Hand Gesture Synthesis for Conversational Characters

Michael Neff

Key words: gesture, character animation, nonverbal communication, virtual agents, embodied conversational agents

Abstract This chapter focuses on the generation of animated *gesticulations*, co-verbal gestures that are designed to accompany speech. It begins with a survey of research on human gesture, discussing the various forms of gesture, their structure and timing requirements relative to speech. The two main problems for synthesizing gesture animation are determining what gestures a character should perform (the specification problem) and then generating appropriate motion (the animation problem). The specification problem has used a range of input, including speech prosody, spoken text and a communicative intent. Both rule-based and statistical approaches are employed to determine gestures. Animation has also used a range of procedural, physics-based and data driven approaches in order to solve a significant set of expressive and coordination requirements. Fluid gesture animation must also reflect the context and include listener behavior and floor management. The chapter concludes with a discussion of future challenges.

1 Introduction

Do gestures communicate? Yes, they do. This has been the conclusion of several meta-studies on the impact of gesture (8; 18; 13). It is also one of the distinguishing features of gestures in animation. While all movement communicates to some degree, gestures often play a role that is explicitly communicative. Another distinguishing feature for the gestures that we are most often interested in is that they are *co-verbal*. That is, they occur with speech and they are inextricably linked to that

Michael Neff
University of California - Davis, Dept. of Computer Science, 1 Shields Ave. Davis, CA, 95616, USA e-mail: mpneff@ucdavis.edu

speech in both content and timing. McNeill argues that gestures and language are not separate, but gestures are part of language (37).

There are different forms of movement that can broadly be called “gesture”. Building on the categories of Kendon (17), McNeill defined “Kendon’s Continumm” (36; 37) to capture the range of gesture types people employ:

- *Gesticulation*: gesture that conveys a meaning related to the accompanying speech
- *Speech-like gestures*: gestures that take the place of a word(s) in a sentence
- *Emblems*: conventionalized signs, like a thumbs up
- *Pantomime*: gestures with a story and are produced without speech
- *Sign Language*: signs are lexical words

As you move along the continuum, the degree to which speech is obligatory decreases and the degree to which gestures themselves have the properties of a language increases. This article will focus on *gesticulations*, which are gestures that co-occur with speech as they are most relevant to conversational characters. Synthesis of the whole spectrum; however, presents worthwhile animation problems. Emblems and pantomimes are useful in situations where speech may not be possible. Sign languages are the native language of many members of the deaf community and sign synthesis can increase their access to computational sources. The problems of gesticulations are unique, however, since they are co-present with speech and do not have linguistic structure on their own.

Kendon introduced a three level hierarchy to describe the structure of gestures (16). The largest structure is the *gesture unit*. Gesture units start in a retraction or rest pose, continue with a series of gestures and then return to a rest pose, potentially different from the initial rest pose. A *gesture phrase* encapsulates an individual gesture in this sequence. Each gesture phrase can in turn be broken down into a sequence of *gesture phases*. A *preparation* is a motion that takes the hands to the required position and orientation for the start of the gesture stroke. A *prestroke hold* is a period of time in which the hands are held in this configuration. The *stroke* is the main meaning carrying movement of the gesture and has the most focused energy. It may be followed by a *poststroke hold* in which the hands are held at the end position. The final phase is a *retraction* that returns the hands to a rest pose. All phases are optional except the stroke. There are some gestures in which the stroke does not involve any movement (e.g. a raised index finger). These are variously called an *independent hold* (22) or a *stroke hold* (37). The pre- and poststroke holds were proposed by Kita (21) and act to synchronize the gesture with speech. The prestroke hold delays the gesture stroke until the corresponding speech begins and the poststroke hold occurs while the corresponding speech is completing. Much like they allow mental processing in humans, they can be used in synthesis systems to allow time for planning or other processing to take place.

The existence of gesture units is important for animation systems as it indicates a potential need to avoid generating a sequence of singleton gestures that return to a rest pose after each gesture. While this would offer the simplest synthesis solution, people are quite sensitive to the structure of gestural communication. A study (20) showed that people found a character that used multiple phrase gesture units more natural, friendly and trustworthy than a character that performed singleton gestures, which was viewed as more nervous. These significant differences in appraisal occurred despite only one of twenty five subjects being able to actually identify the difference between the multi-phrase g-unit clips and single phrase g-unit clips. This illustrates what appears to be a common occurrence in our gesture research: people will react to differences in gesture performance without being consciously aware of what those differences are.

Gestures are synchronized in time with their co-expressive speech . About 90% of the time, the gesture occurs slightly before the co-expressive speech (43) and rarely occurs after (16). Research on animated characters does indicate a preference for this slightly earlier timing of gesture, but also suggests that people may not be particularly sensitive to errors in timing, at least within a +/- .6 second range (51).

A number of categorizations of gesture have been proposed. One of the best known is from McNeill and Levy (38; 36) and contains the classes *iconics*, *metaphorics*, *deictics* and *beats*. Iconic gestures create images of concrete objects or actions, such as illustrating the size of a box. Metaphorics create images of the abstract. For instance, a metaphoric gesture could make a cup shape with the hand, but refer to holding an idea rather than an actual object. Metaphoric gestures are also used to locate ideas spatially, for instance putting positive things on the left and negative to the right and then using this space to categorize future entities in the conversation. Deictics locate objects and entities in space, as with pointing, creating a reference and context for the conversation. They are often performed with a hand that is closed except for an extended index finger, but can be performed with a wide range of body parts. Deixis can be abstract or concrete. Concrete deixis points to an existing reference (e.g. an object or person) in space whereas abstract deixis creates a reference point in space for an idea or concept. Beats are small back and forth or up and down movements of the hand, performed in rhythm to the speech. They serve to emphasize important sections of the speech.

In later work, McNeill (37) argued that it is inappropriate to think of gesture in terms of categories, but the categories should instead be considered dimensions. This reflects the fact that any individual gesture may contain several of these properties (e.g. deixis and iconicity). He suggests additional dimensions of *temporal highlighting* (the function of beats) and *social interactivity*, which helps to manage turn taking and the flow of conversation.

2 State of the art

Generation of conversational characters has achieved substantial progress, but the bar for success is extremely high. People are keen observers of human motion and will make judgments based on subtle details. By way of analogy, people will make judgments between good and bad actors, and actors being good in a particular role, but not another - and actors are human, with all the capacity for naturalness and expressivity that comes with that. The bar for conversational characters is that of a good actor, effectively performing a particular role. The field remains a long way from being able to do this automatically, for a range of different characters and over prolonged interactions with multiple subjects.

3 Gesture Generation Tasks

3.1 Gesture Specification

When generating virtual conversational characters, one of the primary challenges is determining what gestures a character should perform. Different approaches have trade-offs in terms of the type of input information they require, the amount of processing time needed to determine a gesture and the quality of the gesture selection, both on grounds of accurately reflecting a particular character personality and being appropriate for the co-expressed utterance.

One approach is to generate gestures based on prosody variations in the spoken audio signal. Prosody includes changes in volume and pitch. Such approaches have been applied for head nods and movement (39), as well as gesture generation (33; 32). A main advantage of the approach is that good quality audio can be highly expressive and using it as an input for gesture specification allows the gestures to match the expressive style of the audio. Points of emphasis in the audio appear to be good landmarks for placing gesture and their use will provide uniform emphasis across the channels. Prosody-based approaches have been used to generate gesture in real-time as a user speaks (33; 32). The drawback of only using prosody is that it does not capture semantics, so the gestures will likely not match the meaning of the audio and certainly not supplement the underlying meaning that is being conveyed in the utterance with information not present in the audio. This concern can be at least partially addressed by also parsing the spoken text (35). It is believed that in human communication, the brain is co-planning the gesture and the utterance (37), so approaches that do not use future information about the planned utterance may be unlikely to match the sophistication of human gesture-speech coordination.

Another approach generates gesture based on the text of the dialogue that is to be spoken. A chief benefit of these techniques is that text captures much of the information being conveyed, so these techniques can generate gestures that aid the semantics of the utterance. Text can also be analyzed for emotional content and rhetorical style, providing a rich basis for gesture generation. Rule-based approaches (4; 31; 34; 35) can determine both the gesture locations and the type of gestures to be performed. Advantages of these techniques are that they can handle any text covered by their knowledge bases and are extensible in flexible and straightforward ways. Disadvantages include that some amount of manual work is normally required to create the rules and it is difficult to know how to author the rules to create a particular character, so behavior tends to be generic. Other work uses statistical approaches to predict the gestures that a particular person would employ (19; 42; 3). These techniques support the creation of individualized characters, which are essential for many applications, such as anything involving storytelling. Individualized behavior may also outperform averaged behavior (3), as would be contained in generic rules. These approaches, however, are largely limited to reproducing characters like the subjects modeled and creating arbitrary characters remains an open challenge. Recent work has begun applying deep learning to the mapping from text and prosody to gesture (6). This is a potentially powerful approach, but requires a large quantity of data and ways to produce specific characters must be developed. While the divide between prosody-driven and rule-based approaches is useful for understanding techniques, current approaches are increasingly relying on a combination of text and prosody information (e.g. (34; 35)).

Techniques based on generating gesture from text are limited to ideas expressed in the text. The information we convey through gesture is sometimes redundant with speech, although expressed in a different form, but often expresses information that is different to that in speech (37). For example, I might say "I saw a [monster.]", with the square brackets indicating the location of a gesture that holds my hand above my head, with my fingers bent 90 degrees at the first knuckle and then held straight. The gesture indicates the height of the monster, information completely lacking from the verbal utterance. Evidence suggests that gestures are most effective when they are non-redundant (13; 9; 46). This implies the need to base gesture generation on a deeper notion of a "communicative intent, which may not solely be contained in the text and describes the fully message to be delivered.

The SAIBA (Situation, Agent, Intention, Behavior, Animation) framework represents a step towards establishing a computational architecture to tackle the fundamental multimodal communication problem of moving from a communicative intent to output across the various agent channels of gesture, text, prosody, facial expressions and posture (44). The approach defines stages in production and markup languages to connect them. The first stage is planning the communicative intent. This is communicated using the Function Markup Language (12) to the Behavior Planner, which decides how to achieve the desired functions using the agent modalities available. The final behavior is then sent to a Behavior Realizer for generation using the Behavior Markup Language (25; 50). Such approaches echo, at least at the broad conceptual level, theories of communication like McNeill's Growth Point hypothesis that argue gesture and language emerge in a shared process from a communicative intent (37). Recent work has sought to develop cognitive (24) and combined cognitive and linguistic models (2) to explore the distribution of communicative content across output modalities.

3.2 Gesture Animation

Generate high quality gesture animation must satisfy a rich set of requirements:

- match the gesture timing to that of the speech

- connect individual gestures into fluent gesture units
- adjust the gesture to the character's context (e.g. to point to a person or object in the scene)
- generate appropriate gesture forms for the utterance (e.g. show the shape of an object, mime an action being performed, point)
- vary the gesture based on the personality of the character
- vary the gesture to reflect the character's current mood and tone of the speech

While a wide set of techniques have been used for gesture animation, the need for precise agent control, especially in interactive systems, has often favored the use of kinematic procedural techniques (e.g. (5; 27; 10)). For example, Kopp and Wachsmuth (27) present a system that uses curves derived from neurophysiological research to drive the trajectory of gesturing arm motions. Procedural techniques allow full control of the motion, making it easy to adjust the gesture to the requirements of the speech, both for matching spatial and timing demands.

While gesture is less constrained by physics than motions like tumbling, physical simulation has still been used for gesture animation and can add important nuance to the motion (40; 41; 42; 49). These approaches generally include balance control and a basic approximation to muscle, such as a proportional derivative controller. The balance control will add full body movement to compensate for arm movements and the controllers can add subtle oscillations and arm swings. These effects require proper tuning.

Motion capture data has seen increasing use in an attempt to improve the realism of character motion. These techniques often employ versions of motion graphs (28; 30; 1) which concatenate segments of motion to create a sequence, such as in (47; 7). The motion capture data can provide very high quality motion, but control is more limited, so it can be a challenge to adapt the motion to novel speech or generated different characters. Gesture relies heavily on hand shape and it can be a challenge to capture good quality hand motion while simultaneously capturing body motion. Some techniques seek to synthesize acceptable hand motion using the body motion alone (14). For a fuller discussion of the issues around hand animation, please refer to (53).

As part of the SAIBA effort, several research groups have developed "behavior realizers" , animation engines capable of realizing commands in the Behavior Markup Language (50) that is supplied by a higher level in an agent architecture. These systems emphasize control and use a combination of procedural data and motion clips (e.g. (15; 48; 49; 11; 45)). The SmartBody system, for example, uses a layering approach based on a hierarchy of controllers for different tasks (e.g. idle motion, locomotion, reach, breathing). These controllers may control different or overlapping parts of the body, which creates a coordination challenge. They can be combined or one controller may override another (45).

Often gesture specification systems will indicate a particular gesture form that is required, e.g. a conduit gesture in which the hand is cupped and moves forward. Systems often employ a dictionary of gesture forms that can be used in syntheses. These gestures have been encoded using motion capture clips, hand animation or numerical spatial specifications. Some techniques (26) have sought to generate the correct forms automatically, for example based on a description of the image trying to be created by the gesture.

Gesture animation is normally deployed in scenarios where it is desirable for the characters to portray clear personalities and show variations in emotion and mood . For these reasons, controlling expressive variation of the motion has been an important focus. A set of challenges must be solved. These include determining how to parameterize a motion to give expressive control, understanding what aspects of motion must be varied to generate a desired impact, ensuring consistency over time, determining how to expose appropriate control structures to the user or character control system and finally, synthesizing the motion to contain the desired properties. Chi et al. (5) use the Effort and Shape components of Laban Movement Analysis to provide an expressive parameterization of motion. Changing any of the four Effort Qualities (Weight, Space, Time and Flow) or the Shape Qualities (Rising-Sinking, Spreading-Enclosing, Advancing-Retreating) will vary the timing and path of the gesture, along with the engagement of the torso. Hartmann et al. (10) use Tension, Continuity and Bias splines (23) to control arm trajectories and provide expressive control through parameters for activation, spatial and temporal extent, fluidity and repetition. Neff and Fiume (41) develop an extensible set of movement properties that can be varied and a system that allows users to write character sketches that reflect a particular character's movement tendencies and then layer additional edits on top.

While gestures are often largely thought of as movements of the arms and hands, and often represented this way in computational systems, they can indeed use the whole body . A character can nod its head, gesture with its toe, etc. More importantly, while arms are the dominant appendages for a motion, engaging the entire body can lead to more clear and effective animation. Lamb called this engagement of the whole body during gesturing Posture-Gesture Merger and argued that it led to a more fluid and attractive motion (29).

3.3 Additional Considerations

Conversations are interactions between people and this must be reflected in the animation. Both the speaker(s) and listener(s) have roles to play . Visual attention must be managed through appropriate gaze behavior to indicate who is paying attention and how actively, along with indicating who is thinking or distracted. Attentive listeners will provide back channel cues, like head nods, to indicate that they are listening and understanding. These must be appropriately timed with the speaker's dialog. Holding the floor is also actively managed. Speakers may decide to yield their turn to another. Listeners may interrupt, and the speaker may yield in response or refuse to do so. Floor management relies on both vocal and gestural cues. Proxemics are also highly communicative to an audience and must be managed appropriately. This creates additional animation challenges in terms of small scale locomotion in order to fluidly manage character placement.

Gestural behavior must adapt to the context . Gestures will be adjusted based on the number of people in the conversation and their physical locations relative to one another. As characters interact, they may also begin to mirror each other's behavior and postures. Gestures are also often used to refer to items in the environment and hence must be adapted based on the character's location. Finally, characters will engage in conversations while also simultaneously performing other activities, such as walking, jogging or cleaning the house. The gesture behavior must be adapted to the constraints of this other behavior, for example gestures performed while jogging tend to be done with more bent arms and are less frequent than standing gestures (52).

4 Future Directions

While significant progress has been made, the bar for conversational gesture animation is very high. We are a long way from being able to easily create synthetic characters that match the expressive quality, range and realism of a skilled actor, and applications that rely on synthetic characters are impoverished by this gap. Some of the key issues to address include :

Characters with Large Gesture Repertoires: It currently takes a great deal of work to build a movement set for a character, generally involving recording, cleaning and retargeting motion capture or hand animating movements. This places a practical limitation on the the number of gestures that they can perform. Methods that allow large sets of gestures to be rapidly generated are needed. A particular challenge is being able to synthesize novel gestures on the fly to react to the character's current context.

Motion Quality: While motion quality has improved, it remains well short of photo-realism, particularly for interactive characters. Hand motion remains a particular challenge, as is appropriate full body engagement. Most systems focus on standing characters whereas people engage in a wide range of activities while simultaneously gesturing. A significant challenge is correctly orchestrating a performance across the various movement modalities (breath, arm movements, body movements, facial expressions, etc.), especially when the motion diverges from playback of a recording or hand-animated sequence.

Planning from Communicative Intent: Systems that can represent an arbitrary communicative intent and can distribute it across various communication modes, and do so in different ways for different speakers, remain a long term goal. This will likely require both improved computational models and a more thorough understanding of how humans formulate communication.

Customization for Characters and Mood: While people tend to have their own, unique gesturing style, it is a challenge to imbue synthetic characters with this expressive range without an enormous amount of manual labor. It is also a challenge to accurate reflect a character's current mood; anger, sadness, irritation, excitement, etc.

Authoring Controls: If a user wishes to create a particular character with a given role, personality, etc., there must be tools to allow this to be authored. Substantial work is required to allow authors to go from an imagined character to an effective realization.

References

1. O. Arikan and D. A. Forsyth. Interactive motion generation from examples. *ACM Transactions on Graphics*, 21(3):483–490, 2002.
2. K. Bergmann, S. Kahl, and S. Kopp. Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints. In *Intelligent Virtual Agents*, pages 203–216. Springer, 2013.
3. K. Bergmann, S. Kopp, and F. Eyssel. Individualized gesturing outperforms average gesturing–evaluating gesture production in virtual humans. In *International Conference on Intelligent Virtual Agents*, pages 104–117. Springer Berlin Heidelberg, 2010.
4. J. Cassell, H. Vilhjálmsson, and T. Bickmore. BEAT: the Behavior Expression Animation Toolkit. In *Proceedings of SIGGRAPH 2001*, pages 477–486, 2001.
5. D. M. Chi, M. Costa, L. Zhao, and N. I. Badler. The EMOTE model for effort and shape. In *Proc. SIGGRAPH 2000*, pages 173–182, 2000.
6. C.-C. Chiu, L.-P. Morency, and S. Marsella. Predicting co-verbal gestures: A deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents*, pages 152–166. Springer International Publishing, 2015.
7. A. Fernández-Baena, R. Montaño, M. Antonijoan, A. Roversi, D. Miralles, and F. Alías. Gesture synthesis adapted to speech emphasis. *Speech Communication*, 57:331–350, 2014.
8. S. Goldin-Meadow. *Hearing gesture: How our hands help us think.* Harvard University Press, 2005.
9. S. Goldin-Meadow. Talking and thinking with our hands. *Current Directions in Psychological Science*, 15(1):34–39, 2006.
10. B. Hartmann, M. Mancini, and C. Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In *Proc. Gesture Workshop 2005*, volume 3881 of *LNAI*, pages 45–55, Berlin; Heidelberg, 2006. Springer.
11. A. Heloir and M. Kipp. EMBR–A Realtime Animation Engine for Interactive Embodied Agents. In *Intelligent Virtual Agents 09*, pages 393–404. Springer, 2009.
12. D. Heylen, S. Kopp, S. C. Marsella, C. Pelachaud, and H. Vilhjálmsson. The next step towards a function markup language. In *International Workshop on Intelligent Virtual Agents*, pages 270–280. Springer, 2008.
13. A. B. Hostetter. When do gestures communicate? a meta-analysis. *Psychological bulletin*, 137(2):297, 2011.
14. S. Jörg, J. Hodgins, and A. Safonova. Data-driven finger motion synthesis for gesturing characters. *ACM Transactions on Graphics (TOG)*, 31(6):189, 2012.
15. M. Kallmann and S. Marsella. Hierarchical motion controllers for real-time autonomous virtual humans. In *Proceedings of the 5th International working conference on Intelligent Virtual Agents (IVA'05)*, pages 243–265, Kos, Greece, September 12-14 2005.
16. A. Kendon. Some relationships between body motion and speech. *Studies in dyadic communication*, 7(177):90, 1972.
17. A. Kendon. How gestures can become like words. *Cross-cultural perspectives in nonverbal communication*, 1:131–141, 1988.
18. A. Kendon. Do gestures communicate? a review. *Research on language and social interaction*, 27(3):175–200, 1994.
19. M. Kipp. *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers, 2005.

20. M. Kipp, M. Neff, K. Kipp, and I. Albrecht. Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In *Proceedings of Intelligent Virtual Agents (IVA07)*, volume 4722 of *LNAI*, pages 15–28. Association for Computational Linguistics, 2007.
21. S. Kita. The temporal relationship between gesture and speech: A study of japanese-english bilinguals. *MS, Department of Psychology, University of Chicago*, 90:91–94, 1990.
22. S. KITA, I. VAN GIJN, and H. VAN DER HULST. Movement phase in signs and co-speech gestures, and their transcriptions by human coders. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 23–35. Springer-Verlag, 1998.
23. D. H. U. Kochanek and R. H. Bartels. Interpolating splines with local tension, continuity, and bias control. *Computer Graphics (Proceedings of SIGGRAPH 84)*, 18(3):33–41, 1984.
24. S. Kopp, K. Bergmann, and S. Kahl. A spreading-activation model of the semantic coordination of speech and gesture. *Proceedings of the 35th Annual Conference of the Cognitive Science Society (CogSci 2013). Cognitive Science Society, Austin (in press, 2013)*, 2013.
25. S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *International Workshop on Intelligent Virtual Agents*, pages 205–217. Springer, 2006.
26. S. Kopp, P. Tepper, and J. Cassell. Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 97–104. ACM, 2004.
27. S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15:39–52, 2004.
28. L. Kovar, M. Gleicher, and F. Pighin. Motion graphs. *ACM Transactions on Graphics*, 21(3):473–482, 2002.
29. W. Lamb. *Posture and gesture: an introduction to the study of physical behavior*. Duckworth, London, 1965.
30. J. Lee, J. Chai, P. S. A. Reitsma, J. K. Hodgins, and N. S. Pollard. Interactive control of avatars animated with human motion data. *ACM Transactions on Graphics*, 21(3):491–500, 2002.
31. J. Lee and S. Marsella. Nonverbal behavior generator for embodied conversational agents. In *Intelligent virtual agents*, pages 243–255. Springer, 2006.
32. S. Levine, P. Krahenbuhl, S. Thrun, and V. Koltun. Gesture controllers. *ACM Transactions on Graphics (TOG)*, 29(4):1–11, 2010.
33. S. Levine, C. Theobalt, and V. Koltun. Real-time prosody-driven synthesis of body language. *ACM Transactions on Graphics (TOG)*, 28(5):1–10, 2009.
34. M. Lhommet and S. C. Marsella. Gesture with meaning. In *Intelligent Virtual Agents*, pages 303–312. Springer, 2013.
35. S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 25–35. ACM, 2013.
36. D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.
37. D. McNeill. *Gesture and thought*. University of Chicago Press, 2005.
38. D. McNeill and E. Levy. Conceptual representations in language activity and gesture. In R. J. Jarvella and W. Klein, editors, *Speech, Place, and Action*, pages 271–295. Wiley, Chichester, 1982.
39. L.-P. Morency, I. de Kok, and J. Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In *International Workshop on Intelligent Virtual Agents*, pages 176–190. Springer Berlin Heidelberg, 2008.
40. M. Neff and E. Fiume. Modeling tension and relaxation for computer animation. In *Proc. ACM SIGGRAPH Symposium on Computer Animation 2002*, pages 81–88, 2002.
41. M. Neff and E. Fiume. AER: Aesthetic Exploration and Refinement for expressive character animation. In *Proc. ACM SIGGRAPH / Eurographics Symposium on Computer Animation 2005*, pages 161–170, 2005.

42. M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 27(1):5:1–5:24, Mar. 2008.
43. S. Nobe. Where do most spontaneous representational gestures actually occur with respect to speech. *Language and gesture*, 2:186, 2000.
44. SAIBA. Working group website, 2012. http://wiki.mindmakers.org/projects:saiba:main.
45. A. Shapiro. Building a character animation system. In *International Conference on Motion in Games*, pages 98–109. Springer Berlin Heidelberg, 2011.
46. M. A. Singer and S. Goldin-Meadow. Children learn when their teacher's gestures and speech differ. *Psychological Science*, 16(2):85–89, 2005.
47. M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler. Speaking with hands: creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics*, 23(3):506–513, 2004.
48. M. Thiebaux, A. Marshall, S. Marsella, and M. Kallman. Smartbody: Behavior realization for embodied conversational agents. In *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, pages 151–158, 2008.
49. H. Van Welbergen, D. Reidsma, Z. Ruttkay, and J. Zwiers. Elckerlyc-A BML realizer for continuous, multimodal interaction with a virtual human. *JMUI*, 2010.
50. H. Vilhjalmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. Marshall, C. Pelachaud, et al. The behavior markup language: Recent developments and challenges. In *Intelligent Virtual Agents*, pages 99–111. Springer, 2007.
51. Y. Wang and M. Neff. The influence of prosody on the requirements for gesture-text alignment. In *Intelligent Virtual Agents*, pages 180–188. Springer, 2013.
52. Y. Wang, K. Ruhland, M. Neff, and C. O'Sullivan. Walk the talk: coordinating gesture with locomotion for conversational characters. *Computer Animation and Virtual Worlds*, 27(3-4):369–377, 2016.
53. N. Wheatland, Y. Wang, H. Song, M. Neff, V. Zordan, and S. Jrg. State of the Art in Hand and Finger Modeling and Animation. *Computer Graphics Forum*, 2015.