Multimodal Expression in Virtual Humans

Celso de Melo Ana Paiva

IST – Technical University of Lisbon and INESC-ID

Avenida Prof. Cavaco Silva – Taguspark

2780-990 Porto Salvo, Portugal

cmme@mega.ist.utl.pt ana.paiva@inesc-id.pt

Abstract. This work proposes a real-time virtual human multimodal expression model. Five modalities explore the affordances of the body: deterministic, non-deterministic, gesticulation, facial and vocal expression. Deterministic expression is keyframe body animation. Non-deterministic expression is robotics-based procedural body animation. Vocal expression is voice synthesis, through Festival, and parameterization, through SABLE. Facial expression is lip-synch and emotion expression through a parametric muscle-based face model. Inspired by psycholinguistics, gesticulation expression is unconventional, idiosyncratic and unconscious hand gestures animation described as sequences of Portuguese Sign Language hand shapes, position and orientation. Inspired by the arts, one modality goes beyond the body to explore the affordances of the environment and express emotions through camera, lights and music. To control multimodal expression, this work proposes a high-level integrated synchronized markup language – Expressive Markup Language. Finally, three studies, involving a total of 197 subjects, evaluated the model in storytelling contexts and produced promising results.

Keywords: Multimodal expression, virtual humans, gesticulation expression, environment expression, facial expression, vocal expression

Introduction

Humans express themselves through their bodies. Speech and gesticulation express thought. Face, body and voice express emotions. As digital technology evolves, it is natural to explore in virtual human systems aspects of human expressiveness. In this sense, this work proposes a virtual human real-time multimodal expression model which integrates:

- Feature-based gesticulation expression as arbitrary sequences of Portuguese Sign Language hand shapes, orientations and positions;
- Vocal expression including speech synthesis and voice parameterization;
- Muscle-based facial expression including lip-synch and emotion expression;
- Multimodal expression control through a high-level synchronized markup language.

Furthermore, the human mind affords expression beyond the body. For long the arts have realized this and explored the surrounding environment to express emotions. This work introduces *environment expression* as a new modality for emotion expression through three environment channels – camera, illumination and music.

Related Work

Building a virtual human is a multidisciplinary challenge [1]. Computer graphics provides the means to model and animate bodies. Human sciences inform emotion and nonverbal behavior expression. Artificial Intelligence endows virtual humans with the capability to act, react and hold a conversation. This work touches all these areas.

Traditionally, virtual human animation is separated into body animation and facial animation. Body animation consists of animating a hierarchical skeleton. Control techniques include [2]: (a) *motion capture*, where animation is driven by a human actor; (b) *keyframe animation*, where a human artist defines keyframes and in-between frames are automatically generated; (c) *inverse kinematics*, where animation of the body's extremities automatically animate the rest of the chain; (d) *dynamics-based animation*, which generates physically realistic animation. This work explores the second and third techniques. Facial animation consists of deforming the face. Control techniques include [3]: (a) *motion capture*, where animation is driven by a human actor; (b) *keyframe animation*, where static face poses are interpolated; (c) *muscle-based animation*, where facial muscles are simulated and control parameters provided. This work explores the last technique.

Several computer graphics virtual human systems have been developed. Blumberg [4] proposes a layered architecture where the geometry defines skeleton and skin, the animation layer defines motor skills and an arbitration mechanism and the behavior layer defines goal-oriented behaviors which are converted into motor skills for execution. Perlin's *Improv* [5] system supports animation combination through a layering mechanism where final animation results from the weighted combination of all layers' animations and supports scripting to produce animation sequences. *Jack* [6] bases the skeleton on human anthropometry and anatomy studies so as to simulate correct proportions and motion ranges. Finally, *MIRALab*'s virtual humans [7] define detailed face and hand models and define an intermediate muscle layer thus, leading to more realistic animation.

This work explores gesticulation animation. Gesticulation is the kind of gestures humans do in conversation or narration contexts [8]. They tend to be focused on the arms and hands, though other body parts may be involved [9]. Gesticulation and speech co-express the same underlying idea unit synchronizing at the semantic and pragmatic levels. According to how it unfolds in time, gesticulation can be structured into phases. The stroke phase is where actual meaning is conferred and is synchronous with its co-expressive speech 90% of the time. McNeill and colleagues characterize gesticulation according to four dimensions [8][9]: *Iconicity*, which refer to characteristics of the action or event being described; *Metaphoricity*, which refer to abstract concepts; *Deixis*, which refer to concrete or abstract entities in the physical space surrounding the speaker; *Beats*, which refer to small baton like movements. According to McNeill ([9], p.42), “multiplicity of semiotic dimensions is an almost universal occurrence in gesture”. Thus, it makes more sense to speak of dimensions and saliency rather than categories and hierarchy.

Several virtual human gesticulation systems have been developed. Cassell and colleagues [10] proposed Animated Conversation, a rule-based system capable of synchronizing gestures of the right type with co-occurring speech. In [11] *Real Estate Agent (Rea)*, an embodied conversational agent is presented which is capable of multimodal input recognition, distribution of communicative intent across modalities and multimodal output synthesis. Kopp and colleagues [12] developed a comprehensive model for gesture animation based on research in psycholinguistics and motor control theory. Here, gesture production starts with gesture planning which selects appropriate gesture templates according to communicative intent, proceeds by instantiating the templates according to context and concludes by assigning appropriate temporal constraints. The gesture plan is then fed into a motor planner for execution. Finally, Cassell, Kopp and colleagues are currently developing *NUMACK* [13], a system capable of synthesizing in real-time co-verbal context-sensitive iconic gestures without relying on a library of predefined gestures.

Several researchers have explored *motion modifiers* to add emotive qualities to existent body motion data. Signal-processing techniques [14] were used to extract information from motion data which is used to generate emotional variations of neutral motion. Rose and colleagues [15] generate new motion with a certain mood or emotion from motion data interpolation based on radial functions and low order polynomials. Chi and colleagues [16] propose a system which adds expressiveness to existent motion data based on the effort and shape parameters of a dance movement observation technique called Laban Movement Analysis. However, in digital worlds, motion modifiers need not be limited to body parameters. In this sense, this work explores environment expression.

The problem of controlling multimodal expression has led to the development of several high-level languages [17]. Of particular relevance for this work are VHML [18], SMIL [19] and MURML [20]. These are all XML-based languages. The first, *Virtual Human Markup Language (VHML)* proposes six virtual human control subsystems: (1) Dialogue Manager Markup Language; (2) Speech Markup Language; (3) Facial Animation Markup Language; (4) Emotion Markup Language; (5) Body Animation Markup Language; (6) Gesture Markup Language. The second, *Synchronized Multimedia Integration Language (SMIL)* is oriented to audiovisual interactive presentations and proposes a sophisticated modality synchronization mechanism. Finally, *Multimodal Utterance Representation Markup Language (MURML)* defines a convenient notation for gesture definition and synchronization with co-verbal speech.

Model Overview

This work proposes a real-time multimodal expression model for virtual humans including deterministic, non-deterministic, gesticulation, facial, vocal and environment expression – see Figure 1. Virtual humans are structured according to a three-layer architecture [4][5].

The *geometry layer* defines a 54-bone human-based skeleton. The *animation layer* defines deterministic and non-deterministic animation mechanisms. The *behavior layer* supports a language for integrated synchronized multimodal expression.

[Figure 1: The multimodal expression model.]

Deterministic Expression

Deterministic expression supports human-dependent keyframe animation and several animation combination mechanisms. This modality revolves around *animation players* which animate subsets of the skeleton's bones according to specific animation mechanisms. Several players can be active at the same time and thus, as they may compete for the same bones, an arbitration mechanism based on priorities is defined. Supported animation mechanisms include: (a) *weighted combined animation*, where the resulting animation is the "weighted average" of animations placed on several weighted layers; (b) *body group animation*, where disjoint sets of skeleton's bones – body groups – execute independent animations; (c) *pose animation*, which applies stances to bones, supports combination between two stances and provides a parameter to control interpolation between them.

Non-Deterministic Expression

Non-deterministic expression applies robotics to virtual humans thus, laying the foundations for non-deterministic animation, i.e., human-free procedural animation. In the geometry layer, six revolute joint robotic manipulators are integrated with the skeleton to control the limbs and joint limits are defined according to anthropometry data in [21]. In the animation layer, three inverse kinematics and one inverse velocity primitives are defined: (1) *joint interpolation*, which animates the manipulator's target through interpolation in the joint space; (2) *function based interpolation*, which animates the target according to a transformation defined, at each instant, by a mathematical function; (3) *frame interpolation*, which animates the target according to interpolation between the current frame and the intended frame; (4) *Jacobian-based animation*, which applies inverse velocity algorithms to animate the target according to intended Cartesian and angular velocities.

Vocal Expression

Vocal expression integrates the Festival [22] text-to-speech system. This integration involves four aspects, Figure 2: (1) the notion of speech; (2) Festival's voice synthesis pipeline extension; (3) a communication protocol between Festival and virtual humans; (4) a new behavior layer API for voice control. A speech is modeled as a set of files including: (a) utterance structure, i.e., phonemes, words and times; (b) utterance waveforms; (c) a configuration file with information about all files. Festival's voice synthesis pipeline is extended, after natural language and signal processing, with the following steps: after each utterance synthesis, save its structure and waveform and inform the virtual human that an utterance is ready to play; after all utterances synthesis, save the speech file and communicate about speech completion. Integration builds on Festival's ability to function according to the server/client model. Thus, a communication protocol was developed which is characterized as follows: (a) supports voice synthesis primitives; (b) supports utterance conclusion communication throughout synthesis; (c) supports communication of speech synthesis conclusion. Finally, at the virtual human side, the behavior layer was extended to support two voice primitives: (a) *synchronous text-to-speech*, which initiates voice synthesis with real-time feedback as utterances are synthesized; (b) *preprocess text*, which synthesizes speech and saves it in a persistent format for posterior playback. All primitives support simple text as well as SABLE [23]. SABLE supports marking of text emphasis, prosodic breaks, velocity, pitch and text volume configuration, among others.

[Figure 2: Integration of Festival with virtual human.]

Facial Expression

Facial expression builds on a parametrically controlled muscle-based model. Thirty seven muscles are defined and several atomic and group deformation parameters provided. Atomic parameters follow Waters [24] deformation model. Group parameters combine the effect of more than one atomic parameter. Parameters for the eyes, chin and tongue are defined including anthropometric limits on the rotation amplitudes. Emotion expression of six basic emotions [25] – anger, disgust, fear, joy, sadness and surprise – is supported. Emotions are implemented using group parameters. Regarding synchronization with vocal expression, a set of visemes was developed based on group deformation parameters. Mappings were defined for English and Portuguese phoneme alphabets. The co-articulation model, which considers two consecutive visemes, is characterized as follows: (1) from start time to half the duration of the respective phoneme, the face interpolates towards the viseme maximum amplitude; (2) from half the duration to end time of the respective phoneme, if the next phoneme is silence, the face interpolates to the neutral viseme, if it is not silence, the face interpolates to the combination between the current and next visemes.

Gesticulation Expression

This work proposes a feature-based model for gesticulation animation, i.e., gesticulation form is modeled as a sequence in time of constraints on hand shape, position and orientation. A feature-based approach is appropriate for several reasons. First, the necessity of describing gesticulation according to dimensions suggests that meaning distributes across the affordances of the upper limbs and hands and thus, rather than overall form a more granular description is possible. Second, a feature-based approach is compatible with most speech and gesture production models: the imagistic component in McNeill's [1] growth points ultimately materializes into gesture features; de Ruiter's model [26] revolves around the concept of gesture templates (in a gestuary) which correspond to constraints on features; finally, Krauss [27] actually considers knowledge representation as feature-based.

The model is further characterized as follows: (a) supports sub-second synchronization between speech and gesture; (b) supports automatic reproduction of annotated gesticulation according to a gesture transcription algorithm; (c) aims at real-time gesture animation; (d) focuses exclusively on gesticulation of the upper limbs and hands which account for most of the gesticulation seen in humans; (e) focuses on behavior rather than realism and thus, accurate anatomic models ([28][29]) are beyond the scope.

Features

Gesticulation is modeled as a sequence in time of *gesticulation keyframes*. Gesticulation keyframes define static (hand shape, position, orientation palm axis, orientation angle, and handedness) and dynamic (motion velocity) constraints on gesticulation features. Regarding implementation, the model relies on deterministic and non-deterministic expression. In concrete, limb robotic manipulators control the arms, hand position and orientation while pose animation players the hand shape.

The *hand shape* feature can assume any Portuguese Sign Language hand shape [30]. Furthermore, any two shapes can be combined and a parameter is provided to define how much each contributes. Implementation relies on pose player ability to combine stances and on a library of stances for Portuguese Sign Language shapes. The *position* feature is defined in Cartesian coordinates in three-dimensional space. Both a global world and local speaker references can be used. Hand shape orientation is associated with two features: *orientation palm axis*, which defines the normal to the palm; and *orientation angle* which defines a left handed angle about the normal. Implementation relies on inverse kinematics primitives. The *handedness* feature defines whether the gesticulation keyframe applies to the left, right or both hands. In the last case, features apply to the speaker's dominant hand and *symmetrical* values apply to the non-dominant hand. Symmetry can intuitively be understood as the gesticulation which would result if a mirror stood on the sagittal plane. Finally, *motion velocity* defines arm trajectories' velocity. Either joint angle or Cartesian velocity can be defined. Implementation of the joint angle and Cartesian velocity rely, respectively, on the joint-based interpolation and the Jacobian-based animation primitives.

As synchronization between speech and gesture is conveniently described at gesture phase level [1], the model supports *gesticulation phase keyframes*. The phase keyframe extends regular keyframes as follows: (a) a *duration* feature is added which defines total phase time; (b) *sequences of constraints* can now be defined for shape, position and orientation; (c) constraints within a sequence can be set to start at absolute time offsets relative to phase start time or at percentages of total phase duration. Notice, however, that phase keyframes could be converted into an equivalent sequence of regular keyframes.

Automatic Reproduction of Gesticulation Annotations

The gesticulation model supports automatic reproduction of *Gesture Recording Algorithm (GestuRA)* annotations. GestuRA, based on [9] and [31], is a linguistically motivated iterative algorithm for gesticulation form and meaning transcription which is structured in seven passes. First, speech is transcribed from the video-speech record. Second, text is organized into utterances. Third, utterances are classified according to discourse levels [1]. Fourth, gesticulation is filtered ignoring remaining gestures. Fifth, gesticulation phases are annotated. Sixth, gesticulation form is formally annotated. Finally, seventh, gesticulation is classified according to its dimensions and meaning analyzed. GestuRA integration with the gesticulation model is achieved through Anvil [32], a generic multimodal annotation tool, which exports annotations to a XML format which is, then, converted into the multimodal expression language, described below, for execution in virtual humans.

Environment Expression

The human mind affords emotion expression beyond the body. Soon the arts realized this as, for instance, in theatre where dramatic expression, text, sceneries, illumination, make-up, sound and music work together to tell a story. Leveraging on this knowledge, this work proposes an integrated model for storytelling and emotion expression through camera, lights and music. Stories are modeled as sets of *points of interest* which compete for the audience's attention. These can be of three kinds: (1) *characters*; (2) *dialogues*; (3) *sceneries*. Each point of interest has a priority which may change. The model receives as input a story which is told through the director and expression channels. The *director* focuses, at each instant, the audience to the highest priority point of interest. Then, the *environment expression channels* present it to the audience’s senses. Figure 3 summarizes this component.

[Figure 3: The environment expression model.]

Emotion Synthesis

In this work, all points of interest are endowed with emotional states. Character’s emotion synthesis is based on the Ortony, Clore and Collins (OCC) emotion theory [33]. The dialogue’s or scenery's emotional state is the average of all the participant characters’ emotional states. Explored global variables include arousal and mood.

Cinematography

Camera shots vary, among others, according to *distance* and *angle* [34]. Regarding distance, the closer the camera is, the higher the audience’s attachment to the point of interest. Regarding angle, three shots are representative: (1) *eye level* – the camera is at the same height as the point of interest, giving a neutral view; (2) *high angle* – the camera films the point of interest from above creating the impression of smallness and isolation; (3) *low-angle* – the camera films from below creating the impression of a powerful point of interest. In this work, the camera reflects the focused point of interest's strongest emotion as follows: (1) If it is anger or pride, a low-angle shot is chosen; (2) If it is fear, a high-angle shot is chosen; (3) The higher the emotion's intensity, the closer the chosen shot is.

Illumination

The *three-point-lighting* technique is widely used in movies to illuminate characters [36]. It is a configuration composed of the following light “roles”: (1) *key light*, which is the main source of light; (2) *fill light*, which is a low-intensity light that fills an area which otherwise would be in the dark; (3) *back light,* which separates character from background. This work applies this technique. The key light is a point light placed between the point of interest and the camera and emotion expression is achieved through parameter manipulation. Light color varies with strongest emotion according to research on emotion-color relations [35]. For instance, humans associate yellow with the sun and, thus, to something joyful. As it is known that well illuminated scenes are happy and poorly illuminated scenes are mysterious and sad [36], light brightness varies with strongest emotion intensity and valence.

Music

The relationship between music and emotion as been widely explored [37]. Regarding structural features, tempo is one of the most influencing factors. Fast tempo may be associated with happy/exciting emotions and slow tempo with sad/calm emotions. In this work music reflects the focused point of interest's mood valence – positive, neutral and negative. To convey mood valence, music, with the same valence, is randomly selected from a library. To fill the library, music was selected according to the following criteria: (1) Positive songs have fast tempo and, if applicable, positive lyrics; (2) Neutral songs have medium tempo; (3) Sad songs have slow tempo and, if applicable, negative lyrics.

Multimodal Expression

This work proposes a high-level integrated synchronized language - *Expression Markup Language (EML)* – to control virtual human multimodal expression. The language can be used in two ways, Figure 4: (1) as an *interface for a mind* which needs to express synchronously, in real-time and multimodaly through the body; (2) as a *script* which describes a story, written by a human or digital author, in real-time or not, where the virtual human expresses multimodaly. Synchronization between modalities, which is based in W3C's SMIL 2.0 [19], supports execution time definition relative to other modalities, in particular, relative to word or phonemes in vocal expression.

[Figure 4: EML integration with virtual humans.]

The "Hello World!" Example

Let us suppose the virtual human wishes to express "Hello world!". This distributes across modalities as follows: (1) Vocal expression must synthesize "Hello world!" emphasizing "world"; (2) Gesticulation expression must perform a beat gesture, synchronizing with "world", superimposed on a conventional hand salutation gesture; (3) Facial expression includes appropriate lip-synch and emotional expression of joy; (4) Environment expression must reflect the emotion of joy. One possible EML codification is as follows:

```
<voice-text time='0' timeId='t1'>
 <div>Hello<emph><tm name='t2' event='onStart'/>world!</emph></div>
</voice-text>
<gesture-key time='t1-0.1s' duration='0.2' handedness='right'>
 <hand-shapes><key time='0' id='7'/></hand-shapes>
 <palms><key time='0' a='200' x='0.0' y='0.0' z='-1.0'/></palms>
 <motion coords='speaker'>
  <key time='0' x='-7.0' y='45.0' z='3.0'/></motion>
</gesture-key>
<gesture-key time='t2' duration='0.2' handedness='right'>
 <hand-shapes><key time='0' id='7'/></hand-shapes>
```

```
12:   <palms><key time='0' a='120' x='0.0' y='0.0' z='-1.0'/></palms>
13:   <motion coords='speaker' >
14:    <key time='0.1' x='-15.0' y='40.0' z='3.0'/></motion>
15:  </gesture-key>
16:  <body-set-control-parameter time='t1'
17:   parameterId='Joy' value='1.0'/>
```

Lines 1-3 define vocal expression. The SABLE element “emph” is used to emphasize “world” and a time marker is associated with the start of the word. Gesticulation expression is divided into two keyframes. The first, in lines 4-9, raises the right hand above the head with an open-like claw Portuguese Sign Language hand shape. Notice the gesture starts 0.1s before “Hello”. The second, in lines 10-15, represents the beat-like gesticulation where the right hand abruptly moves to the right while still maintaining overall salutation form. Synchronization occurs with “world”. Regarding facial expression, lip-synch occurs automatically and the appropriate parameter for joy expression is set in lines 16-17. Synchronization occurs with “Hello”. Finally, regarding environment expression, even though not explicitly represented, the idea is that the joy emotion would be elicited at the start of “Hello” leading to a closer camera shot, an increase in illumination brightness, a change in light color to yellow and, finally, a positively valenced music fade in.

Evaluation

Two studies were conducted to evaluate gesticulation expression. In both cases, the idea consisted of comparing the narration of the Portuguese traditional story “The White Rabbit” by a human storyteller with a version by a virtual storyteller. The first study, conducted in the scope of the Papous project at Inesc-ID, consisted of presenting the full narration video to the subject and, then, presenting a questionnaire to classify real and synthetic gestures according to contribution to story comprehension, emotion expression success, believability and subject satisfaction. The study was presented to 108 subjects. Results show that synthetic gestures fared well when compared to real gestures. The second study, presented the narration video divided into segments. Segments' narration was equally distributed between the real and synthetic storytellers. After each segment a set of interpretation questions were posed. In the end, the subject was asked to choose the preferred storyteller. The study was presented to 39 subjects. Results indicate that the human storyteller was preferred to the synthetic and that interpretation did not significantly differ between storytellers. Further details relating to these studies can be found in [38].

One study was conducted to evaluate environment expression. The study was based on cartoon-like story application. The study evaluated emotion interpretation with varying configurations of expression channels, music valence interpretation and story preference with or without environment expression. The study was presented to 50 subjects. Results indicate that illumination was the most influencing channel in emotion interpretation, that tempo and lyrics reasonably predict music valence and that people prefer stories told with environment expression. Further details relating to this study can be found in [39].

Conclusions and Future Work

This work proposed a real-time virtual human multimodal expression model. Six modalities are explored. Deterministic expression supports keyframe animation and several animation combination mechanisms. Non-deterministic expression supports procedural animation based on robotics algorithms. Vocal expression supports speech synthesis and voice parameterization through SABLE. Facial expression builds on a muscle-based model and supports lip-synch and emotion expression. Gesticulation expression supports a feature-based description of gesticulation as sequences of Portuguese Sign Language hand shapes, hand positions and orientations. Finally, environment expression goes beyond the body to express emotions through three environment channels – camera, lights and music.

Naturally, this work can be improved. Regarding non-deterministic expression, first, seven degrees-of-freedom limb manipulators should be considered as this allows for better elbow and knee control which leads to more natural motion; second, dynamics could be modeled to produce physically realistic expression. Regarding voice expression, high-level emotional parameters, which SABLE doesn’t support, should be considered. Regarding facial expression, more emotions could be explored as well as emotion combination. Regarding gesticulation expression, first, it is necessary to go beyond arms and hands and explore other body parts; second, further features could be explored like, for instance, velocity profiles [12]; third, preparation and retraction motion could be automatically generated; finally, an anatomic-based hand model with appropriate constraints, muscle and tendons simulation ([28][29]) would lead to more realistic gesticulation. Regarding environment expression, the cinematography channel mapping between emotions and camera shots can be extended to include more emotions and more kinds of shots; the illumination channel should explore shadows as these can be very expressive [36]; the music channel should explore other music parameters [37] (mode, loudness, rhythm, etc.) and relate them to more emotion properties (arousal, emotion type, etc.). Finally, regarding multimodal synchronization, automatic generation of gesticulation co-articulation effects would improve integration with speech.

Acknowledgments

This research was partially supported by the Papous project at Inesc-ID (Ref.: POSI / SRI / 41071 / 2001).

References

[1] Gratch J, Rickel J, Andre E, Badler N, Cassell J, Petajan E. Creating Interactive Virtual Humans: Some Assembly Required. In *IEEE Intelligent Systems*, 17(4):54-63, 2002

[2] Cavazza M, Earnshaw R, Magnenat-Thalmann N, Thalmann D. Motion Control of Virtual Humans. In *IEEE Computer Graphics and Applications*, vol.18 (5), pp.24-31, 1998

[3] Noh J, Neumann U. A Survey of Facial Modeling and Animation Techniques. USC Technical Report, pp.99-705, 1998

[4] Blumberg B, Galyean T. Multi-Level Direction of Autonomous Creatures for Real-Time Virtual Environments. In *Proceedings of SIGGRAPH '95*, 30(3):47-54, 1995

[5] Perlin K, Goldberg A. Improv: A System for Scripting Interactive Actors. *Virtual Worlds in Proceedings of SIGGRAPH'96*, pp.205-216, 1996

[6] Badler N, Phillips C, Webber B. Simulating Humans: Computer Graphics, Animation, and Controls. Oxford University Press, 1992

[7] Kalra P, Magnenat-Thalmann N, Moccozet L, Sannier G, Aubel A. Real-Time Animation of Realistic Virtual Humans. In *IEEE Computer Graphics and Applications*, Vol.18, No.5, pp.42-55; 1998

[8] McNeill D. Hand and Mind: What gestures reveal about thought. University of Chicago Press, 1992

[9] McNeill D. Gesture and Thought. University of Chicago Press, 2005

[10] Cassell J, Pelachaud C, Badler N, Steedman M, Achorn B, Becket T, Douville B, Prevost S, Stone M. Animated Conversation: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agent. In *Proceedings of SIGGRAPH'94*, pp.413-420, 1994

[11] Cassell J, Bickmore T, Billinghurst M, Campbell L, Chang K, Vilhjálmsson H, Yan H. Embodiment in Conversational Interfaces: Rea. In *Proceedings of the CHI'99 Conference*, pp. 520-527, Pittsburgh, PA, 1999

[12] Kopp S, Wachsmuth I. A knowledge-based approach for lifelike gesture animation. In *Proceedings of the 14th European Conference on Artificial Intelligence*, Amsterdam, IOS Press, 2000

[13] Kopp S, Tepper P, Cassell J. Towards Integrated Microplanning of Language and Iconic Gesture for Multimodal Output. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI'04)*, pp.97-104, ACM Press, 2004

[14] Amaya K, Bruderlin A, and Calvert T. Emotion from motion. *Proceedings Graphics Interface'96*, pages 222-229, 1996

[15] Rose C, Bodenheimer B, Cohen M. Verbs and Adverbs: Multidimensional Motion Interpolation. In *IEEE Computer Graphics and Applications*, vol. 18 (5), pp.32-40, 1998

[16] Chi D, Costa M, Zhao L, Badler N. The EMOTE model for effort and shape. *Proceedings of SIGGRAPH '00*, pp.173-182, New Orleans LA, 2000

[17] Arafa Y, Kamyab K, Mamdani E. Character Animation Scripting Languages: A Comparison, 2003

[18] VHML. VHML – Virtual Human Markup Language. www.vhml.org/

[19] SMIL. SMIL: Synchronized Multimedia. www.w3.org/AudioVideo/

[20] Krandsted A, Kopp S, Wachsmuth I. MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents. In *Proceedings of the AAMAS Workshop on "Embodied conversational agents -- Let's specify and evaluate them!"*, Bologna, 2002

[21] NASA Man-Systems Integration Manual (NASA-STD-3000)

[22] Festival. The Festival Speech Synthesis Systems. www.cstr.ed.ac.uk/projects/festival/

[23] SABLE. SABLE: A Synthesis Markup Language – version 1.0. www.bell-labs.com/project/tts/sable.html

[24] Waters K. A muscle model for animation three-dimensional facial expression. *Proceedings of the 14th Annual Conference on Computer Graphics and interactive Techniques M. C. Stone, Ed. SIGGRAPH 1987*. ACM Press, New York, NY, 17-24, 1987

[25] Ekman P. Facial Expressions. In T. Dalgleish, M. Power, eds., *Handbook of Cognition and Emotion*, New York: John Wiley & Sons Ltd, 1999

[26] de Ruiter J. The production of gesture and speech. In D. McNeill, ed., *Language and gesture*, pp.284-311. Cambridge University Press, 2000

[27] Krauss M, Chen Y, Gottesman R. Lexical gestures and lexical access: A process model. In D. McNeill, ed., *Language and gesture*, pp.261-283. Cambridge University Press, 2000

[28] Thompson D, Buford W, Myers L, Giurintano D, Brewer III J. A Hand Biomechanics Workstation. *Computer Graphics*, vol.22, no.4, pp.335-343, 1988

[29] Albrecht I, Haber J, Seidel H. Construction and Animation of Anatomically Based Human Hand Models. *SIGGRAPH 2003*, pp. 98-109, 2003

[30] Secretariado Nacional para a Reabilitação e Integração das Pessoas com Deficiência. Gestuário - Língua Gestual Portuguesa – 5th edition.

[31] Gut U, Looks K, Thies A, Trippel T, Gibbon D. CoGesT – Conversational Gesture Transcription System. Technical Report, University of Bielefeld, 1993

[32] Kipp M. ANVIL – A Generic Annotation Tool for Multimodal Dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg, pp.1367-1370, 2001

[33] Ortony A, Clore G, Collins A. The Cognitive Structure of Emotions. Cambridge University Press, 1988

[34] Arijon D. Grammar of the Film Language. Silman-James Press, 1976

[35] Kaya N. Relationship between color and emotion: a study of college students. College Student Journal, 2004

[36] Birn J. [digital] Lighting & Rendering. New Riders, 2000

[37] Juslin P, Sloboda J. Music and Emotion: theory and research. Oxford University Press, 2001

[38] de Melo C, Paiva A. A Story about Gesticulation Expression. Submitted, 2006

[39] de Melo C, Paiva A. Environment Expression: Expressing Emotions through Cameras, Lights and Music. *Proceedings of Affective Computing Intelligent Interaction (ACII05)*, pp.715-722, Beijing, China, 2005

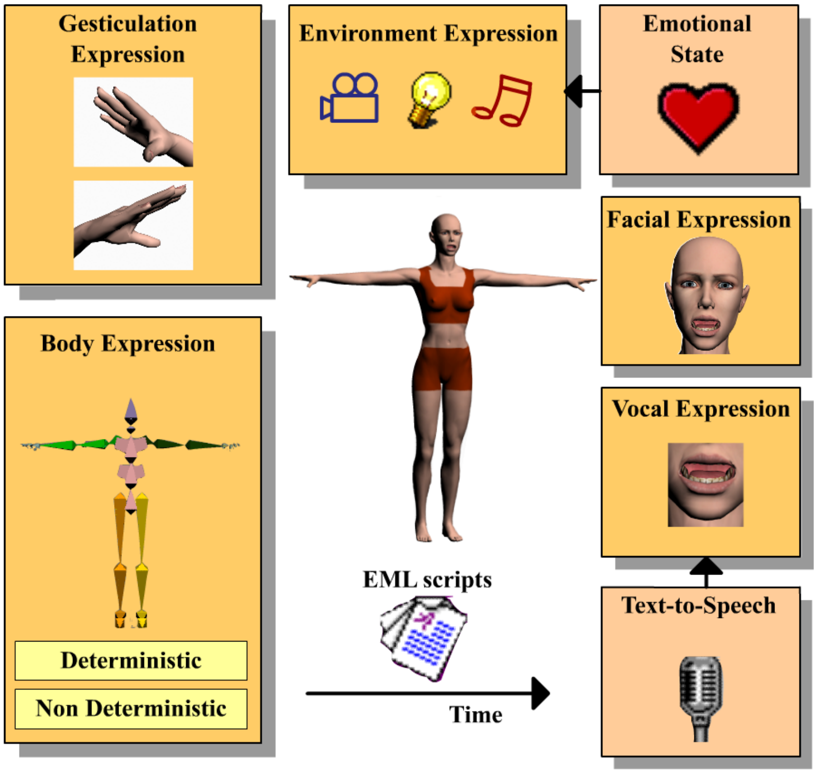

Figure 1 The multimodal expression model.

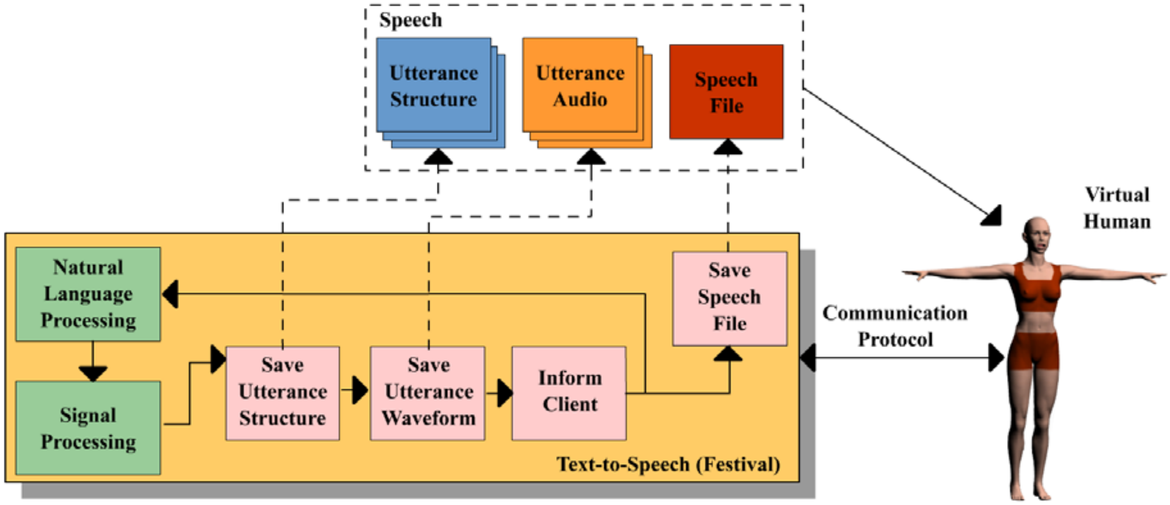

Figure 2 Integration of Festival with virtual humans.

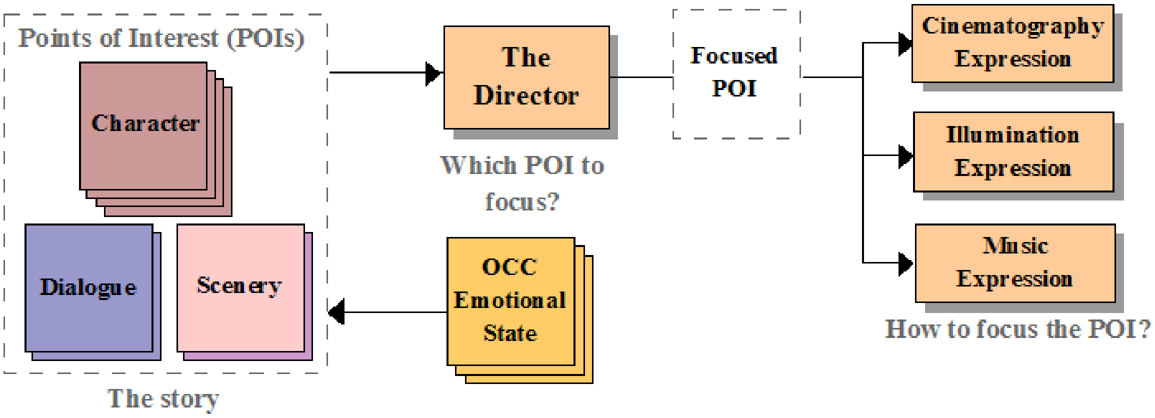

Figure 3 The environment expression model.

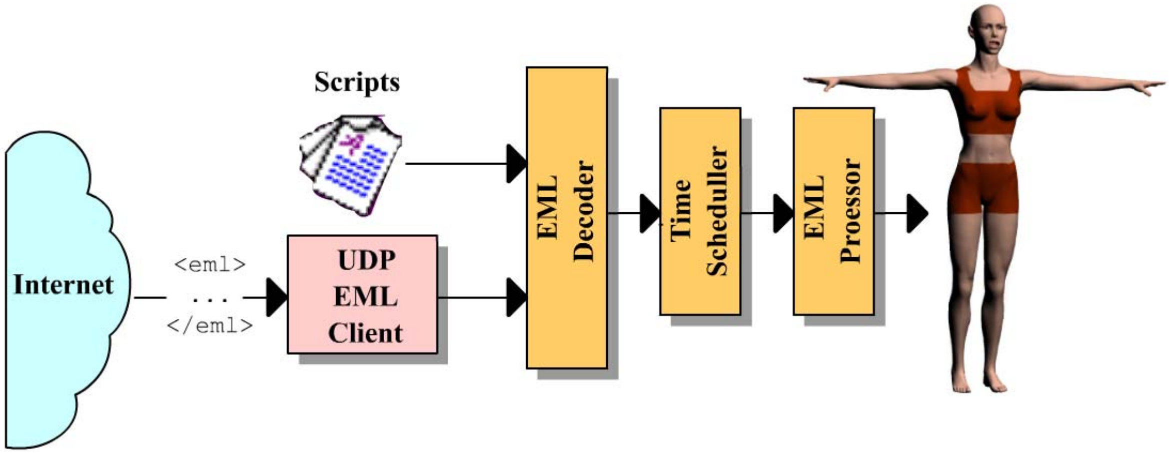

Figure 4 EML integration with virtual humans.