A Story about Gesticulation Expression

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. Gesticulation is essential for the storytelling experience thus, virtual storytellers should be endowed with gesticulation expression. This work proposes a gesticulation expression model based on psycholinguistics. 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 (motion profiles) features; (b) multimodal synchronization between gesticulation and speech; (c) automatic reproduction of annotated gesticulation according to GestuRA, a gesture transcription algorithm. To evaluate the model two studies, involving 147 subjects, were conducted. 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. Results indicate that synthetic gestures fared well when compared to real gestures however, subjects preferred the human storyteller.

1 Introduction

Gesticulation is essential for the storytelling experience. Gesticulation is the kind of gestures humans do in a conversation or narration context [1]. These are idiosyncratic, unconventional and unconscious gestures which reveal the imagery of the story and, thus, support suspension of disbelief. As virtual storytelling systems harness the benefits of traditional storytelling, it is important to endow virtual storytellers with comprehensive models, inspired in humans, for gesticulation expression.

This work proposes a gesticulation expression model which supports:

- Real-time gesticulation animation described as sequences of constraints on static (Portuguese Sign Language hand shapes, orientations and positions) and dynamic (motion profiles) features;
- Multimodal synchronization between gesticulation and speech;
- Automatic reproduction of annotated gesticulation according to GestuRA, a gesture transcription algorithm.

This paper is organized as follows. Section 2 describes relevant research on gesticulation. Section 3 describes the gesticulation expression model. Section 4 describes two studies conducted to evaluate the proposed model in storytelling contexts. Finally, section 5 draws some conclusions and discusses future work.

2 Background and Related Work

Gesticulation is the kind of gestures humans do in a conversation or narration context [1]. They tend to focus on arms and hands, though other body parts may be involved [2]. 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 ([3,4] in [1]): preparation; pre-stroke hold; stroke; post-stroke hold; retraction. The stroke is where actual meaning is conferred and is synchronous with its co-expressive speech 90% of the time [5]. Thus, the proposed gesticulation expression model focuses on arms and hands and supports sub-second gesticulation phase synchronization with speech.

The proposed model is feature-based, i.e., gesticulation is modeled as sequences of static (hand shape, orientation and position) and dynamic (motion profiles) constraints. A feature-based approach is appropriate for several reasons. First, according to McNeill [2] it makes more sense to describe gesticulation according to dimensions and saliency rather than categories and hierarchy. This 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 [1,2] ultimately materializes into gesture features; de Ruiter's sketch model [6] revolves around the concept of gesture templates (in a gestuary) which correspond to constraints on features; Krauss [7] actually considers knowledge representation as feature-based; finally, Kita & Özyürek [8] even though not detailing gesture morphology, motivate their models with motion gestures described according to features.

Regarding related work, several computational psycholinguistics systems have been proposed. Animated Conversation [9], 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)* [10,11] presents an embodied conversational agent capable of proper distribution and realization of communicative intent across speech and gesture. In [12] Cassell et al 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 [13,14] 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 [6], 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 [15]. Recently, Cassell, Kopp and colleagues brought together the best from the aforementioned systems in *NUMACK* [16], 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 and speech production process is beyond the scope of this work, the underlying gesticulation animation model in these systems shares several aspects with the proposed

model, namely: its requisites are strictly based on psycholinguistics research and similar static and dynamic features are explored.

The problem of controlling and integrating gesticulation expression with other modalities is usually solved through markup languages [17]. This work also proposes a control language – *Expression Markup Language (EML)* – which is particularly influenced by: VHML [18], SMIL [19] and MURML [20]. From *Virtual Human Markup Language (VHML)* this work uses the notion of dividing control into subsystems. From *Synchronized Multimedia Integration Language (SMIL)*, which is oriented for audiovisual interactive presentations, this work benefits from the sophisticated modality synchronization mechanism. From *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 [21], which tries to capture an individual's expression style, and APML [22], which represents, among others, communicative intent, emotions, interaction and cultural aspects, the proposed language focuses on low-level body control such as gesticulation animation as sequences of constraints on static and dynamic features and the generation of speech in a text-to-speech system.

Finally, this work supports automatic reproduction from a gesture transcription algorithm. Usually, these algorithms are used to learn aspects from human gesticulation expression and, then, generate databases or explicit rules for virtual humans. However, the added value of being able to automatically reproduce such annotations is flexibility. This idea relates to efforts in automatic gesture recognition [23,24]. Such systems accurately recognize form but, still lag with respect to meaning. In contrast, gesture transcription algorithms rely on knowledge from (human) analysts to interpret meaning and, thus, reproduction from the final annotation, though less accurate in form, is more flexible.

3 The Model

The gesticulation model fits into a broad virtual human real-time multimodal expression model which includes deterministic, non-deterministic, gesticulation, facial, vocal and environment expression [25]. This paper will focus on the first three. The model also supports automatic reproduction of gesticulation annotations according to GestuRA, a gesture transcription algorithm.

The virtual human is structured according to a three-layer architecture [26,27]. The *geometry layer* defines a 54-bone human-based skeleton. The *animation layer* defines deterministic and non-deterministic animation mechanisms. The *behavior layer* defines gesticulation expression and supports a language for integrated synchronized multimodal expression.

3.1 Deterministic Expression

Deterministic expression is about deterministic animation, i.e., sequences of keyframes usually exhaustively conceived by human artists. 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.

3.2 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 [28]. In the animation layer, three inverse kinematics and one inverse velocity primitives are defined, namely: (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 Jacobian-based inverse velocity algorithms to animate the target according to intended Cartesian and angular velocities.

3.3 Gesticulation Expression

The gesticulation expression model controls arms and hands and is built on top of deterministic expression and non-deterministic expression. In concrete, 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 3.3.1. The model supports multimodal synchronization, in particular, between speech and gesture. Synchronization is described on subsection 3.3.2. 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 3.3.3.

3.3.1 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 keyframe interpolation motion profiles.

Regarding static features, the *hand shape* feature can assume any Portuguese Sign Language hand shape [32]. 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 several kinds of (keyframe) interpolators, namely: *linear*, which defines linear interpolation; *cosine*, which defines cosine interpolation; and *parametric cubic curves*, which can represent any kind of velocity profile. Furthermore, interpolators can be structured into hierarchies thus, leading to sophisticated motion profiles. Furthermore, either Cartesian or joint angle velocity can be used. Currently, deceleration near the target position and overshooting effects have been simulated using Bézier and Hermite cubic curves.

3.3.2 Synchronization

To support sub-second synchronization of gesture phases, a control markup language – Expression Markup Language (EML) – supporting phoneme-level synchronization is proposed. The language integrates with SABLE [15] and thus, supports synchronization with speech properties such as intonation contour. Similarly to SMIL [33], 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 to a named timestamp: (a) start of a word; (b) end of a word; (c) start of a phoneme. EML is further described on subsection 3.4.

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 the 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.

In the current implementation, the Festival [29] text-to-speech system has been used to generate speech, retrieve phoneme information and render SABLE text.

3.3.3 Automatic Reproduction of Gesticulation Annotations

The gesticulation model supports automatic reproduction of *Gesture Recording Algorithm (GestuRA)* annotations. This constitutes an important evaluation tool. As speech

and gesture production from communicative intent is not simulated, an alternative to evaluating the model is to compare it to real life situations.

GestuRA, based on [2] and [30], is a linguistically motivated iterative algorithm for gesticulation form and meaning transcription. It 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 – narrative, metanarrative and paranarrative [1]. Fourth, gesticulation is filtered ignoring remaining gestures (such as adaptors, emblems, signs). Fifth, gesticulation phases are annotated. Sixth, gesticulation form is formally annotated. Finally, seventh, gesticulation is classified according to its dimensions and its meaning analyzed. GestuRA integration with the model is achieved through *Anvil* [31], a generic multimodal annotation tool, which exports annotations to a XML format which is, then, converted into EML for immediate execution in virtual humans - Fig. 1.

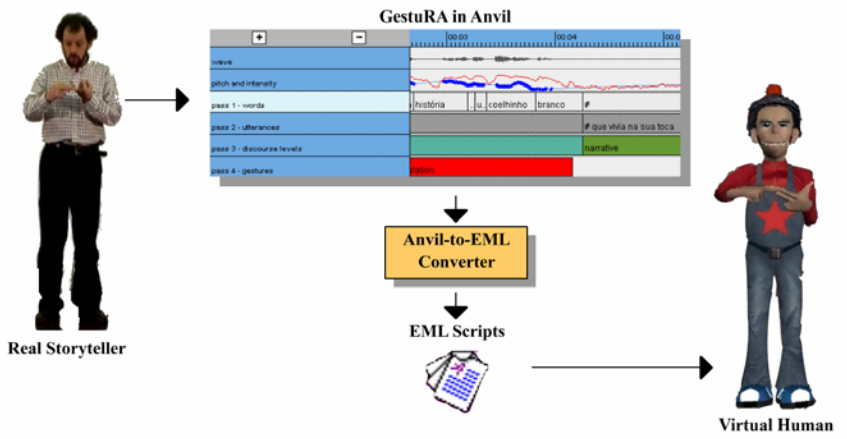

Fig. 1. GestuRA integration with the model

3.4 Multimodal Expression

This work proposes a markup, integrated and synchronized language – Expression Markup Language (EML) – which serves as a control interface for virtual human bodies. The language can be used in two ways, Fig. 2: (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. 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. 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) sup-

ports 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 [33]; (3) *body*, controls both deterministic and non-deterministic body expression; (4) *gesture*, controls gesticulation expression.

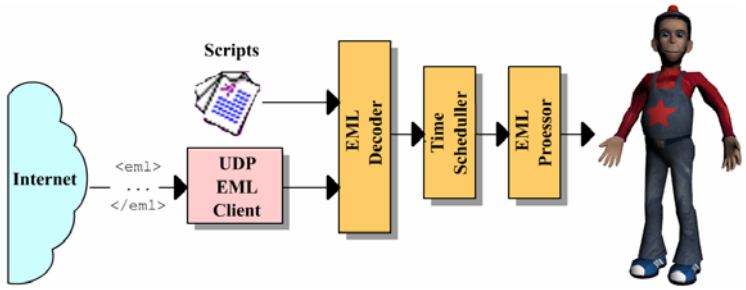

Fig. 2. EML integration with the model

4 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 all forms of expression while the second focused only on gesticulation.

4.1 First Study

The first study was conducted in the scope of the Papous project at Inesc-ID. This project compares a human storyteller with a virtual storyteller with respect to story comprehension, emotion expression, believability and subject satisfaction for each of body, facial and vocal expression. This paper focuses on body 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 synthesized speech audio records. Facial expression was based on a muscular model capable of proper lip-synch and emotion expression. Body expression relied 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 and 5% through keyframe deterministic animation.

Regarding structure, first the subject visualized the story video and, then, answered to a questionnaire. Each subject was presented one of four video versions: (1) *CRVR* – Human narrator with real voice; (2) *CRVS* – Human narrator with synthetic voice;

(3) *CSVR* – Virtual narrator with real voice; (4) *CSVS* – Virtual narrator with synthetic voice. The questionnaire had twelve questions where the subject classified, from 1 (totally disagree) to 7 (totally agree), whether each modality help understand the story, express emotions properly, is believable and is to his liking.

The study was presented to 108 students at the Technical University of Lisbon. Average age was 21 years and 89% of which were males. Most students related to computer science courses. Each video version was presented to 27 students.

Body expression results are summarized in Table 1. In general synthetic gestures are classified lower than real gestures. However, classification differs only in about 0.45 points. Finally, notice that real gesture classification (about 5) was well below 7.

Table 1. Body expression average classifications (scale goes from 1 to 7)

	CRVR	CSVR	CRVS	CSVS
Gestures helped to understand the story	5.19	4.91	5.04	4.82
Gestures expressed the story's emotions	5.15	4.76	5.30	4.82
Gestures were believable	5.07	4.30	5.30	4.61
I liked the gestures	4.89	4.49	5.22	4.82

From these results it is possible to conclude that synthetic gestures fared well when compared to real gestures. Furthermore, in absolute terms, a classification of about 4.6 is reasonably good. However, this study had some limitations. Firstly, subjects were asked to evaluate gestures explicitly when it is known that gesture interpretation is essentially unconscious [1]. Secondly, subject to multiple interpretations, the notion of "believability" is hard to define thus, results related to the question "Gestures were believable" must be interpreted with caution.

4.2 Second Study

So as to further assess the gesticulation model's expressiveness and to correct some of the flaws in the previous study, a second study was conducted. Here, 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 storyteller comparison. Finally, as the study focused on gesticulation, the real voice is used for both storytellers and three variations of the virtual storyteller are defined: (1) *ST*, where feature-based and keyframe gesticulation are expressed; (2) *SF*, where only feature-based gesticulation is expressed; (3) *SN*, where no gesticulation is expressed.

The evaluation is structured into three parts. In part 1 – *profile* – the subject profile is assessed. In part 2 – *story interpretation* – the story is presented to the subject in 8 segments. Segments are narrated either by the human or one, randomly selected at the

start, of the three kinds of virtual storytellers. In concrete, the third and sixth segments are narrated by a storyteller selected by the subject, while the rest is arbitrarily narrated either by the human or virtual storyteller provided that in the end each gets 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, focused on information specially marked in gestures, i.e., information which was 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 which is the best and worst feature of each storyteller.

The study was presented to 39 subjects, 90% of which were male, average age was 23 years and mostly had higher education. The study was fully automated 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 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 the results, the human storyteller is better than the virtual storyteller. Interpretation with the human storyteller is better, but not that much (*diff* of 4.69% for ST). Furthermore, when given a choice, subjects almost always chose 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 Conclusions and Future Work

This paper proposed a model for a feature-based real-time gesticulation animation model. Static features include Portuguese Sign Language hand shapes, position, orientation palm axis, orientation angle, and handedness. Dynamic features include motion profiles. For phoneme-level speech-gesture synchronization, a multimodal expression language, which integrates with SABLE, is proposed. Moreover, the model supports automatic reproduction of annotated gesticulation according to GestuRA. Finally, results from two studies indicate that the model's gesticulation expression fares well when compared to real gesticulation in a storytelling context. Still, the human storyteller was consistently preferred to the virtual storyteller hinting that the model can be improved.

Altogether the model seems to be ready to support gesticulation production models thus, moving from automatic reproduction to automatic generation. Regarding de Ruiter's model [6], the gestuary can mostly be implemented through feature-based and keyframe gesticulation and signal passing synchronization is straightforwardly supported. Krauss' model [7] which is feature-based is also compatible. The language effect on gesture in Kita and Özyürek's model [8] occurs early on the production process and, ultimately, materializes into specific features which this model supports. McNeill's growth point model [1,2] lacks details on morphology generation however, if the dialectic materializes into features and synchronization can be described with respect to a finite number of specific synchronization points, then this model may support it.

Regarding future work, first, gesticulation needs to go beyond arms and hands and explore other body parts. Second, some features' implementation restrict expressiveness. Nothing guarantees that Portuguese Sign Language hand shapes and combinations thereof suffice to model all relevant shapes. Furthermore, lack of redundancy, or elbow control, in the upper limb manipulator limits naturalness. In this sense, seven degrees-of-freedom manipulators should be explored. Third, preparation and retraction motion and co-articulation effects could be automatically generated. Finally, a more anatomically correct hand model with appropriate constraints ([34,35]) would lead to more realistic gesticulation simulation.

6 Acknowledgments

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

7 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.: Sign languages of Aboriginal Australia: Cultural, semiotic and communicative perspectives. Cambridge University Press (1988)
4. Kita, S.: The temporal relationship between gesture and speech: A study of Japanese-English bilingual. MhD, Department of Psychology, University of Chicago (1990)
5. 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
6. de Ruiter, J.: The production of gesture and speech in D. McNeill (ed.), Language and gesture, Cambridge University Press (2000) 284-311
7. 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
8. 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 48 (2003) 16-32
9. 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
10. 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
11. 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
12. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: the Behavior Expression Animation Toolkit in Proc. of SIGGRAPH'01 (2001) 477-486
13. 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)
14. Wachsmuth, I., Kopp, S.: Lifelike Gesture Synthesis and Timing for Conversational Agents in Wachsmusth, Sowa (eds.), Gesture and Sign Language in Human-Computer Interaction, International Gesture Workshop (GW 2001). Springer-Verlag, (2002) 120-133
15. SABLE: A Synthesis Markup Language (v. 1.0). www.bell-labs.com/project/tts/sable.html
16. 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
17. Arafa, Y., Kamyab, K., Mamdani, E.: Character Animation Scripting Languages: A Comparison in Proc. of the 2nd Intl. Conference of Autonomous Agents and Multiagent Systems (2003) 920-921
18. VHML: VHML – Virtual Human Markup Language. www.vhml.org/
19. SMIL: SMIL - Synchronized Multimedia. www.w3.org/AudioVideo/
20. 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)
21. 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', 62(2), (2005) 211-229
22. 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)

23. Pavlovic, V., Sharma, R., Huang, T.: Visual Interpretation of hand gestures for human computer interaction: A review in IEEE Trans. Pattern Analysis Machine Intelligence, vol.19, July (1997) 677-695
24. Gavrila, D.: The visual analysis of human movement: A survey in Computer Vision and Image Understanding, vol.73, Jan. (1999) 82-98
25. de Melo, C., Paiva, A.: Multimodal Expression in Virtual Humans. Accepted for Computer Animation & Social Agents 2006 (CASA2006) and Journal of Computer Animation and Virtual Worlds (2006)
26. Blumberg, B., Galyean, T.: Multi-Level Direction of Autonomous Creatures for Real-Time Virtual Environments in Proc. of SIGGRAPH '95, 30(3) (1995) 47-54
27. Perlin, K., Goldberg, A.: Improv: A System for Scripting Interactive Actors in Virtual Worlds in Proc. of SIGGRAPH'96 (1996) 205-216
28. NASA Man-Systems Integration Manual (NASA-STD-3000)
29. Black, A.; Taylor, P.; Caley, R.; Clark, R.: Festival. www.cstr.ed.ac.uk/projects/festival/
30. Gut, U., Looks, K., Thies, A., Trippel, T., Gibbon, D.: CoGesT – Conversational Gesture Transcription System. Technical Report, University of Bielefeld (1993)
31. Kipp, M.: ANVIL – A Generic Annotation Tool for Multimodal Dialogue in Proc. of the 7th European Conference on Speech Comm. and Technology, Aalborg, (2001) 1367-1370
32. Secretariado Nacional para a Reabilitação e Integração das Pessoas com Deficiência. Gestuário – Língua Gestual Portuguesa – 5th edition
33. SMIL. "SMIL: Synchronized Multimedia"; www.w3.org/AudioVideo/
34. Thompson, D., Buford, W., Myers, L., Giurintano, D., Brewer III, J.: A Hand Biomechanics Workstation in Computer Graphics, vol.22, no.4 (1988) 335-343
35. Albrecht, I., Haber, J., Siedel, H.: Construction and Animation of Anatomically Based Human Hand Models in SIGGRAPH 2003 (2003) 98-109