# Towards Meaningful Robot Gesture

Maha Salem, Stefan Kopp, Ipke Wachsmuth, Frank Joublin

**Abstract** Humanoid robot companions that are intended to engage in natural and fluent human-robot interaction are supposed to combine speech with non-verbal modalities for comprehensible and believable behavior. We present an approach to enable the humanoid robot ASIMO to flexibly produce and synchronize speech and co-verbal gestures at run-time, while not being limited to a predefined repertoire of motor action. Since this research challenge has already been tackled in various ways within the domain of virtual conversational agents, we build upon the experience gained from the development of a speech and gesture production model used for our virtual human Max. Being one of the most sophisticated multi-modal schedulers, the Articulated Communicator Engine (ACE) has replaced the use of lexicons of canned behaviors with an on-the-spot production of flexibly planned behavior representations. As an underlying action generation architecture, we explain how ACE draws upon a tight, bi-directional coupling of ASIMO's perceptuo-motor system with multi-modal scheduling via both efferent control signals and afferent feedback.

Maha Salem
Research Institute for Cognition and Robotics, Bielefeld University, Germany, e-mail: msalem@cor-lab.uni-bielefeld.de

Stefan Kopp
Sociable Agents Group, Bielefeld University, Germany, e-mail: skopp@techfak.uni-bielefeld.de

Ipke Wachsmuth
Artificial Intelligence Group, Bielefeld University, Germany, e-mail: ipke@techfak.uni-bielefeld.de

Frank Joublin
Honda Research Institute Europe, Offenbach, Germany, e-mail: frank.joublin@honda-ri.de

## 1 Introduction

Non-verbal expression via gesture is an important feature of social interaction, frequently used by human speakers to emphasize or supplement what they express in speech. For example, pointing to objects being referred to or giving spatial directions conveys information that can hardly be encoded solely by speech. Accordingly, humanoid robot companions that are intended to engage in natural and fluent human-robot interaction must be able to produce speech-accompanying non-verbal behaviors from conceptual, to-be-communicated information. Forming an integral part of human communication, hand and arm gestures are primary candidates for extending the communicative capabilities of social robots.

According to McNeill [13], co-verbal gestures are mostly generated unconsciously and are strongly connected to speech as part of an integrated utterance, yielding semantic, pragmatic and temporal synchrony between both modalities. This suggests that gestures are influenced by the communicative intent and by the accompanying verbal utterance in various ways. In contrast to task-oriented movements like reaching or grasping, human gestures are derived to some extent from a kind of internal representation of shape [8], especially when iconic or metaphoric gestures are used. Such characteristic shape and dynamical properties exhibited by gestural movement enable humans to distinguish them from subsidiary movements and to perceive them as meaningful [17]. Consequently, the generation of co-verbal gestures for artificial humanoid bodies, e.g., as provided for virtual agents or robots, demands a high degree of control and flexibility concerning shape and time properties of the gesture, while ensuring a natural appearance of the movement.

In this paper, we first discuss related work, highlighting the fact that not much research has so far focused on the generation of robot gesture (Section 2). In Section 3, we describe our multi-modal behavior realizer, the Articulated Communicator Engine (ACE), which implements the speech-gesture production model originally designed for the virtual agent Max and is now used for the humanoid robot ASIMO. We then present a concept for the generation of meaningful arm movements for the humanoid robot ASIMO based on ACE in Section 4. Finally, we conclude and give an outlook of future work in Section 5.

## 2 Related Work

At present, the generation together with the evaluation of the effects of robot gesture is largely unexplored. In traditional robotics, recognition rather than synthesis of gesture is mainly brought into focus. In existing cases of gesture synthesis, however, models typically denote object manipulation serving little or no communicative function. Furthermore, gesture generation is often based on prior recognition of perceived gestures, hence the aim is often to imitate these movements. In many cases in which robot gesture is actually generated with a communicative intent, these arm movements are not produced at run-time, but are pre-recorded for demonstration purposes and are not finely coordinated with speech. Generally, only a few approaches share any similarities with ours, however, they are mostly realized on

less sophisticated platforms with less complex robot bodies (e.g., limited mobility, less degrees of freedom (DOF), etc.). One example is the personal robot Maggie [6] whose aim is to interact with humans in a natural way, so that a peer-to-peer relationship can be established. For this purpose, the robot is equipped with a set of pre-defined gestures, but it can also learn some gestures from the user. Another example of robot gesture is given by the penguin robot Mel [16] which is able to engage with humans in a collaborative conversation, using speech and gesture to indicate engagement behaviors. However, gestures used in this context are predefined in a set of action descriptions called the "recipe library". A further approach is that of the communication robot Fritz [1], using speech, facial expression, eye-gaze and gesture to appear livelier while interacting with people. Gestures produced during interactional conversations are generated on-line and mainly consist of human-like arm movements and pointing gestures performed with eyes, head, and arms.

As Minato et al. [14] state, not only the behavior but also the appearance of a robot influences human-robot interaction. Therefore, the importance of the robot's design should not be underestimated if used as a research platform to study the effect of robot gesture on humans. In general, only few scientific studies regarding the perception and acceptance of robot gesture have been carried out so far. Much research on the human perception of robots depending on their appearance, as based on different levels of embodiment, has been conducted by MacDorman and Ishiguro [12], the latter widely known as the inventor of several android robots. In their testing scenarios with androids, however, non-verbal expression via gesture and gaze was generally hard-coded and hence pre-defined. Nevertheless, MacDorman and Ishiguro consider androids a key testing ground for social, cognitive, and neuroscientific theories. They argue that they provide an experimental apparatus that can be controlled more precisely than any human actor. This is in line with initial results, indicating that only robots strongly resembling humans can elicit the broad spectrum of responses that people typically direct toward each other. These findings highlight the importance of the robot's design when used as a research platform for the evaluation of human-robot interaction scenarios.

While being a fairly new area in robotics, within the domain of virtual humanoid agents, the generation of speech-accompanying gesture has already been addressed in various ways. Cassell et al. introduced the REA system [2] over a decade ago, employing a conversational humanoid agent named Rea that plays the role of a real estate salesperson. A further approach, the BEAT (Behavior Expression Animation Toolkit) system [3], allows for appropriate and synchronized non-verbal behaviors by predicting the timing of gesture animations from synthesized speech in which the expressive phase coincides with the prominent syllable in speech. Gibet et al. generate and animate sign-language from script-like specifications, resulting in a simulation of fairly natural movement characteristics [4]. However, even in this domain most existing systems either neglect the meaning a gesture conveys, or they simplify matters by using lexicons of words and canned non-verbal behaviors in the form of pre-produced gestures.

In contrast, the framework underlying the virtual agent Max [9] is geared towards an integrated architecture in which the planning of both content and form across both

modalities is coupled [7], hence giving credit to the meaning conveyed in non-verbal utterances. According to Reiter and Dale [15], computational approaches to generating multi-modal behavior can be modeled in terms of three consecutive tasks: firstly, determining *what* to convey (i.e., content planning); secondly, determining *how* to convey it (i.e., micro-planning); finally, realizing the planned behaviors (i.e., surface realization). Although the Articulated Communicator Engine (ACE) itself operates on the surface realization layer of the generation pipeline, the overall system used for Max also provides an integrated content planning and micro-planning framework [7]. Within the scope of this paper, however, only ACE is considered and described, since it marks the starting point required for the interface endowing the robot ASIMO with similar multi-modal behavior.

## 3 An Incremental Model of Speech-Gesture Production

Our approach is based on straightforward descriptions of the designated outer form of the to-be-communicated multi-modal utterances. For this purpose, we use MURML [11], the XML-based Multi-modal Utterance Representation Markup Language, to specify verbal utterances in combination with co-verbal gestures [9]. These, in turn, are explicitly described in terms of form features (i.e., the posture aspired for the gesture stroke), specifying their affiliation to dedicated linguistic elements based on matching time identifiers. Fig. 1 shows an example of a MURML specification which can be used as input for our production model. For more information on MURML see [11].

```
<definition><utterance>
 <specification>
 And now take the object <time id="t1" chunkborder="true"/>
 and make it <time id="t2"/> this big. <time id="t3"/>
 </specification>
 <behaviorspec>
  <gesture id="gesture_1" scope="hand">
   <affiliate onset="t2" end="t3" focus="this"/>
    <constraints>
     <symmetrical dominant="right_arm" symmetry="SymMS">
      <parallel>
       <static slot="HandShape" value="BSflat(FBround all o)"/>
       <static slot="ExtFingerOrientation" value="DirA"/>
       <static slot="PalmOrientation" value="DirL"/>
       <static slot="HandLocation" value="LocChest LocCenterRight LocNorm"/>
      </parallel>
     </symmetrical>
    </constraints>
  </gesture>
 </behaviorspec>
</utterance></definition>
```

**Fig. 1** Example of a MURML specification for multi-modal utterances.

The concept underlying the multi-modal production model is based on an empirically suggested assumption referred to as *segmentation hypothesis* [13], according to which the co-production of continuous speech and gesture is organized in successive segments. Each of these, in turn, represents a single idea unit which we refer

to as a *chunk* of speech-gesture production. A given chunk consists of an intonation phrase and a co-expressive gesture phrase, concertedly conveying a prominent concept [10]. Within a chunk, synchrony is mainly achieved by gesture adaptation to structure and timing of speech, while absolute time information is obtained at phoneme level and used to establish timing constraints for co-verbal gestural movements. Given the MURML specification shown in Fig. 1, the correspondence between the verbal phrase and the accompanying gesture is established by the <time id="..."/> tag with unique identifier attributes. Accordingly, the beginning and ending of the affiliate gesture is defined using the <affiliate onset="..." end="..."/> tag. The incremental production of successive coherent chunks is realized by processing each chunk on a separate 'blackboard' running through a sequence of states (Fig. 2). These states augment the classical two-phase planning - execution procedure with additional phases, in which the production process of subsequent chunks can interact with one another.

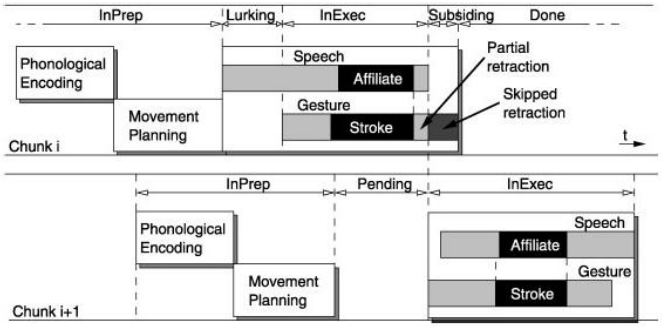


**Fig. 2** Blackboards run through a sequence of processing states for incremental production of multi-modal chunks.

This approach for gesture motor control is based on a hierarchical concept: During higher-level planning, the motor planner is provided with timed form features as described in the MURML specification. This information is then transferred to independent motor control modules. Such a functional-anatomical decomposition of motor control aims at breaking down the complex control problem into solvable sub-problems [18]. ACE [10] provides specific motor planning modules, amongst others, for the arms, the wrists, and the hands, which instantiate local motor programs (LMPs). These are used to animate required sub-movements and operate within a limited set of DOFs and over a designated period of time (Fig. 3). For each limb's motion, an abstract motor control program (MCP) coordinates and synchronizes the concurrently running LMPs for an overall solution to the control problem. The overall control framework, however, does not attend to how such submovements are controlled. To allow for an effective interplay of the LMPs within a MCP, the planning modules arrange them into a controller network which defines

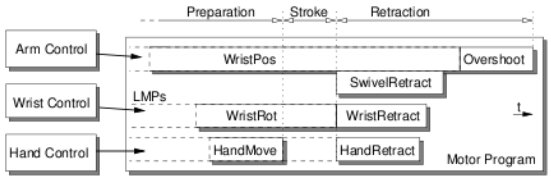


**Fig. 3** Composition of motion control in Local Motor Programs (LMPs) for a hand-arm gesture.

their potential interdependencies for mutual (de-)activation. LMPs are able to transfer activation between themselves and their predecessors or successors to ensure context-dependent gesture transitions. Consequently, they can activate or deactivate themselves at run-time based on feedback information on current movement conditions. Once activated, LMPs are continuously applied to the kinematic skeleton in a feedforward manner.

The on-the-fly timing of gestures is accomplished by the ACE engine as follows: The gesture stroke phase (the expressive 'core' phase) is set to accompany the co-expressive phase in speech (the 'affiliate') as annotated in the MURML specification. The ACE scheduler retrieves timing information about the synthetic speech at the millisecond level and defines the gesture stroke to start and end accordingly. These temporal constraints are automatically propagated down to each single gesture component (e.g. how long the hand has to form a certain shape). The motor planner then creates the LMPs that meet both the temporal constraints and the form constraints. The second aspect of scheduling, namely, the decision to skip preparation or retraction phases, emerges automatically from the interplay of motor programs at run-time. Motor programs monitor the body's current movements and decide when to activate themselves and to take action in order to realize the planned gesture stroke as scheduled. A retraction phase is skipped when the motor program of the next gesture takes over the control of the effectors from the previous program. This online scheduling creates fluent and continuous multi-modal behavior. It is possible because of the interleaved production of successive chunks of multi-modal behavior in ACE and has been employed successfully in several virtual humans.

## 4 Control Architecture for Robot Gesture

By re-employing existing concepts from the domain of virtual conversational agents, our goal is to similarly enable the robot to flexibly produce speech and co-verbal gesture at run-time. This requires a robot control architecture that combines conceptual representation and planning with motor control primitives for speech and arm movements, thereby endowing ASIMO with 'conceptual motorics'.

Since gesture generation with ACE is based on external form features as given in the MURML description, arm movement trajectories are specified directly in task space. The calculated vector information is passed on to the robot motion control

module which instantiates the actual robot movement. For this purpose, externally formulated locol motor programs (for wrist position and preparation/stroke of wrist flexion and swivel movement) are invoked first. Subsequently, inverse kinematics (IK) of the arm is solved on the velocity level using the ASIMO whole body motion (WBM) controller framework [5]. WBM aims to control all DOF of the humanoid robot by given end-effector targets, providing a flexible method of controlling upper body movement by specifying only relevant task dimensions selectively in real-time. For this purpose, task-specific command elements can be assigned to the command vector at any time, enabling the system to control one or multiple effectors while generating a smooth and natural movement. Redundancies are optimized regarding joint limit avoidance and self-collision avoidance. For more details on WBM control for ASIMO see [5].

Once IK has been solved for the internal body model provided for WBM control, the joint space description of the designated trajectory is applied to the real ASIMO robot. Due to constraints imposed by the robot's physical architecture and motor control, however, the inner states represented within the WBM controller might deviate from the actual motor states of the real robot during run-time. For this reason, a bi-directional interface for both efferent and afferent signaling is required. This is realized by a feedback loop, updating the internal model of ASIMO in WBM as well as the kinematic body model coupled to ACE at a sample rate $r$. Note, however, that for successful integration, this process needs to synchronize the two competing sample rates of the ACE framework on the one hand, and the WBM software controlling ASIMO on the other hand. Fig. 4 illustrates our proposed control architecture for robot gesture.

A main advantage of this approach to robot control in combination with ACE is the formulation of the trajectory in terms of effector targets in task space, which are then used to derive a joint space description using the standard WBM controller for ASIMO. An alternative method would aim at the extraction of joint angle values from ACE and a subsequent mapping onto the robot body model. This, however, might lead to joint states that are not feasible on the robot, since ACE was originally designed for a virtual agent application and does not entirely account for certain physical restrictions such as collision avoidance. As a result, solving IK using ASIMO's internally implemented WBM controller ensures safer postures for the robot. However, due to deviations from original postures and respective joint angles, the outer form of a gesture might be distorted such that its original meaning is altered. Therefore, whether and how the form and meaning of gestures are affected will be subject to further evaluation as work progresses.

## 5 Conclusion and Future Work

We presented a robot control architecture to enable the humanoid robot ASIMO to flexibly produce and synchronize speech and co-verbal gestures at run-time. The framework is based on a speech and gesture production model originally developed for a virtual human agent. Being one of the most sophisticated multi-modal

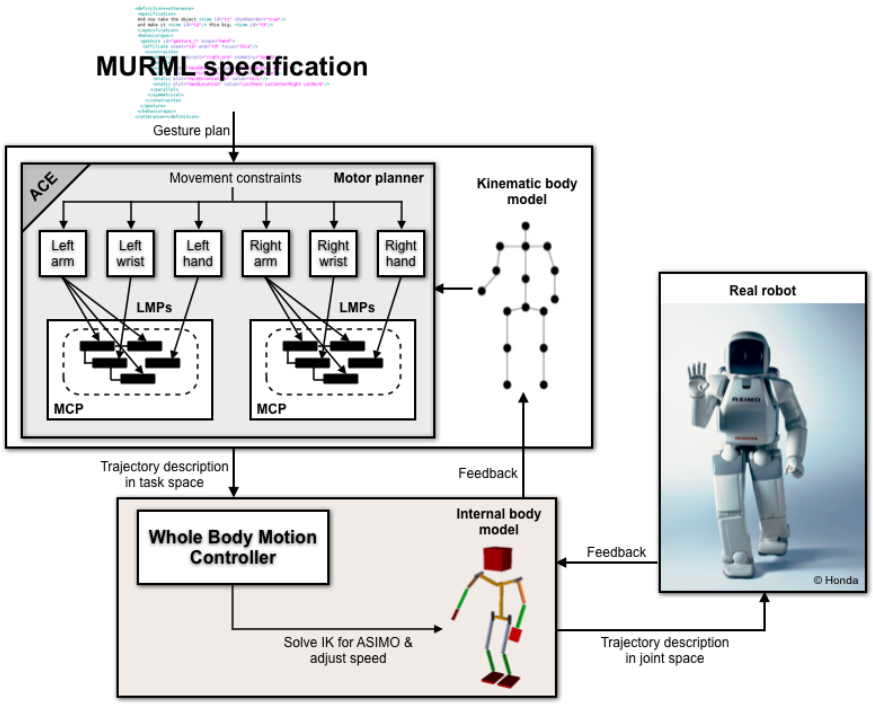


**Fig. 4** Control architecture for robot gesture.

schedulers, the Articulated Communicator Engine (ACE) allows for an on-the-spot production of flexibly planned behavior representations. By re-employing ACE as an underlying action generation architecture, we draw upon a tight coupling of ASIMO's perceptuo-motor system with multi-modal scheduling. This has been realized in a bi-directional robot control architecture which uses both efferent actuator control signals and afferent sensory feedback. Our framework combines conceptual, XML-based representation and planning with motor control primitives for speech and arm movements. This way, pre-defined canned behaviors can be replaced by conceptual motorics generated at run-time.

The requirement to meet strict synchrony constraints to ensure temporal and semantic coherence of communicative behavior presents a main challenge to our framework. Clearly, the generation of finely synchronized multi-modal utterances proves to be more demanding when realized on a robot with a physically constrained body than for an animated virtual agent, especially when communicative signals are to be produced at run-time. Currently, synchrony is mainly achieved by gesture adaptation to structure and timing of speech, obtaining absolute time information at phoneme level. To tackle this challenge the cross-modal adaptation mechanisms applied in ACE will be extended to allow for a finer mutual adaptation between robot gesture and speech.

**Acknowledgements** The research project "Conceptual Motorics" is based at the Research Institute for Cognition and Robotics, Bielefeld University, Germany. It is supported by the Honda Research Institute Europe.

## References

1. M. Bennewitz, F. Faber, D. Joho, and S. Behnke. Fritz – A humanoid communication robot. In *RO-MAN 07: Proc. of the 16th IEEE International Symposium on Robot and Human Interactive Communication*, 2007.
2. J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan. Human conversation as a system framework: desigining embodied conversational agents. In *Embodied Conversational Agents*, pages 29–63. MIT Press: Cambridge, MA, 2000.
3. J. Cassell, H. Vilhjalmsson, and T. Bickmore. Beat: the behavior expression animation toolkit. In *Proceedings of SIGGRAPH'01*, 2001.
4. S. Gibet, T. Lebourque, and P.-F. Marteau. High-level specification and animation of communicative gestures. *Journal of Visual Languages and Computing*, 12(6):657–687, 2001.
5. M. Gienger, H. Janßen, and S. Goerick. Task-oriented whole body motion for humanoid robots. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, Tsukuba, Japan, 2005. Accepted.
6. J.F. Gorostiza, R. Barber, A.M. Khamis, M. Malfaz, R. Pacheco, R. Rivas, A. Corrales, E. Delgado, and M.A. Salichs. Multimodal human-robot interaction framework for a personal robot. In *RO-MAN 06: Proc. of the 15th IEEE International Symposium on Robot and Human Interactive Communication*, 2006.
7. S. Kopp, K. Bergmann, and I. Wachsmuth. Multimodal communication from multimodal thinking - towards an integrated model of speech and gesture production. *Semantic Computing*, 2(1):115–136, 2008.
8. S. Kopp and I. Wachsmuth. A Knowledge-based Approach for Lifelike Gesture Animation. In W. Horn, editor, *ECAI 2000 - Proceedings of the 14th European Conference on Artificial Intelligence*, pages 663–667, Amsterdam, 2000. IOS Press.
9. S. Kopp and I. Wachsmuth. Model-based Animation of Coverbal Gesture. In *Proceedings of Computer Animation 2002*, pages 252–257. IEEE Press, 2002.
10. S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52, 2004.
11. A. Kranstedt, S. Kopp, and I. Wachsmuth. MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents. In *Proceedings of the AAMAS02 Workshop on Embodied Conversational Agents - let's specify and evaluate them*, Bologna, Italy, July 2002.
12. K.F. Macdorman and H. Ishiguro. The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, 7(3):297–337, 2006.
13. D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.
14. T. Minato, M. Shimada, H. Ishiguro, and S. Itakura. Development of an android robot for studying human-robot interaction. *Innovations in Applied Artificial Intelligence*, pages 424–434, 2004.
15. E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Cambridge Univ. Press, 2000.
16. C.L. Sidner, C. Lee, and N. Lesh. The role of dialog in human robot interaction. In *International Workshop on Language Understanding and Agents for Real World Interaction*, 2003.
17. I. Wachsmuth and S. Kopp. Lifelike Gesture Synthesis and Timing for Conversational Agents. In I. Wachsmuth and T. Sowa, editors, *Gesture and Sign Language in Human-Computer Interaction*, LNAI 2298, pages 120–133, Berlin, 2002. Springer.
18. D. Zeltzer. Motor control techniques for figure animation. *IEEE Computer Graphics Applications*, 2(9):53–59, 1982.