# EXPRESSIVE AVATARS IN MPEG-4

*M. Mancini, B. Hartmann, C. Pelachaud*
IUT of Montreuil - University of Paris 8

{m.mancini, c.pelachaud}@iut.univ-paris8.fr

*A. Raouzaiou, K. Karpouzis*
Image, Video and Multimedia Systems Laboratory, Natl Technical University of Athens
{araouz, kkarpou}@image.ece.ntua.gr

## ABSTRACT

Man-Machine Interaction (MMI) Systems that utilize multimodal information about users' current emotional state are presently at the forefront of interest of the computer vision and artificial intelligence communities. A lifelike avatar can enhance interactive applications. In this paper, we present the implementation of GretaEngine and synthesized expressions, including intermediate ones, based on MPEG-4 standard and Whissel's Emotion Representation.

## 1. INTRODUCTION

Research in facial expression analysis and synthesis has mainly concentrated on archetypal emotions. In particular, sadness, anger, joy, fear, disgust and surprise are categories of emotions that attracted most of the interest in human computer interaction environments. Moreover, the MPEG-4 indicates an alternative way of modeling facial expressions and the underlying emotions, which is strongly influenced from neurophysiological and psychological studies (FAPs). The adoption of token-based animation in the MPEG-4 framework [1] benefits the definition of emotional states, since the extraction of simple, symbolic parameters is more appropriate to analyze, as well as synthesize facial expression and hand gestures.

In this paper we describe the implementation of GretaEngine and an approach to synthesize expressions, including intermediate ones, via the tools provided in the MPEG-4 standard based on real measurements and on universally accepted assumptions of their meaning, taking into account results of Whissel's study [1]. The results of the synthesis process can then be applied to avatars, so as to convey the communicated messages more vividly than plain textual information or simply to make interaction more lifelike.

## 2. EMOTION REPRESENTAION

The obvious goal for emotion analysis applications is to assign category labels that identify emotional states. However, labels as such are very poor descriptions, especially since humans use a daunting number of labels to describe emotion.

Activation-evaluation space [3] is a representation that is both simple and capable of capturing a wide range of significant issues in emotion. A basic attraction of that arrangement is that it provides a way of describing emotional states which is more tractable than using words, but which can be translated into and out of verbal descriptions. Translation is possible because emotion-related words can be understood, at least to a first approximation, as referring to positions in activation-emotion space. Various techniques lead to that conclusion, including factor analysis, direct scaling, and others.

## 3. FACIAL EXPRESSION IN MPEG-4

### 3.1. Modeling Primary Expressions Using Motion Capture Data

We currently use a system based on key frame animation, where an expression is defined by 3 temporal parameters, namely onset, apex and offset. But such a specification does not allow one to capture the subtlety of facial expression dynamism. In order to improve these animations, we are studying real data of facial movements coming from motion capture sequences (we are very grateful to Franck Multon of University of Rennes 2 for the recording of the motion captured data) we have recorded using an Oxford Metrics Vicon system (www.vicon.com). Our data are organized into 78 sequences performed by two actors, a man and a woman, each having 33 markers on the face, 21 of which correspond to FAPs (Facial Animation Parameter) locations. These sequences are simple basic movements, like raising eyebrows or smiling, and basic emotions such as anger, happiness, surprise. Finally we recorded two sequences of monologues in which extreme expressions of emotions were displayed.

Having filtered the data and taken out head movements, we compute the X, Y and Z displacement of each marker in all frames. The displacements are normalized by FAPUs (Facial Animation Parameter Units) defined for each actor's face. From the extracted FAPs values, we study their movements. We can notice that FAPs from the same facial area, like the four ones describing the movement of an eyebrow, have the same movement with a proportional factor. After simulating each displacement curves using the ADSR (Attack, Decay, Sustain, Release) model, we found out that the model is a good approximation of the captured data. For example, most of the movements follow the three phases, Attack-Sustain-Release, and an intense movement always has a Decay phase. To generate the FAPs values, we determine the desired phases (A, D, S or R), an intensity value and a duration values. The FAP values are then calculated with control points linked by a Piecewise cubic Hermite interpolation. The analysis is performed on data coming from both actors to find out individual differences.

### 3.2. Modeling Primary and Intermediate Expressions Using Computer Vision

In order to model an emotional state in a MMI context, we must first describe the six archetypal expressions (joy, sadness, anger, fear, disgust, surprise) in a symbolic manner, using easily and robustly estimated tokens. FAPs representations [1] make good candidates for describing quantitative facial and hand motion features. The use of these parameters serves several purposes such as compatibility of created synthetic sequences with the MPEG-4 standard and increase of the range of the described emotions – archetypal expressions occur rather infrequently and in most cases emotions are expressed through variation of a few discrete facial features related with particular FAPs.

Based on elements from psychological studies [2], we have described the six archetypal expressions using MPEG-4 FAPs [4]. In general, these expressions can be uniformly recognized across cultures and are therefore invaluable in trying to analyze the users' emotional state.

The initial range of variation for the FAPs has been computed as follows: Let $m_{i,j}$ and $\sigma_{i,j}$ be the mean value and standard deviation of FAP $F_j$ for the archetypal expression $i$ (where $i$={1→Anger, 2→Sadness, 3→Joy, 4→Disgust, 5→Fear, 6→Surprise}), as estimated in [4].

Apart from archetypal expressions, in everyday life one can meet a variety of other expressions, not always possible to be defined. These expressions can belong to the same category with one of the six archetypal expressions or can lye between them. As a general rule, one can define six general categories, each characterized by an archetypal emotion; within each of these categories, intermediate expressions are described by different emotional intensities, as well as minor variation in expression details. From the synthetic point of view, emotions belonging to the same category can be rendered by animating the same FAPs using different intensities. This ensures that the synthesis does not render "robot-like" animation, but drastically more realistic results.

For example, the emotion group *fear* also contains *worry* and *terror* [4] which can be synthesized by reducing or increasing the intensities of the employed FAPs, respectively.

On the other hand, creating profiles for emotions that do not clearly belong to a universal category is not straightforward. Apart from estimating the range of variations for FAPs, one should first define the vocabulary of FAPs for the particular emotion. In order to proceed we utilize the *activation* parameter of Emotion's Wheel. Here we have to notice that the combination of two archetypal expressions is better defined if these two expressions are belonging to the same *evaluation hemicycle*; to combine a negative and a positive emotion is not always well defined, although in some cases the derived expressions are very well illustrated.

As already stated, defining a vocabulary is not enough for modeling expressions; profiles should be created for this purpose. This poses a number of interesting issues in the case of different FAPs employed in the animation of individual profiles: in our approach, FAPs that are common in both emotions are retained during synthesis, while FAPs used in only one emotion are averaged with the respective neutral position. The same applies in the case of mutually exclusive FAPs: averaging of the intensities usually favors the most exaggerated of the emotions that are combined, whereas FAPs with contradicting intensities are cancelled out. In practice, this approach works successfully, as shown in the animated profiles in Section 5.3. The combination of different, perhaps contradictory or exclusive, FAPs can be used to establish a distinct emotion categorization, similar to the semantic one, with respect to the common or neighboring FAPs that are used to synthesize and animate emotions.

It should be noted that the profiles, created using the above procedure, have to be animated for testing and correction purposes; the final profiles are those that present an acceptable visual similarity with the requested real emotion.

## 4. EXPRESSIVE GESTURE MODELING

Our approach to gesture expressivity is driven by a perceptual standpoint -- how expressivity is perceived by others; not what internal muscle activation patterns underlie these signals. Researchers in social psychology have investigated how various influences affect perceived bodily behaviours, mostly through ad-hoc measuring instruments constructed by narrowing down an extensive list of choices through coder reliability testing.

Wallbot and Scherer [7] had judges encode their impressions of behaviour with respect to speed, volume, style (weak or energetic), small or large movement activity and pleasantness. Gallaher [8] found four significant dimensions of variability in personal encoding style:

- expressiveness - energetic communication;
- animation - energy in acts not directly related to communication;
- expansiveness - use of space, elbow position;
- co-ordination - smoothness, fluidity.

Human movement observation is an active field within the dance community. The most prominent system of notation is Laban movement analysis (kinetography) [9]. Laban uses five dimensions of classification: Body, Space, Shape, Effort, and Relationship. Each dimension is further subdivided into a set of parameters, e.g., four each for the Effort and Shape dimensions. Spatial extension (Space) is captured as well as movement information (Shape) and intention (Effort).

## 5. IMPLEMENTATION

This sections focuses on the parameters that we use in our facial and gesture engine and the methods used to calculate motion based from these parameters. Our animation engine takes one APML [10] tagged text (a language defined using XML that specifies the communicative function of the text) and outputs two animation files, one for the face and one for the gesture animation. Very important is the notion we called "expressivity", that starting from some influences like the agent's personality, social role and so on, will affect the process of calculating the agent's animation (therefore the agent's behaviour perceived by the user)[6].

We propose a set of six attributes which can be considered as a basic qualitative representation of human expressivity (see [5]) and which will influence both the facial and gesture animation generation:

- Overall activation: amount of activity across several modalities during a conversational turn
- Spatial extent: amplitude of movements, e.g. amount of space taken up by body
- Temporal: duration of movements
- Fluidity: smoothness and continuity of overall movement
- Power/Energy: dynamic properties of the movement, e.g. weak/relaxed versus strong/tense
- Repetitivity: tendency to rhythmic repeats of specific movements along specific modalities

### 5.1. Facial animation

A facial expression is characterised providing its temporal parameters and its shape, that is the quantity of displacement for all the involved FAPs. Each expression is temporally described by four phases:

- attack: the time that, starting from the neutral face, the expression takes to reach its maximal intensity
- decay: the time during which expression intensity lightly decreases to reach the “sustain” value
- sustain: the time during which the expression is maintained (and is the more visible part)
- release: the time needed for an expression to return to neutral, starting from the maximal intensity

The “Spatial extent” expressivity parameter will influence the way in which the peak value of the attack phase is calculated. Then the duration of each of the four phases will be scaled accordingly to the “Temporal” parameter. “Fluidity” will increase or decrease of the slope of the A, D and R phases of the ADSR curve while “Power” will affect the lip movement controlling the lip muscle tension that may appear for the expressions of emotions like fear and anger. Finally “Repetitivity” will repeat the movement, for example increasing the number of head nods or eyebrows raisings accompanying emphasis. We are currently expanding this model to include the results from the analysis of the motion capture data (see Section 3.1.). An expression would be described as an n-tuple of elements A, D, S or R.

### 5.2. Gesture animation

Each gesture is defined in a library using a descriptive language that allows one to create gestures as a set of keyframes, using the composition of some basic hand shapes, palm orientations, wrist rotations and arm shapes [5]. Once keyframes are defined, the engine performs proper interpolation between them. “Spatial extent” expressivity parameter will expand or contract the entire space that is used for gesturing in front of the agent. Wrist positions in our gesture language are defined in terms of sectors of this space. Then we will use the “Temporal” parameter as a high-level notion of how quickly the gesture phases should be performed. “Fluidity” will capture the continuity between movements by modifying the continuity of the arms' trajectory paths as well as the acceleration and deceleration of the limbs. These effects will be obtained modifying the control point locations for the kinematics Bezier splines or, for convenience, choosing TCB splines for the kinematics interpolants and adjusting their tension, continuity and bias parameters. “Power” will determine the amount of energy and tension invested into a movement. We will once again look at the dynamic properties of gestures (powerful movements will be expected to have higher acceleration and deceleration magnitudes) and also at the inter-gestural rest phases. “Repetitivity” will enable the generation of gestures which are composed by a sequence of a variable number of strokes, usually very close one to each other, generated by the engine by "contracting" the stroke phase length of the gesture and repeating it many times instead of only one.

### 5.3. Results

Figure 1(a) shows a particular profile for the archetypal expression *anger*, while Fig. 1(b) and (c) show alternative profiles of the same expression, differentiating on FAP intensities.

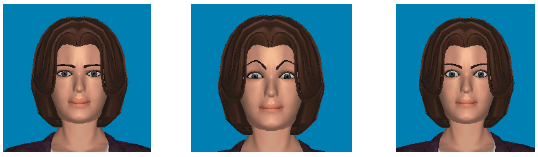
(a) (b) (c)

**Figure 1. Profile: (a)-(c) anger**

Animated profile of the expressions *terrified* and *worried* are illustrated in Figure 2. Both these expressions belong to the emotion category *fear*.

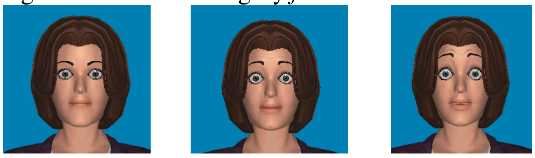
(a) (b) (c)

**Figure 2. Profiles for (a) worried, (b) afraid, (c) terrified**

Using the rules described above, *depression* (Figure 3b) is animated using *fear* (Fig.3a) and *sadness* (Fig.3c) and *suspicious* (Figure 4b) using *anger* (Fig. 4a) and *disgust* (Fig. 4c).

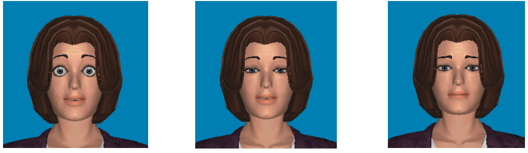
(a) (b) (c)

**Figure 3. Profiles for (a) fear, (b) depressed (c) sadness**

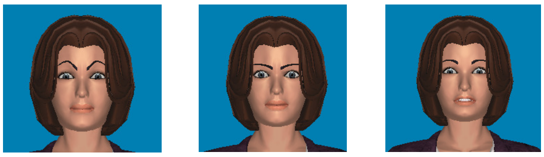
(a) (b) (c)

**Figure 4. Profiles for (a) anger, (b) suspicious (c) disgust**

## 6. CONCLUSIONS

Expression synthesis provides a powerful and universal means of expression and interaction in HCI applications. In this paper we presented a method of synthesizing realistic expressions using GretaPlayer. This method employs concepts included in established standards, such as MPEG-4, which are widely supported in modern computers and standalone devices.

## 7. REFERENCES

[1] M. Preda and F. Prêteux, “Advanced animation framework for virtual characters within the MPEG-4 standard”, *ICIP 2002*, Rochester, New York.

[2] P. Ekman, “Facial expression and Emotion,” *Am. Psychologist*, Vol. 48, pp.384-392, 1993.

[3] C.M. Whissel, “The dictionary of affect in language,” *Emotion: Theory, research and experience: Vol 4, The measurement of emotions,* Academic Press, 1989.

[4] A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis and S. Kollias, “Parameterized facial expression synthesis based on MPEG-4,” *EURASIP JASP*, 2002, No. 10.

[5] B. Hartmann, M. Mancini, C. Pelachaud, “Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis”, *Computer Animation 2002*, pp. 111. 3, 6, 7

[6] V. Maya, M. Lamolle, C. Pelachaud, “Influences and Embodied Conversational Agents”, *AAMAS 2004,* 1306-1307

[7] H. G. Wallbott and K. R. Scherer, “Cues and Channels in Emotion Recognition”, *Journal of Personality and Social Psychology*, 1986, v. 51, n. 4, p. 690-699

[8] P. E. Gallaher, “Individual Differences in Nonverbal Behavior: Dimensions of Style”, *Journal of Personality and Social Psychology*, 1992, v. 63, n. 1, p. 133-145

[9] R. Laban and F.C. Lawrence, “Effort: Economy in body movement”, Plays, Inc, 1974

[10] B. DeCarolis, C. Pelachaud, I. Poggi and M. Steedman, *APML, A mark-up language for believable behavior generation, Life-Like Characters*, Springer, 2004