Modeling Gesticulation Expression in Virtual Humans

Celso M. de Melo[1], Ana Paiva[2]

[1] USC, University of Southern California
demelo@usc.edu
[2] IST – Technical University of Lisbon and INESC-ID,
Avenida Prof. Cavaco Silva – Taguspark,
2780-990 Porto Salvo, Portugal
ana.paiva@inesc-id.pt

Abstract. Gesticulation is the kind of unconscious, idiosyncratic and unconventional gestures humans do in conversation or narration. This chapter reviews efforts made to harness the expressiveness of gesticulation in virtual humans and proposes one such model. First, psycholinguistics research is overviewed so as to understand how gesticulation occurs in humans. Then, relevant computer graphics and computational psycholinguistics systems are reviewed. Finally, a model for virtual human gesticulation expression is presented which supports: (a) real-time gesticulation animation described as sequences of constraints on static (Portuguese Sign Language hand shapes, orientation palm axis, orientation angle and handedness) and dynamic features; (b) synchronization between gesticulation and synthesized speech; (c) automatic reproduction of annotations in GestuRA, a gesticulation transcription algorithm; (d) expression control through an abstract integrated synchronized language – Expression Markup Language (EML). Two studies, which were conducted to evaluate the model in a storytelling context, are also described.

1 Introduction

Humans express thought through gesticulation. Gesticulation is the kind of unconscious, idiosyncratic and unconventional gestures humans do in conversation or narration. They tend to focus on the arms and hands, though other body parts may be involved. Furthermore, gesticulation and speech, which are believed to be different sides of the same mental process, co-express the same underlying idea unit and synchronize at various levels. [1,2]

The problem of modeling gesticulation can be divided into the sub-problems of generation and execution. *Gesticulation generation* concerns with the simulation of the speech and gesture production process, i.e., the distribution of communicative intent across modalities and selection of proper surface realizations which, in the case of gestures, correspond to constraints on static and dynamic features of the arms and hands. *Gesticulation execution* is more akin to the body and concerns with the actual animation, in a synchronized fashion, of the static and dynamic constraints which define the gesture. This chapter will focus on the latter but, overview the former.

To clarify the challenges involved in this endeavor, a virtual human model for gesticulation expression is described. The model supports: (a) real-time gesticulation animation described as sequences of constraints on static (Portuguese Sign Language hand shapes, orientations and positions) and dynamic features; (b) multimodal synchronization, including between gesticulation and speech; (c) automatic reproduction of annotated gesticulation according to *GestuRA*, a gesture transcription algorithm; (d) expression control through a markup integrated synchronized language.

The rest of the text is organized as follows: section 2 overviews gesticulation research in psycholinguistics; section 3 describes relevant computational models; section 4 presents a gesticulation expression model for virtual humans as well as an evaluation study; finally, section 5 draws conclusions and discusses future work.

2 Gesticulation and Psycholinguistics

Human gestures can be categorized into three subclasses [3]: gesticulation; emblems; and signs. *Emblems* are culturally dependent gestures which have conventionalized meaning. An example is the American V (of victory) gesture, executed with the palm facing the listener. *Sign languages* consist of communication languages expressed through visible hand gestures. Examples are languages used by the deaf, such as the Portuguese Sign Language [4]. Finally, *gesticulation*, which is the focus of this chapter, is the kind of idiosyncratic, unconventional and unconscious gestures humans do in narrations or conversations [1,2]. Gesticulation is tightly synchronized with speech, is structured in phases and can be interpreted according to several dimensions.

2.1 Gesticulation and Speech

Gestures which occur when a person is speaking manifest verbal thought. Verbal thought, which does not include all forms of thought, nor all forms of speech, is the kind of thought which resides in the intersection between thought and speech. It is believed that speech and gesticulation are manifestations of the same underlying process [1,2]. Thus, gesticulation and speech co-express the same underlying idea unit possibly in non-redundant ways, as they synchronize at the semantic and pragmatic levels, develop together in childhood and deteriorate together in aphasia. Through gesticulation, however, information is conveyed in a fundamentally different way than through speech: (a) gesticulation is not combinatoric – two gestures produced together do not combine to form a larger one with a complex meaning; (b) there is no hierarchical structure in gesticulation as in language; (c) gesticulation does not share the linguistic properties found on verbal communication.

2.2 Gesticulation Structure

According to how it unfolds in time, gesticulation can be structured hierarchically into units, phrases and phases [5,6]. A *unit*, which is the highest level in the hierar-

chy, is the time interval between successive rests of the limbs. A unit may contain various phrases. A *phrase* is what is intuitively called 'gesture' [2]. A phrase consists of various *phases*: (a) preparation, where the limbs position themselves to initiate the gesture; (b) pre-stroke hold, where a hold occurs just before the stroke; (c) stroke, which is the only obligatory phase, where actual meaning is conferred. The stroke is synchronous with its co-expressive speech 90% of the time [7] and, when asynchronous, precede the semantically related speech; (d) post-stroke hold, where a hold occurs after the stroke, before initiating retraction; (e) retraction, where the limbs return to the resting position. Preparation, stroke and retraction were introduced by Kendon [8] and the holds by Kita [9].

2.3 Gesticulation Dimensions

McNeill and colleagues characterize gesticulation according to four dimensions [1,2]: (1) *iconicity*, which refers to gesticulation features which demonstrate through its shape some characteristic of the action or event being described; (2) *metaphoricity*, which is similar to iconics however, referring to abstract concepts; (3) *deixis*, which refers to features which situate in the physical space, surrounding the speaker, concrete and abstract concepts in speech; (4) *beats*, which refer to small baton like movements that do not change in form with the accompanying speech. They serve a pragmatic function occurring, for instance, with comments on one's own linguistic contribution, speech repairs, and reported speech. According to McNeill ([2], 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 exclusive categories and hierarchy.

2.4 Gesticulation Models

Several gesticulation production models have been proposed. McNeill's *growth point model* [1,2] explains verbal thinking through growth points, which represent idea units. In a growth point two unlike modes of thinking – linguistic and imagistic – are active and this instability resolves by accessing stable language forms and materializing into gesture. This materialization increases with the unpredictability of the idea unit, i.e., with its opposition to current context. In contrast, extending Levelt's *speaking* model [10], various modular information processing models have been proposed including by de Ruiter, Krauss and Kita & Özyürek. In *de Ruiter's sketch model* [11] the conceptualizer – which transforms communicative intent into a propositional form called the preverbal message – receives as input communicative intent and outputs a *sketch* holding gesture form specifications. These specifications rely on a *gestuary* which stores predefined gesture templates which impose constraints on features. Synchronization is achieved through signal passing between modules. Krauss's model [12], contrasting to de Ruiter's and McNeill's assumption of imagistic knowledge, is a *featural model*, i.e., concepts are represented as propositional and non-propositional (visuospatial) features. During gesture production, a subset of the non-propositional features is selected to pass down to a motor planner which generates form. The model

also describes gesture effects on lexical retrieval. As in de Ruiter's model, synchronization is achieved through signal passing. Kita and Özyürek [13] propose a model which says, contrasting to Krauss' and de Ruiter's models, that gestures are influenced by the speaker's language.

2.5 Implications for Computer Science

The psycholinguistics research presented in this section leads to several requisites for a computational model of gesticulation:

- *Gesticulation should, at least, span arms and hands*, as it tends to focus in these body parts;
- *Gesticulation and speech should be able to synchronize at the sub-second time granularity*, as they are believed to be different sides of the same underlying mental process and synchronize at the semantic and pragmatic levels;
- *It should be possible to describe gesticulation at the phase level*, as they distinguish parts which are motivated by physical, synchronization or meaning constraints. Phases are also crucial for gesture fusion in co-articulation effects;
- *Gesticulation can be described through constraints on its features*, in concrete, as sequences of static (hand shape, orientation and position) and dynamic constraints (motion profiles). The feature-based approach is justified for several reasons. First, describing gesticulation according to dimensions and saliency suggests that meaning distributes across the affordances of the upper limbs and hands and thus, rather than overall form a more granular (or feature-based) description is possible. Second, a feature-based approach is compatible with most speech and gesture production models: the imagistic component in McNeill's growth points ultimately materializes into gesture features; de Ruiter's sketch model revolves around the concept of gesture templates (in a gestuary) which correspond to constraints on features; Krauss actually considers knowledge representation as feature-based; finally, Kita & Özyürek even though not detailing gesture morphology, motivate their model with motion gestures described according to features.

3 Gesticulation and Computer Science

Building a gesticulation expression computational model comes with many challenges, Fig. 1. First, it is necessary to build a *virtual human* which has a body which can be animated to gesticulate. This challenge is in the domain of computer graphics. Second, it is necessary to solve the *gesticulation execution* problem, which concerns with animating a gesticulation plan. Third, it is necessary to solve the *gesticulation production* problem, which isn't independent of speech production and concerns with converting communicative intent into synchronized verbal and gesticulation plans. Building on virtual human models, computational psycholinguistics systems address these last two issues. Finally, *interfaces* should be built between these layers to pro-

mote modularity. In this regard, several markup languages have been proposed. Section 4 describes one approach which focuses on the gesticulation execution problem.

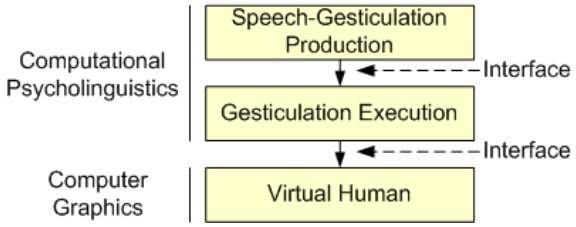

Fig. 1. A framework for gesticulation expression

3.1 Gesticulation and Computer Graphics

In its simplest form, building a virtual human consists of defining a hierarchical skeleton and a mesh for the skin. Animating the skeleton leads to skin mesh deformation using the vertex blending technique [14]. Several animation mechanisms have been explored [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.

Particularly relevant to gesticulation are specialized models of the hands. Thompson [16] proposes an anatomically accurate hand model based on tomographic scans. Wagner [17] argues for individual models of the hand by comparing the anthropometry of pianists with regular people. Moccozet [18] proposes a three-layer hand model – skeleton, muscle and skin – where Dirichlet free-form deformations are used to simulate muscle and realistic skin deformation. Sibille [19] proposes a real-time generic anatomical hand model. Hand motion is based on dynamics, mass-spring meshes are used to calculate soft tissue deformations and, finally, the system handles collision detection. Finally, Albrecht [20] also proposes a real-time anatomical human hand model. Motion relies on a realistic muscle model based on anatomical data, mechanical laws and a mass-spring system. Even though these models have great potential to generate realistic gesticulation, thus far, most computational psycholinguistics systems have used far simpler hand models.

3.2 Gesticulation and Computational Psycholinguistics

Building on the aforementioned graphic models, several computational psycholinguistics systems have been proposed to address the gesticulation production and execution problems. Animated Conversation [21], developed by Cassell and colleagues, is a rule-based system capable of synchronizing gestures of the right type with co-occurring speech. *Real Estate Agent (Rea)* [22,23] presents an embodied conversa-

tional agent capable of proper distribution and realization of communicative intent across speech and gesture. Cassell et al. [24] also propose the *Behavior Expression Animation Toolkit (BEAT)* which receives as input text and, based on rules, automatically generates appropriate synchronized nonverbal behavior. Kopp and colleagues [25,26] developed a comprehensive model for gesture animation based on research in psycholinguistics and motor control theory. Here, a knowledge base, similar to de Ruiter's gestuary [11], holds gesture templates which consist of hierarchies of constraints on static and dynamic features of the stroke phase. Gesture production instantiates templates and feeds them into a motor planner for execution. Preparation, retraction and co-articulation effects are automatically appended. The model supports sophisticated arm trajectories including velocity profiles. The system also supports speech parameterization through SABLE [27]. Recently, Cassell, Kopp and colleagues brought together the best from the aforementioned systems in *NUMACK* [28], a system capable of synthesizing in real-time co-verbal context-sensitive iconic gestures without relying on a library of predefined gestures. Though the gesture-speech production process is not the focus of the chapter, the underlying gesticulation animation model in these systems shares several aspects with the model presented in section 4, namely: the requisites are based on psycholinguistics research and similar static and dynamic features are explored.

3.3 Interface Languages

Controlling and integrating gesticulation expression with other modalities is usually solved through markup languages [29]. The idea is that the gesticulation production process communicates the gesticulation plan, created from the communicative intent, to the gesticulation execution process, which animates it, through this language. The language, thus, supports a convenient clear-cut separation between these processes. Presently, no such standard language exists. The community has acknowledged this and has begun to address it. A promising effort is the SAIBA framework [30] which brings together several research groups. Unfortunately, this standard is still in its infancy and, therefore, the model presented in section 4 requires, for the time being, yet another control language – *Expression Markup Language (EML)*. This language is particularly influenced by: VHML [31], SMIL [32] and MURML [33]. Regarding Virtual Human Markup Language (VHML), this work reuses the notion of organizing control according to modality-specific modules. Regarding Synchronized Multimedia Integration Language (SMIL), which is oriented towards audiovisual interactive presentations, this work uses a similar modality synchronization mechanism. Regarding Multimodal Utterance Representation Markup Language (MURML), this work defines a similar notation for gesture specification and synchronization with co-verbal speech. Finally, in contrast to high-level languages such as GESTYLE [34] which tries to capture the individual's expression style and APML [35] which represents, among others, communicative intent, emotions, interaction and cultural aspects, the proposed language focuses on speech synthesis and low-level body control such as gesticulation animation as sequences of constraints on static and dynamic features.

4 A Model for Gesticulation Expression

This section describes a gesticulation expression model for virtual humans which supports: (a) real-time gesticulation animation described as sequences of constraints on static (Portuguese Sign Language hand shapes, orientations and positions) and dynamic features; (b) multimodal synchronization between gesticulation and speech; (c) automatic reproduction of annotated gesticulation according to GestuRA, a transcription algorithm; (d) expression control through an abstract integrated synchronized language. The model builds on top of a virtual human architecture, which provides keyframe and procedural animation mechanisms, and integrates with other expression modalities including speech, Fig. 2.

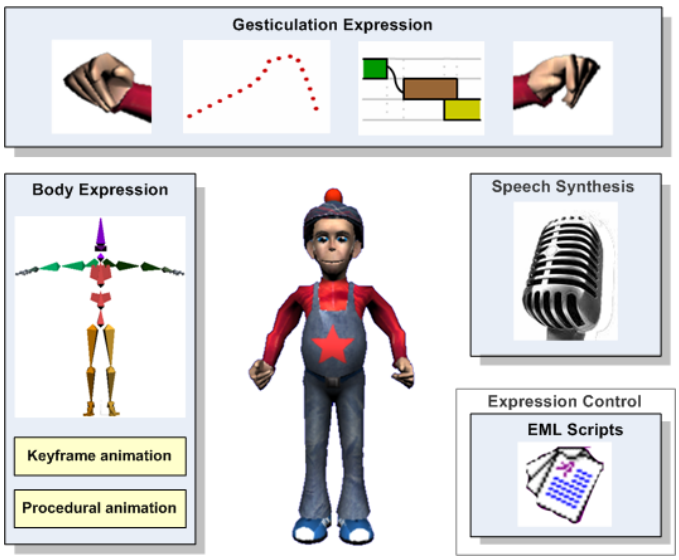

Fig. 2. Gesticulation expression model overview

The remainder of this section is organized as follows: subsection 4.1 describes the virtual human architecture; subsection 4.2 describes speech synthesis and integration with gesticulation; subsection 4.3 describes the feature-based gesticulation model itself; subsection 4.4 describes multimodal expression control using a markup scripting language and subsection 4.5 describes two evaluation studies.

4.1 Virtual Humans

The virtual human is structured according to a three-layer architecture [36,37]. The *geometry layer* defines a 54-bone human-based skeleton. The *animation layer* defines keyframe and procedural animation mechanisms. The *behavior layer* defines speech and gesticulation expression and supports a language for integrated synchronized multimodal expression.

4.1.1 Keyframe animation

Keyframe animation animates the virtual human according to predefined sequences of poses usually designed by human artists. The model generates in-between poses, supports several animation combination mechanisms but, ultimately, is not independent of the human creator and, thus, is not very flexible. Still, keyframe animation is useful for gesticulation expression in the following situations: (a) animation of complex gesticulation which is cumbersome to model through features; (b) animation of gesticulation which involves body parts other than arms and hands. Keyframe animation 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.

4.1.2 Procedural animation

Procedural animation consists of animating the virtual humans by controlling the limbs' extremities. Procedural animation is at the core of flexible gesticulation expression as it provides the means to position and orient the hands arbitrarily in space according to specific motion profiles. Notice this flexibility isn't possible using keyframe animation. Procedural animation is based on robotics techniques [38]. 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 [39]. 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.

4.2 Voice Synthesis

Voice synthesis is based on the Festival [40] text-to-speech system. Festival features facilitate integration with gesticulation as they include: (a) a simple Scheme programming interface; (b) server/client interaction through sockets thus, supporting clients in other programming languages; (c) access to synthesized utterance structure (words, phonemes, times, etc.), which synchronizes with gesticulation phases, and the ability to save this data in files; (d) incremental real-time synthesis, thus, allowing the virtual human to schedule gesticulation while speech is being synthesized; (e) limited

support for SABLE [27] which allows definition of speech emphasis, prosodic breaks, velocity, pitch, text volume configuration, among others.

Festival integration with the virtual human involves four aspects, Fig. 3: (1) the notion of speech; (2) an extension to Festival's voice synthesis pipeline; (3) a communication protocol; (4) a new behavior layer API for speech 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. Using Festival's programming interface, the voice synthesis pipeline is extended, after natural language and signal processing, with the following steps: after each utterance has been synthesized, its structure and waveform are saved and the virtual human is informed that an utterance is ready to play; after all utterances have been synthesized, the speech file is saved and the virtual human is informed about speech synthesis completion. The communication protocol is characterized as follows: (a) supports voice synthesis primitives; (b) supports incremental utterance conclusion communication; (c) supports communication of speech synthesis conclusion. At the virtual human side, the behavior layer was extended to support two voice primitives: (1) *synchronous text-to-speech*, which initiates voice synthesis with real-time feedback as utterances are synthesized; (2) *preprocess text*, which synthesizes speech and saves it in a persistent format for posterior playback. Both primitives support SABLE.

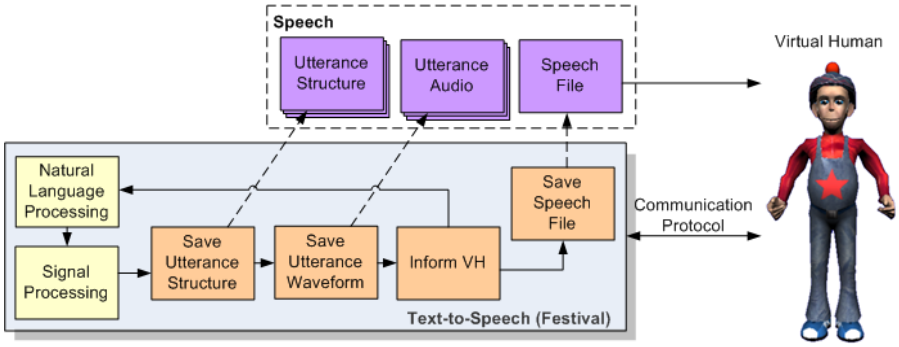

Fig. 3. Speech-synthesis integration with Festival

4.3 Gesticulation expression

The gesticulation expression model controls arms and hands and relies on keyframe and procedural animation. Precisely, limb manipulators control the arms, hands' position and orientation while pose animation players control the hands' shape. The model is feature-based, i.e., gesticulation form is modeled as a sequence in time of constraints on static and dynamic features. Features are described on subsection 4.3.1. Motion modifiers influence the interpretation of otherwise neutral gesticulation. Modifiers are described on subsection 4.3.2. The model supports multimodal synchronization, in particular, between speech and gesture. Synchronization is described

on subsection 4.3.3. Finally, the model supports automatic reproduction of annotated gesticulation according to GestuRA, a gesture transcription algorithm. GestuRA and its integration with the model are described on subsection 4.3.4.

4.3.1 Features

Gesticulation is modeled as a sequence in time of constraints on static and dynamic features. Static features are represented in *gesticulation keyframes* and include: hand shape, position, orientation palm axis, orientation angle, and handedness. Dynamic features define motion profiles for keyframe interpolation.

Regarding static features, the *hand shape* feature can assume any Portuguese Sign Language hand shape [4]. 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 world and speaker references can be used. Hand shape orientation is defined by two features: *orientation palm axis*, which defines the palm's normal; 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, remaining features apply to the speaker's dominant hand and symmetrical values apply to the non-dominant hand. Symmetry is intuitively understood as the gesticulation which would result if a mirror stood on the sagittal plane.

Regarding dynamic features, the model supports (keyframe) interpolation through parametric cubic curves, which can represent any kind of velocity profile, such as deceleration near the target position and overshooting effects which we see in humans [41]. Currently, the model supports Bézier and Hermite cubic curves, as well as piecewise combinations thereof. Furthermore, interpolators can be structured into hierarchies thus, leading to sophisticated motion profiles. Moreover, either Cartesian or joint angle velocity can be used. Implementation of interpolation in Cartesian and joint angle space relies, respectively, on the frame interpolation and joint interpolation procedural animation control primitives.

4.3.2 Modifiers

Several researchers have explored *motion modifiers* which add emotive qualities to existent motion data. Signal-processing techniques [42,43,44] were used to extract information from motion data which is used to generate emotional variations of neutral motion. Rose and colleagues [45] generate new motion with a certain mood or emotion from motion data interpolation based on radial functions and low order polynomials. Chi and colleagues [46] 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. Finally, Hartmann [47] draws from psychology six parameters for gesture modification: *overall activation*, which refers to the quantity of movement during a conversational turn; *spatial extent*, which refers to the amplitude of movement; *temporal extent*, which refers to the duration of movements; *fluidity*, which refers to smoothness and continuity of movement;

power, which refers to how strong or weak the movement appears; and *repetition*, which refers to rhythmic repeats of specific movement.

The effect of these modifiers can be simulated resorting to the static and dynamic features described above. However, in digital worlds, motion modifiers need not be limited to the body. Thus, inspiring in the arts, we've explored a different set of modifiers which rely on properties of the surrounding environment [48] – such as camera, lights and music – and the screen [49] – such as the virtual human pixels themselves – to convey emotional interpretations to virtual human movement. These modifiers are, however, detailed elsewhere [48, 49].

4.3.3 Synchronization

Sub-second synchronization of gesture phases with speech relies on a control markup language – Expression Markup Language (EML) – which supports phoneme-level synchronization. The language integrates with SABLE [27] and thus, supports synchronization with speech properties such as intonation contour. Similarly to SMIL [32], modality execution time can be set to absolute or modality relative values. Furthermore, named timestamps can be associated with text to be synthesized. The following events can be associated with named timestamps: (a) start of a word; (b) end of a word; (c) start of a phoneme. EML is detailed on subsection 4.4.

As synchronization between speech and gesture is conveniently described at the gesture phase level, the model supports explicit *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 associated to shape, position and orientation features; (c) constraints within a sequence can be set to start at absolute time offsets relative to phase start time or at percentages of the total phase duration. However, phase keyframes do not add expressiveness to the model in the sense that gesticulation described with phase keyframes could be converted into an equivalent sequence of regular keyframes.

4.3.4 Automatic reproduction of gesticulation annotations

The gesticulation model supports automatic reproduction of *Gesture Recording Algorithm (GestuRA) annotations*. GestuRA, based on [2] and [50], is a linguistically motivated iterative algorithm for gesticulation *form* and *meaning* transcription. The former refers to the kinesthetic properties, whereas the latter to the interpretation of the gesture. GestuRA is structured into seven passes, Fig. 4. First, speech is transcribed from the video-speech record. Second, text is organized into utterances. Third, utterances are classified according to discourse levels – narrative, metanarrative and paranarrative [1]. Fourth, gesticulation is filtered ignoring remaining gestures (such as emblems, for instance). Fifth, gesticulation phases are annotated. Sixth, gesticulation form is formally annotated. Finally, seventh, gesticulation is classified according to its dimensions and its meaning analyzed.

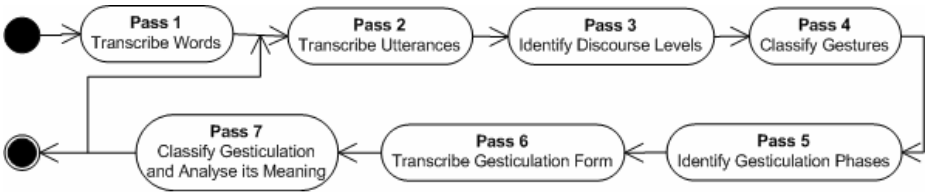

Fig. 4. Overview of the Gesture Recording Algorithm (GestuRA)

GestuRA integration with the gesticulation model is achieved through *Anvil* [51], a generic multimodal annotation tool. In concrete, implementing GestuRA in Anvil benefits from its capability of exporting annotations to a XML format. This format can, then, be converted into EML for immediate execution in virtual humans, Fig. 5.

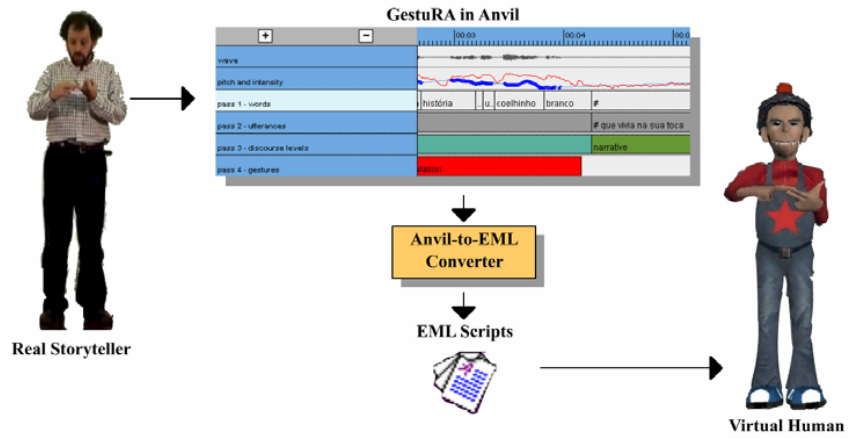

Fig. 5. Gesticulation model integration with GestuRA

Automatic reproduction from GestuRA is valuable for various reasons. First, reproduction from transcribed annotations is flexible. Usually annotation algorithms are used to build databases of human gestures. Thus, all gesture details need to be formalized and are, usually, classified according to form and meaning. Therefore, reproduction from such an annotation can selectively choose which information to use according to context. For instance, if we want to provide a virtual human with a certain style, we could disregard form annotation and simply reproduce gestures from the annotated meaning but, using a stylized form. This flexibility contrasts with reproduction from automatic gesture recognition algorithms [52,53], which accurately recognize form but, are still limited with respect to meaning interpretation. Automatic reproduction is also useful to test transcription accuracy and, furthermore, constitutes an important evaluation tool for the gesticulation expression model. As speech and gesture production from communicative intent is not simulated, an alternative to evaluating the model is to compare it to actual real-life videos.

4.4 Multimodal Expression control

This work proposes a markup, integrated and synchronized language – *Expression Markup Language (EML)* – which serves as a control interface for the body. The language can be used in two ways, Fig. 6: (1) as an *interface for a mind* which needs to express, in real-time, synchronously and multimodaly through the body; (2) as a *script* which describes a story, written by an author, where the virtual human expresses multimodaly. In the first case, the mind communicates to the body in real-time, through a socket or API, a set of EML clauses which are immediately executed. The gesticulation production process is meant to integrate with the execution process in this way (see section 3). In the second case, the script defines a sequence of clauses, temporally ordered, which defines a story which can be played later by different virtual humans. Regarding specification, EML is a markup language structured into modules: (1) *core*, defines the main elements; (2) *time and synchronization*, defines multimodal synchronization and is characterized as follows: (a) supports execution time definition relative to other clauses; (b) supports execution time definition relative to word or phoneme in vocal expression clauses; (c) supports loops; (d) supports parallel and sequential execution. This module is based on W3C's SMIL 2.0 specification [32]; (3) *body*, controls both keyframe and procedural animation; (4) *voice*, controls speech synthesis; (5) *gesture*, controls gesticulation expression.

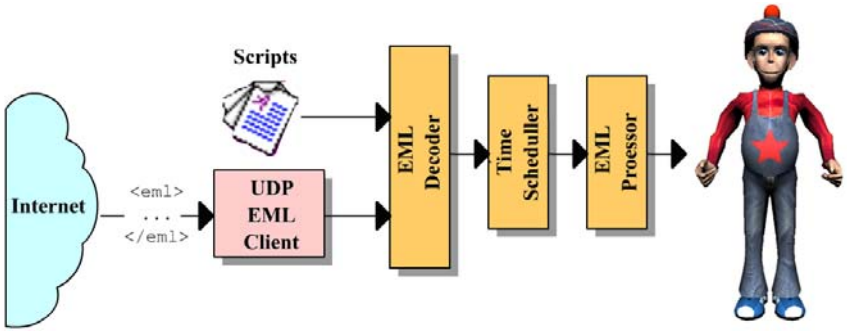

Fig. 6. EML integration with the virtual human

4.5 Evaluation

Two studies were conducted to assess the model's expressiveness. 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, aimed at evaluating several expression modalities while the second focused only on gesticulation.

4.5.1 First study

The first study was conducted in the scope of the "Papous" project at Inesc-ID[1] and aimed at comparing a human storyteller with a virtual storyteller with respect to story comprehension, emotion expression, credibility and subject satisfaction for each of gesticulation, facial and vocal expression. This document will focus only on the gesticulation expression results. The human storyteller was a non-professional actor which was simply asked to tell the story in an expressive way without imposing any requirements on gesticulation expression. Regarding the virtual storyteller, the voice consisted of modulated synthesized speech audio records. Facial expression, including proper lip-synch and emotion expression, was generated from a pseudo-muscular model [54]. Gesticulation expression was based on a GestuRA transcription of the human storyteller video, lasting 7 minutes and 30 seconds. In total, 286 gestures were transcribed of which 95% were automatically reproduced through feature-based gesticulation expression and 5% through keyframe animation.

Regarding structure, the subject begins by visualizing the story video and, then, answers to a questionnaire. Each subject is presented with one of four video versions: (1) *CRVR*, which uses the human narrator with real voice; (2) *CRVS*, which uses the human narrator with synthetic voice; (3) *CSVR*, which uses the virtual narrator with real voice; (4) *CSVS*, which uses the virtual narrator with synthetic voice. The questionnaire consists of twelve classification questions where the subject is asked to classify, from 1 (totally disagree) to 7 (totally agree), whether each modality helps understand the story, expresses emotions properly, is believable and is to his liking.

The study was presented to 108 students at IST-Technical University of Lisbon. Average age was 21 years and 89% were males. Most students were enrolled in technology-related courses. Each video version was presented to 27 students.

Gesticulation expression results are summarized in Table 1. Comparing real and synthetic gestures classifications, it is clear that real gestures got the best classification for every question. Nevertheless, synthetic gestures were positively classified by the majority of subjects with respect to contribution to story comprehension, emotion expression, believability and liking (positive answers above 50% for all questions).

From these results it is possible to conclude that synthetic gestures contribute to story comprehension, emotion expression and believability. Still, synthetic gestures do not capture all the subtleties of its real counterpart as these are better classified in general. Furthermore, this study had some limitations. Firstly, subjects were asked to evaluate gestures explicitly when it is known that gesture interpretation is essentially unconscious [1,2]. Secondly, subject to multiple interpretations, the notion of "believability" is hard to define thus, results related to the question "Gestures were believable" should be interpreted with caution.

[1] Papous: Papous project at Inesc-ID. Ref.: POSI / SRI / 41071 / 2001

Table 1. Summary of results for gesture classification questions in the first study

	CRVR	CSVR	CRVS	CSVS
Did gestures help understand the story?				
Negative (%)	7.4	22.2	11.1	14.8
Neutral (%)	7.4	3.7	7.4	7.4
Positive (%)	85.2	74.1	81.5	77.8
Did gestures express the story's emotions?				
Negative (%)	7.4	29.6	3.7	18.5
Neutral (%)	11.1	7.4	3.7	14.8
Positive (%)	81.5	63	92.6	66.7
Were gestures believable?				
Negative (%)	11.1	37	11.1	18.5
Neutral (%)	14.8	3.7	11.1	25.9
Positive (%)	74.1	59.3	77.8	55.6
Did you like the gestures?				
Negative (%)	11.1	29.6	7.4	11.1
Neutral (%)	22.2	11.1	3.7	22.2
Positive (%)	66.7	59.3	88.9	66.7

4.5.2 Second study

To further assess the model's expressiveness and to correct some of the flaws in the previous study, a second study was conducted. In this study, first, subjects are told that the evaluation is about virtual storytelling and "gesticulation expression" is never mentioned throughout. Second, synthetic gestures are indirectly evaluated through story interpretation questions. Third, each subject sees the story alternatively narrated by the human or virtual storyteller thus, allowing for direct comparison. Finally, as the study focuses on gesticulation expression, the real voice is used for both storytellers and three variations of the virtual storyteller are defined: (1) *ST*, which uses both feature-based and keyframe gesticulation; (2) *SF*, which uses only feature-based gesticulation; (3) *SN*, which uses no gesticulation.

The evaluation is structured into three parts. In part 1 – *profile* – the subject profile is assessed. In part 2 – *story interpretation* – the whole story is presented. To facilitate remembering, the story is divided into 8 segments of 30 seconds each. Segments are narrated by either the human storyteller or one of the three kinds of virtual storytellers randomly selected at the start. In concrete, the third and sixth segments are narrated by a subject selected storyteller, while the rest is arbitrarily narrated either by

the human or virtual storyteller provided that in the end each narrates an equal number of segments. After each segment, multiple choice interpretation questions are posed. In total 32 questions were formulated. Importantly, a subset, named the *highly bodily expressive (HBE)* questions, focuses on information specially marked in gestures, i.e., information which is either redundantly or non-redundantly conveyed through complex gestures like iconics or metaphorics. Finally, in part 3 – *story appreciation* – the subject is asked to choose the preferred storyteller and to describe the best and worst feature of each storyteller.

The study was presented to 39 subjects, 90% of which were male, with average age of 23 years and most had college-level education. The study was fully automated in software and average evaluation time was about 20 minutes. Distribution of virtual storyteller kinds across subjects was: 46% for ST; 31% for SF; 23% for SN. Subject recruitment included personal contact mainly at both campuses of IST-Technical University of Lisbon and distribution of the software through the Web.

Regarding story interpretation results, if we define *diff* to be the difference between the percentage of correct answers following the human storyteller and the percentage of correct answers following the virtual storyteller, then *diff* was: for ST, 4.69%; for SF, -0.68%; for SN, -1.62%. However, if we consider only HBE questions, than distribution is as follows: for ST, 4.75%; for SF, 0.00%; for SN, 9.19%. Regarding subject storyteller selection on the third and sixth segments, the human storyteller was selected about 75% of the time (for ST, 75.00%; for SF, 83.30%; for SN, 72.22%). Regarding subject storyteller preference, the human storyteller was preferred about 90% of the time (for ST, 88.89%; for SF, 83.33%; for SN, 100.00%). Finally, some of the worst aspects mentioned for the virtual storyteller were “body expression limited to arms”, “static/rigid”, “artificial” and “low expressivity”. These relate to the best aspects mentioned for the human storyteller, namely “varied postures”, “energetic/enthusiastic”, “natural” and “high expressivity”.

As can be seen by these results, the human storyteller fares better than the virtual storyteller. Interpretation with the human storyteller is better, though not that much (*diff* of 4.69% for ST). Furthermore, when given a choice, subjects almost always choose the human storyteller. Analyzing the best and worst aspects selected for each storyteller might give insight into this issue. Surprisingly, if all questions are considered, *diff* actually reduces for SN when compared to ST (-1.63% over 4.69%). The fact that the human storyteller’s voice and face were highly expressive and gestures were mostly redundant might help explain this. However, if only HBE questions are considered, *diff* considerably increases for the SN case (from 4.75% to 9.19%). Furthermore, for the SN case, the human storyteller was preferred 100% of the times. This confirms that gesticulation affects interpretation. Finally, comparing ST with SF, *diff* for all questions reduces for the latter case (from 4.69% to -0.68%). This suggests that the lack of feature-based gesticulation support for the small fraction of highly complex gestures does not impede effective interpretation.

5 Discussion and Future Work

This chapter overviews the challenge of building a virtual human computational model of gesticulation expression. First, a virtual human architecture is required with appropriate control mechanisms to support gesticulation animation. Second, the gesticulation execution problem, which refers to converting a gesticulation plan into an animation plan, must be addressed. Finally, the speech-gesticulation problem, which refers to converting communicative intent into verbal and gesticulation plans, should be addressed.

The chapter also proposes a gesticulation expression model which supports: (a) real-time gesticulation animation described as sequences of constraints on static (Portuguese Sign Language hand shapes, orientation palm axis, orientation angle and handedness) and dynamic features; (b) multimodal synchronization between gesticulation and speech; (c) automatic reproduction of GestuRA annotations; (d) expression control through the abstract integrated synchronized Expression Markup Language. The model builds on top of a layered virtual human architecture which supports keyframe and procedural animation. Finally, two studies were conducted to evaluate the model in a storytelling context. In both cases, we compare the expression of a human with a virtual storyteller. Results indicate that synthetic gestures contributed to story comprehension, emotion expression and believability. Furthermore, synthetic gestures fared well against real gestures. However, the fact that the human storyteller was consistently preferred hints that there is still room for improvement.

Therefore, regarding future work, first, gesticulation needs to go beyond arms and hands and explore other body parts. Posture shifts, which relate to discourse structure [55], could be explored. Second, some features' implementation restrict expressiveness. For instance, nothing guarantees that Portuguese Sign Language hand shapes and non-spline parametric curves (such as Bézier and Hermite) and combinations thereof suffice to express, respectively, all shapes and motion profiles. Furthermore, lack of elbow control in the upper limb manipulator limits naturalness [38]. Third, preparation and retraction motion, as well as co-articulation effects, could be automatically generated. Finally, a more anatomically correct hand model with appropriate constraints (subsection 3.1) would lead to more realistic gesticulation simulation.

At a more global level, the next step is to tackle the gesticulation production problem. Altogether, the model seems ready to support speech and gesticulation production models (subsection 2.4). Regarding de Ruiter's model, the gestuary can mostly be implemented through feature-based and keyframe gesticulation; signal passing synchronization is straightforwardly supported. Krauss' model which is feature-based is also compatible with the model but, cross-modal priming is not supported. The language effect on gesture in Kita and Özyürek's model occurs early in the production process and, ultimately, materializes into features which the model supports. McNeill's growth point model doesn't detail morphology generation. However, if the dialectic ultimately materializes into features and synchronization can be described with a finite number of synchronization points, then the model is likely to support it.

6 Acknowledgments

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

References

1. McNeill, D.: Hand and Mind: What gestures reveal about thought. University of Chicago Press (1992)
2. McNeill, D.: Gesture and Thought. University of Chicago Press (2005)
3. Kendon, A.: How gestures can become like words in F. Poyatos (ed.), Cross-cultural perspectives in nonverbal communication, pp.131-141, Hogrefe (1988)
4. Secretariado Nacional para a Reabilitação e Integração das Pessoas com Deficiência: Gestuário – Língua Gestual Portuguesa, fifth edition (1991)
5. Kendon, A.: Some relationships between body motion and speech in A. Siegman and B. Pope (eds.), Studies in dyadic communication. Pergamon Press (1972) 177-210
6. Kendon, A.: Gesticulation and speech: Two aspects of the process of utterance in M. Key (ed.), The Relationship of Verbal and Nonverbal Communication. Mouton and Co. (1980) 207-227
7. Nobe, S.: Where do most spontaneous representational gestures actually occur with respect to speech? in D. McNeill (ed.), Language and Gesture. Cambridge University Press (2000) 186-198
8. Kendon, A.: Sign languages of Aboriginal Australia: Cultural, semiotic and communicative perspectives. Cambridge University Press (1988)
9. Kita, S.: The temporal relationship between gesture and speech: A study of Japanese-English bilingual. MhD thesis, Department of Psychology, University of Chicago (1990)
10. Levelt, W.: Speaking. MIT Press (1989)
11. de Ruiter, J.: The production of gesture and speech in D. McNeill (ed.), Language and gesture, Cambridge University Press (2000) 284-311
12. Krauss, M., Chen, Y., Gottesman, R.: Lexical gestures and lexical access: A process model in D. McNeill (ed.), Language and gesture. Cambridge University Press (2000) 261-283
13. Kita, S., Özyürek, A.: What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking in Journal of Memory and Language (2003) 48:16-32
14. Akenine-Möller, T., Haines, E.: Real-time Rendering, second edition. A K Peters (2002)
15. Cavazza, M., Earnshaw, R., Magnenat-Thalmann, N., Thalmann, D.: Motion Control of Virtual Humans in IEEE Computer Graphics and Applications (1998) 18 (5):24-31
16. Thompson, D.; Buford, W.; Myers, L.; Giurintano, D.; Brewer III, J.: A Hand Biomechanics Workstation in Computer Graphics (1988) 22 (4): 335-343
17. Wagner, C.: The pianist's hand: Anthropometry and biomechanics in Ergonomics (1988) 31(1):97-131
18. Moccozet, L.; Magnenat-Thalmann N.: Dirichlet Free-Form Deformations and their Application to Hand Simulation in Proc. Computer Animation'97 (1997) 93-102
19. Sibille, L.; Teschner, M.; Srivastava, S.; Latombe, J.: Interactive Simulation of the Human Hand in CARS'02 (2002) 7-12
20. Albrecht, I.; Haber, Jörg, H.; Siedel, H.: Construction and Animation of Anatomically Based Human Hand Models in SIGGRAPH'03 (2003) 98-109

21. 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 Proc. of SIGGRAPH'94 (1994) 413-420
22. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., Yan, H.: Embodiment in Conversational Interfaces: Rea in Proc. of the CHI'99 Conference, Pittsburgh, PA (1999) 520-527
23. Cassell, J., Stone, M.: Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems in Proc. of the AAAI 1999 Fall Symposium on Psychological Models of Communication in Collaborative Systems, North Falmouth, MA (1999) 34-42
24. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: the Behavior Expression Animation Toolkit in Proc. of SIGGRAPH'01 (2001) 477-486
25. Kopp, S., Wachsmuth, I.: A knowledge-based approach for lifelike gesture animation in Proc. of the 14th European Conf. on Artificial Intelligence, Amsterdam, IOS Press (2000)
26. Wachsmuth, I., Kopp, S.: Lifelike Gesture Synthesis and Timing for Conversational Agents in Wachsmuth, Sowa (eds.), Gesture and Sign Language in Human-Computer Interaction, International Gesture Workshop (GW 2001). Springer-Verlag, (2002) 120-133
27. SABLE: A Synthesis Markup Language (v. 1.0). www.bell-labs.com/project/tts/sable.html
28. Kopp, S., Tepper, P., Cassell, J.: Towards Integrated Microplanning of Language and Iconic Gesture for Multimodal Output in Proc. of the International Conference on Multimodal Interfaces (ICMI'04). ACM Press (2004) 97-104
29. Arafa, Y.; Kamyab, K.; Mamdani, E; Character Animation Scripting Languages: A Comparison in Proc. of the International Conference on Autonomous Agents 2003, Melbourne, Australia (2003)
30. Kopp, S.; Krenn, B.; Marsella, S.; Marshall, A.; Pelachaud, C.; Pirker, H.; Thórisson, K.; Vilhjálmsson, H.; Towards a Common Framework for Multimodal Generation: The Behavior Markup Language in Proc. of Intelligent Virtual Agents'06 (2006) 205-217
31. VHML: VHML – Virtual Human Markup Language. www.vhml.org/
32. SMIL: SMIL - Synchronized Multimedia. www.w3.org/AudioVideo/
33. Kranstedt, A.; Kopp, S.; Wachsmuth, I.: MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents in AAMAS'02 Workshop Embodied conversational agents-let's specify and evaluate them!, Bologna, Italy (2002)
34. Ruttkay, Z.; Noot, H.: Variations in Gesturing and Speech by GESTYLE in International Journal of Human-Computer Studies, Special Issue on 'Subtle Expressivity for Characters and Robots' (2005) 62(2): 211-229
35. de Carolis, B.; Pelachaud, C.; Poggi, I.; Steedman, M.: APML, a Mark-up Language for Believable Behavior Generation in H. Prendinger (ed.), Life-like Characters. Tools, Affective Functions and Applications. Springer (2004)
36. Blumberg, B., Galyean, T.: Multi-Level Direction of Autonomous Creatures for Real-Time Virtual Environments in Proc. of SIGGRAPH'95 (1995) 30(3):47-54
37. Perlin, K., Goldberg, A.: Improv: A System for Scripting Interactive Actors in Virtual Worlds in Proc. of SIGGRAPH'96 (1996) 205-216
38. Tolani, D.; Goswani, A.; Badler, N.: Real-time inverse kinematics techniques for anthropomorphic limbs in Graphics Models (2000) 62:353-338
39. NASA Man-Systems Integration Manual (NASA-STD-3000)
40. Festival: The Festival Speech Synthesis Systems. www.cstr.ed.ac.uk/projects/festival/
41. Mark L.: Control of Human Movement. Human Kinetics Publishers (1993)
42. Unuma, M.; Anjou, K.; Takeuchi, R.; Fourier principles for emotion-based human figure animation in Proc. of SIGGRAPH'95 (1995) 91-96

43. Brudelin, A.; Williams, L.: Motion signal processing in Proc. of SIGGRAPH'95 (1995) 97-104
44. Amaya, K.; Bruderlin, A.; Calvert, T.: Emotion from motion in Proc. Graphics Interface'96 (1996) 222-229
45. Rose, C.; Bodenheimer, B.; Cohen, M.: Verbs and Adverbs: Multidimensional Motion Interpolation in IEEE Computer Graphics and Applications (1998) 18 (5):32-40
46. Chi, D.; Costa, M.; Zhao, L.; Badler, N.: The EMOTE model for effort and shape in Proc. of SIGGRAPH'00 (2000) 173-182
47. Hartmann, B.; Mancini, M.; Pelachaud, C.: Implementing Expressive Gesture Synthesis for Embodied Conversational Agents in Gesture Workshop. Springer (2005)
48. de Melo, C., Paiva, A.: Environment Expression: Expressing Emotions through Cameras, Lights and Music in Proc. of Affective Computing Intelligent Interaction'05 (2005) 715-722
49. de Melo, C., Paiva, A.: Expression of Emotions in Virtual Humans using Lights, Shadows, Composition and Filters in Proc. of Affective Computing Intelligent Interaction'07 (2007) 546-557
50. Gut, U.; Looks, K.; Thies, A.; Trippel, T.; Gibbon, D.: CoGest – Conversational Gesture Transcription System, Technical Report. University of Bielefeld (1993)
51. Kipp, M.: ANVIL – A Generic Annotation Tool for Multimodal Dialogue in Proc. of the 7th European Conference on Speech Communication and Technology (2001) 1367-1370
52. Pavlovic, V., Sharma, R., Huang, T.: Visual Interpretation of hand gestures for human computer interaction: A review in IEEE Trans. Pattern Analysis Machine Intelligence (1997) 19:677-695
53. Gavrila, D.: The visual analysis of human movement: A survey in Computer Vision and Image Understanding (1999) 73:82-98
54. Raimundo, G.: Real-time Facial Expression and Animation. MSc thesis, Department of Information Systems and Computer Engineering, IST-Technical University of Lisbon (2007)
55. Cassell, J.; Nakano, Y.; Bickmore, T.; Sidner, C.; Rich, C.: Annotating and Generating Posture from Discourse Structure in Embodied Conversational Agents in Workshop on Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents, Autonomous Agents'01 (2001)