between

two images to give the appearance that the first image evolves smoothly into the second image. Since the approach containing the neutral expressions between predicted expressions was considered to render the outcome counterintuitive we only focused on direct tweening of the intermediate expressions. Overall the tweening approach is friendlier to the eye but is still not perceived as a realistic expression generation as we can see in the rating test results that follow.

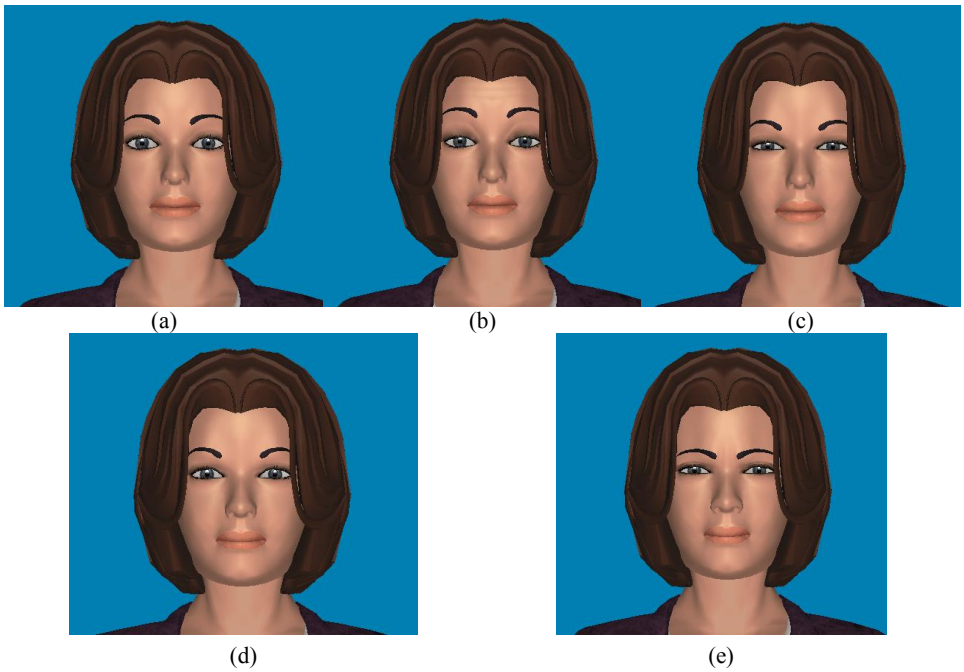

Fig. 2. Intermediate predictions of facial expressions according to Scherer's component process model for the case of hot anger-(a) neutral, (b) novelty-high, (c) goal obstructive, (d) control high/power high, (e) final expression –hot anger.

4.3 *Manipulating animation parameters*

For both animation approaches we have analysed the synthesis method but have yet to clarify the issues of temporal evolution and intensity in facial expression evolution. What we did was to produce four starting video sequences (fear additive, fear sequential, hot anger additive, hot anger sequential). In order to end up with these four sequences we linearly interpolated between the values of the intermediate expressions. We used the typical MPEG-4 25 frames per second animation speed. The values of the facial animation parameters for each intermediate expression were approximated by comparing

snapshots of Greta with those provided by psychological predictions. In order to achieve linear interpolation for each FAP between each intermediate expression the step for each FAP had to be calculated. This was done by taking the difference between the FAP values of two consecutive intermediate expressions and dividing it by the number of the frames in between. We arbitrarily chose 50 frames in between intermediate expressions following the suggestions of experts in the field. In the same way we chose 40 frames for each intermediate expression in order to freeze it for enough time for it to be perceived in the video sequence.

By carefully manipulating these original sequences and their parameters we came up with plenty of variations both in the temporal evolution of the expression and in the intensity of the participating face muscles.

5. Rating test on animation type

There are various questions worth answering through rating tests in the synthesis of emotional expressions. In our case the first question worth asking, in order to continue an in depth analysis of the parameters just mentioned, was the regarding choice of the animation approach. Thus, we conducted a rating test using twenty computer science postgraduate students as participants in order to see how these two types of animation were perceived.

Our independent variable was the animation type and we randomly assigned half the participants to view the videos in the order: neutral, fear, neutral, hot anger, and half the participants to view the videos in the order: neutral, hot anger, neutral, fear. This was done so as to avoid ordering effects. We used a neutral condition of Greta simply blinking and looking straight ahead with a neutral expression as a conceptual baseline.

Participants were then asked, with the use of a questionnaire, which emotion they identified in each of the sequences they viewed. Their replies were constrained with labels. Participants' confidence was measured indirectly by asking them questions such as "how much fear do you think is displayed in this expression?" in the case of the hot anger sequence.

Results from this first rating test showed an above chance recognition of the correct emotions in the case of the additive approach, were as the sequential approach gave recognition results marginally above random choice. Confidence measures were also higher in the additive case for the emotion of hot anger, although in both additive and sequential fear videos sadness was also recognised above average, thus decreasing the overall level of confidence.

Based on these preliminary results, rating tests to follow will be based on sequential animation and will investigate the perception of different temporal evolutions of the same emotions as well as different maximum values of their participating facial animation parameters. More emotion sequences will also be developed according to the available component process model predictions

6.Future Work – Adding Gesture Expressivity

Affective interaction with a virtual agent can be achieved using various modalities. So far we have limited our study in facial expressions for given cognitive appraisals of stimuli. Gestures can be used to reinforce the message that the agent wishes to convey. The way they are acted out (the temporal evolution of a gesture, it's time duration etc) can also differentiate the message put forward. In the literature of perception studies six specific gesture expressivity parameters are defined [12][13]:

- Overall activation
- Spatial extent
- Temporal
- Fluidity
- Power/Energy
- Repetitivity

These expressivity dimensions have been designed for communicative behaviours only. Research on gesture synthesis -using tools provided in the MPEG-4 standard- studies the variations of specific expressivity parameters in respect to patterns identified in the process of human gesture analysis [14]. According to such approaches, we can manipulate the expressivity parameters of gestures in given scenarios/ emotional episodes for different character profiles. A "short – fused" character will react more vividly in a goal obstructive situation than a "serene" one. Using this in conjunction with the predictions on facial expressions of each emotion episode in a situated/ scenario based approach we aspire to simulate different agent behaviours that will eventually correspond to different character profiles.

7.Conclusions

Appraisal theories strive to explain the emotion elicitation process. They have been applied to virtual environments in order to model emotional behaviour in virtual humans. There is rich ongoing research on computational models of emotion based on psychological appraisal theories. EMA, a computational model of emotional behaviour by Cratch and Marsella [15] is a state of the art depictive example. It combines appraisal theory principles and coping predictions in order to model behaviour and emotional states. Other modelling attempts also adopt appraisal theory principles and test them empirically in data collection environments such as games or simple interaction interfaces [16], [17].

The challenge we are taking up in this work is to empirically test the hypothesis that different evaluations of stimulus can be used to define different agent character profiles. Different evaluations of stimulus lead to different facial expressions. These facial expressions can be coupled with matching gestures by choosing the appropriate expressivity parameters. For the time being we plan on dealing with non-verbal agent expressivity where specific scenarios can provide the stimuli for the agent to react to. The

animations computed in the current paper will be imported in a virtual environment where users will be able to interact with the agents and invoke different reactions according to each agent character profile.

Acknowledgments

This research is partly supported by the EC Project HUMAINE (IST-507422).

References

[1] Scherer, K.R.: Appraisal Considered as a Process of Multilevel Sequential Checking. In Scherer, K.R., Schorr, A., & Johnstone, T., (Eds) Appraisal Processes in Emotion: Theory Methods, Research. Oxford, New York: Oxford University Press, 92-129 (2001)

[2] Sander, G., Didier, D. &. Scherer K. (2005), 'A systems approach to appraisal mechanisms in emotion', Neural Networks 18, 317-352.

[3] Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., Kollias, S.: Parameterized facial expression synthesis based on MPEG-4. EURASIP Journal on Applied Signal Processing, Vol. 2002, No. 10. Hindawi Publishing Corporation (2002) 1021-1038.

[4] Picard, R. W.: *Affective Computing*. MIT Press, Cambridge, MA, (1997)

[5] Roseman I.J. and Smith, C.A.: Appraisal Theory: Overview, Assumptions, Varieties, Controversies. In Scherer, K.R., Schorr, A., & Johnstone, T., (Eds) Appraisal Processes in Emotion: Theory Methods, Research. Oxford, New York: Oxford University Press, 3-19 (2001)

[6] Ortony, A., Clore, G.L. and Collins, A.: The Cognitive Structure of Emotions, Cambridge, UK: Cambridge University Press (1988)

[7] Pegna, A. J., Khateb, A., Lazeyras, F., & Seghier, M. L.: Discriminating emotional faces without primary visual cortices involves the right amygdala. *Nature Neuroscience, 8*(1), 24–25 (2004)

[8] Scherer, K.R.: On the Nature and Function of Emotion: A Component Process Approach. In Scherer, K.R., & Ekman, P., (Eds) Approaches to Emotion. Hillsdale, New Jersey, London: Lawrence Erlbaum Associates, Publishers. 293-318 (1984)

[9] Tekalp, M., Ostermann, J.: Face and 2-D mesh animation in MPEG-4. Image Communication Journal, Vol.15, Nos. 4-5 (2000) 387-421

[10] Wehrle, K. S. S. S. &. S. K. R. (2000), 'Studying the dynamics of emotional expression using synthesized facial muscle movements', Journal of Personality and Social Psychology 78(1), 105-119.

[11] Ekman, P.: "Facial expression and Emotion," *Am. Psychologist*, vol. 48 pp.384-392 (1993)

[12] Hartmann, B., Mancini, M. and Pelachaud, C., Implementing Expressive Gesture Synthesis for Embodied Conversational Agents. Gesture Workshop (2005), Vannes.

[13] Wallbott, H.G, Bodily expression of emotion. European Journal of Social Psychology, 28:879–896, 1998.

[14] Caridakis, G.; Raouzaiou, A.; Karpouzis, K. & Kollias, S. (2006), Synthesizing Gesture Expressivity Based on Real Sequences, in , Workshop on multimodal corpora: from multimodal behaviour theories to usable models, LREC 2006 Conference, Genoa, Italy.

[15] Gratch, J. & Marsella, S. (2004), A Domain-Independent Framework for Modeling Emotion, Cognitive Systems Research 5(4):269-306.

[16] Wehrle, T. & Kaiser, S. (2000). Emotion and facial expression. In A. Paiva (Ed.), Affect in Interactions: Towards a new generation of interfaces, (pp. 49-64). Heidelberg: Springer.

[17] Gebhard, P. (2005), ALMA: a layered model of affect, in 'AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems', ACM Press, New York, NY, USA, pp. 29-36.