Generation and Evaluation of Communicative Robot Gesture

Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing & Frank Joublin

International Journal of Social Robotics

ISSN 1875-4791
Volume 4
Number 2

Int J of Soc Robotics (2012) 4:201-217
DOI 10.1007/s12369-011-0124-9

Springer

Int J Soc Robot (2012) 4:201–217
DOI 10.1007/s12369-011-0124-9

Generation and Evaluation of Communicative Robot Gesture

Maha Salem · Stefan Kopp · Ipke Wachsmuth · Katharina Rohlfing · Frank Joublin

Accepted: 7 November 2011 / Published online: 4 February 2012

Abstract How is communicative gesture behavior in robots perceived by humans? Although gesture is crucial in social interaction, this research question is still largely unexplored in the field of social robotics. Thus, the main objective of the present work is to investigate how gestural machine behaviors can be used to design more natural communication in social robots. The chosen approach is twofold. Firstly, the technical challenges encountered when implementing a speech-gesture generation model on a robotic platform are tackled. We present a framework that enables the humanoid robot to flexibly produce synthetic speech and co-verbal hand and arm gestures at run-time, while not being limited to a predefined repertoire of motor actions. Secondly, the achieved flexibility in robot gesture is exploited in controlled experiments. To gain a deeper understanding of how communicative robot gesture might impact and shape human perception and evaluation of human-robot interaction, we conducted a between-subjects experimental study using the humanoid robot in a joint task scenario. We manipulated the non-verbal behaviors of the robot in three experimental conditions, so that it would refer to objects by utilizing either (1) unimodal (i.e., speech only) utterances, (2) congruent multimodal (i.e., semantically matching speech and gesture) or (3) incongruent multimodal (i.e., semantically non-matching speech and gesture) utterances. Our findings reveal that the robot is evaluated more positively when non-verbal behaviors such as hand and arm gestures are displayed along with speech, even if they do not semantically match the spoken utterance.

Keywords Multimodal interaction and conversational skills · Non-verbal cues and expressiveness · Social human-robot interaction · Robot companions and social robots

M. Salem (✉)
Research Institute for Cognition and Robotics, Bielefeld, Germany
e-mail: msalem@cor-lab.uni-bielefeld.de

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

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

K. Rohlfing
Emergentist Semantics Group, Bielefeld University, Bielefeld, Germany
e-mail: kjr@uni-bielefeld.de

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

1 Introduction

One of the main objectives of social robotics research is to design and develop robots that can engage in social environments in a way that is appealing and familiar to human interaction partners. However, interaction is often difficult because inexperienced users do not understand the robot's internal states, intentions, actions, and expectations. Thus, to facilitate successful interaction, social robots should provide communicative functionality that is both natural and intuitive. The appropriate level of such communicative functionality strongly depends on the appearance of the robot and attributions thus made to it. Given the design of humanoid robots, they are typically expected to exhibit human-like communicative behaviors, using their bodies for non-verbal

expression just as humans do. Representing an important feature of human communication, co-verbal hand and arm gestures are frequently used by human speakers to illustrate what they express in speech [24]. Crucially, gestures help to convey information which speech alone cannot provide, as in referential, spatial or iconic information [11]. At the same time, human listeners have been shown to be well-attentive to information conveyed via such non-verbal behaviors [7]. Moreover, providing multiple modalities helps to dissolve ambiguity typical of unimodal communication and, as a consequence, to increase robustness of communication. Thus, it appears reasonable to equip humanoid robots that are intended to engage in natural and comprehensible human-robot interaction with speech-accompanying gestures.

1.1 Gesture in Human Communication

Gesture is a phenomenon of human communication that has been studied by researchers from various disciplines for many years. A multiplicity of hand, arm and body movements can all be considered to be gestures, and although definitions and categorizations vary widely, much gesture research has sought to describe the different types of gesture, e.g., [15, 24]. McNeill [24], for example, categorizes four main types of gesture based on semiotics: (1) iconics, i.e., gestures representing images of concrete entities and/or actions; (2) metaphorics, i.e., gestures whose pictorial content presents abstract ideas rather than concrete objects; (3) deictics, i.e., pointing gestures; and (4) beats, i.e., hand movements performed along with the rhythmical pulsation of speech without conveying semantic information. In his later work, however, McNeill [25] claims that the search for categories actually seems misled: since the majority of gestures are multifaceted, it is more appropriate to think in terms of combinable dimensions rather than categories. In this way, dimensions can be combined without the need for a hierarchy. Unlike task-oriented movements like reaching or object manipulation, human gestures are partly derived from an internal representation of 'shape' [16], which particularly applies to iconic or metaphoric gestures. Such characteristic shape and dynamical properties enable humans to distinguish gestures from subsidiary movements and to perceive them as meaningful [42].

In this paper, we use the term *gesture* to refer specifically to representational gestures [12], i.e., movements that co-express the content of speech by pointing to a referent in the physical environment (deictic gestures) or gestures depicting a referent with the motion or shape of the hands (iconic gestures). Other types of gesture such as beat gestures (movements that emphasize the prosody or structure of speech), emblems (movements that convey conventionalized meanings) and turn-taking gestures (movements that regulate interaction between speakers) fall outside the scope of the present work.

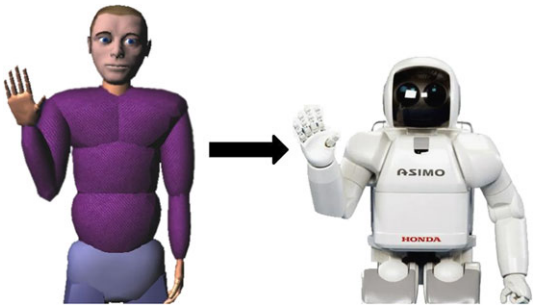

Fig. 1 The goal of the present work is to realize speech and non-verbal behavior generation for the physical Honda humanoid robot (*right*) by transferring an existing virtual agent framework as employed for the agent Max (*left*) and to subsequently evaluate it in controlled experiments of human-robot interaction

1.2 Gesture Behavior for Artificial Communicators

To endow a humanoid robot with communicative co-verbal gestures, it requires a large degree of flexible control especially with regards to shape properties of the gesture. At the same time, adequate timing and natural appearance of these body movements are essential to add to the impression of the robot's liveliness. Since the challenge of multimodal behavior realization for artificial humanoid bodies has already been explored in the context of virtual conversational agents, our approach builds upon an existing solution from this research area [35, 42]. The Articulated Communicator Engine (ACE) [17] implements the speech-gesture production model that was originally designed for the virtual human agent *Max* and is now used as the underlying action generation framework for the Honda humanoid robot (Fig. 1). Based on the implementation of such a speech and gesture production model for humanoid robot gesture [36], we exploit the achieved flexibility in communicative robot behavior in a controlled experimental study to investigate how humans experience a humanoid robot that performs gestures during interaction. This way, we try to shed light onto human perception and understanding of gestural machine behaviors and how these can be used to design more natural communication in social robots.

The rest of this paper is organized as follows. We first discuss related work in Sect. 2, showing that not much research has focused on the generation and evaluation of robot gesture. In Sect. 3, we describe our multimodal behavior realizer, the Articulated Communicator Engine (ACE), which implements the speech-gesture production model originally designed and implemented for the virtual human agent Max and is now used for the Honda humanoid robot (Fig. 1). We then describe our approach to a robot control architecture employing ACE for producing gestural hand and arm movements for the humanoid robot in Sect. 4. Subsequently, gesture representations realized in our controller framework are

presented, evaluated and discussed in Sect. 5. We further describe the empirical study conducted to evaluate robot gesture in a human-robot interaction scenario and present an evaluation and discussion of results in Sect. 6. Finally, we conclude and give an outlook of future work in Sect. 7.

2 Related Work

Two research areas are relevant to the present work: firstly, in the area of computer animation, researchers have developed frameworks to realize multimodal communication behavior in virtual conversational agents; secondly, in the field of robotics, researchers have explored various approaches to generate non-verbal behaviors along with speech in humanoid robots. The challenges are similar in that both research areas demand a high degree of control and flexibility so that human-like motion can be adapted to a system with non-human kinematics. The levels of complexity encountered in each field, however, are not equivalent. Although the range of different body types found in virtual embodied agents is manifold and hence challenging, character animation has less restrictive motion than even the most state-of-the-art humanoid robots [33]. For example, animation of virtual agents reduces or even eliminates the problems of handling joint and velocity limits; in a robot body, however, these have to be explicitly addressed given real physical restrictions.

2.1 Virtual Agents

In contrast to the research field of robotics, the challenge of generating speech and co-verbal gesture has already been tackled in various ways within the domain of virtual human agents. Some of the earliest work includes that of Cassell et al. who presented the *REA* system [5] in which a conversational humanoid agent operates as a real estate salesperson. A more recent approach is that of the interactive expressive system *Greta* [31] which is able to communicate using verbal and non-verbal modalities. Even in the domain of virtual conversational agents, however, most existing systems simplify matters by using lexicons of words and canned non-verbal behaviors in the form of pre-produced gestures [9]. In contrast, the ACE framework underlying the virtual agent *Max* [17] builds upon an integrated architecture in which the planning of both content and form across both modalities is coupled [18], thereby taking into account the meaning conveyed in non-verbal utterances. For this reason, our proposed approach benefits from transferring a sophisticated multimodal behavior scheduler from a virtual conversational agent to a physical robot.

In addition to the technical contributions presented in the area of embodied conversational agents, there has also been active work in evaluating complex gesture models for the animation of virtual characters. Several studies have investigated and compared the human perception of traits such as naturalness in virtual agents. In one such study [19], the conversational agent Max communicated by either utilizing a set of co-verbal gestures alongside speech, typically by self-touching or movement of the eyebrows, or by utilizing speech alone without any such accompanying gestures. Human participants were then invited to rate their perception of Max's behavioral-emotional state, for example, its level of aggressiveness, its degree of liveliness, etc. Crucially, the results of the study suggested that virtual agents are perceived in a more positive light when they are able to produce co-verbal gestures alongside speech (rather than acting in a speech-only modality). In [2] Bergmann et al. modeled the gestures of Max based on real humans' non-verbal behavior and subsequently set out to question the communicative quality of these models via human participation. The main finding was that Max was perceived as more likable, competent and human-like when gesture models based on individual speakers were applied, as opposed to combined gestures of a collection of speakers, random gestures, or no gestures.

2.2 Robotics

Although much of the robotics research has been dedicated to the area of gesture recognition and analysis, only few approaches have pursued both the generation of humanoid robot gesture and the investigation of human perception of such robot behavior. Within the few existing approaches that are actually dedicated to gesture synthesis, the term "gesture" has been widely used to denote object manipulation tasks rather than non-verbal communicative behaviors. For example, Calinon and Billard [4] refer to the drawing of stylized alphabet letters as gestures in their work. Many researchers have focused on the translation of human motion for gesture generation in various robots, usually aiming at imitation of movements captured from a human demonstrator, e.g., [3]. Miyashita et al. [27] and Pollard et al. [33] present further techniques for limiting human motion of upper body gestures to movements achievable by a variety of different robotic platforms. These models of gesture synthesis, however, mainly focus on the technical aspects of generating robotic motion that fulfills little or no communicative function. In addition, they are limited in that they do not combine generated non-verbal behaviors with further output modalities such as speech.

Only a few approaches in robotics incorporate both speech and gesture synthesis; however, in most cases the robots are equipped with a set of pre-recorded gestures that are not generated on-line but simply replayed during human-robot interaction, as seen in [8] or [40]. Moreover, a majority of approaches focusing on gesture synthesis for humanoid robots are limited to the implementation and evaluation of a

single type of gesture, typically deictic (e.g., [32, 41]) or emblematic gestures (e.g., [13]) instead of providing a general framework that can handle all types of gesture. The communication robot presented in [1] is one of the few systems in which different types of gesture are actually generated on-line. These mainly consist of arm movements and pointing gestures performed synchronously with eyes, head, and arms, and are accompanied by speech to make the robot appear livelier. However, all aforementioned approaches are realized on platforms with less complex robot bodies which, for example, comprise fewer degrees of freedom (DOF), have limited mobility, and perform body movements in a rather jerky fashion (as seen in [1]). Moreover, many of these robots expose only little or no humanoid traits.

As stated in [26], however, the appearance of a robot can be just as important as its behavior when evaluating the experience felt by human interaction partners. In other words, the robot's design is crucial if we are to eventually study the effect of robot gesture on humans. MacDorman and Ishiguro [22] have researched human perception of robot appearance as based on different levels of embodiment, with android robots representing the most anthropomorphic form. Although an innovative approach, android robots only feature certain hard-coded gestures and thus still lack any real-time gesture-generating capability. Moreover, findings presented in [38] suggest that the mismatch between the highly human-like appearance of androids and their mechanical, less human-like movement behavior may lead to increased prediction error in the brain, possibly accounting for the 'uncanny valley' phenomenon [28]. Thus, a major advantage of using the Honda humanoid robot as a research platform lies in its humanoid, yet not too human-like appearance and smooth, yet not completely natural movement behavior. Although the Honda robot cannot mimic any facial expression, it is favorable for us to use such a robot, as the focus of the present work lies on hand and arm gestures. This way, the perception of the robot's gestural arm movements can be assessed as the primary non-verbal behavior.

2.3 Evaluation of Robot Gesture

Despite the interesting implications of the evaluation studies conducted with virtual agents, we must be cautious when transferring the findings from the domain of animated graphical characters to the domain of social robots. Firstly, the presence of real physical constraints can alter the perceived level of realism. Secondly, given the greater degree of embodiment that is possible in a real-world system, interaction with a robot is potentially richer; human participants could, for example, walk around or even touch a real robot. This makes the interaction experience more complex and is naturally expected to affect the outcome of the results.

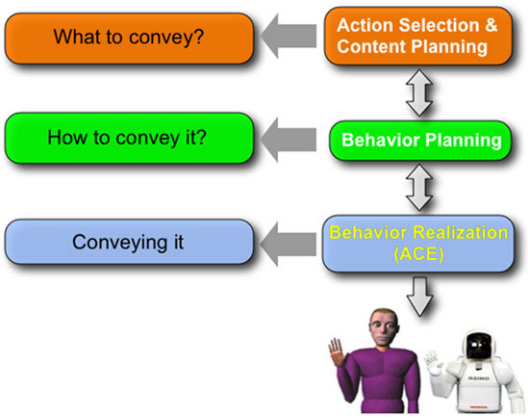

Fig. 2 Behavior generation pipeline adapted from Reiter and Dale [34, 36]

One of the few models that resembles our approach in that it attempts to generate and evaluate a multitude of gesture types for the Honda humanoid robot was presented by Ng-Thow-Hing et al. [30]. Their proposed model reconstructs the communicative intent through text and parts-of-speech analysis to select appropriate gestures. The evaluation of the system, however, was merely undertaken using several video-based studies.

We argue that, in order to obtain a representative assessment of robot gesture and the human perception thereof, it is necessary to evaluate such non-verbal behavior in actual interaction scenarios. As gesture scope and space can only be accurately observed and assessed in a true interaction, we decided to conduct an experimental study using our speech-gesture synthesis model implemented on the Honda humanoid robot. Since the evaluation of the effects and acceptance of communicative robot gesture is still largely unexplored, we attempt to investigate whether multimodal robot behavior, i.e., displaying gesture along with speech, is desired by human interaction partners and favored over unimodal communication.

3 Integrated Model of Speech-Gesture Production

Computational approaches to synthesizing multimodal behavior can be modeled as three consecutive tasks [34] (Fig. 2): firstly, determining *what* to convey (i.e., content planning); secondly, determining *how* to convey it (i.e., behavior planning); finally, conveying it (i.e., behavior realization). Addressing the third task of this behavior generation pipeline, the Articulated Communicator Engine (ACE) operates at the behavior realization layer, yet the overall system used by the virtual agent Max also provides an integrated content planning and behavior planning framework [18]. The present work focuses on ACE which forms the starting point for an interface endowing the humanoid robot with similar multimodal behavior.

Fig. 3 A feature-based MURML specification for multimodal utterances

```
<definition><utterance>
 <specification>
  The bathroom is <time id="t1"/> over there. <time id="t2">
 </specification>
 <behaviorspec>
  <gesture id="gesture_1" scope="hand">
   <affiliate onset="t1" end="t2" focus="there"/>
    <constraints>
     <parallel>
      <static slot="HandShape" value=" BSflat (FBround all o)"/>
      <static slot="ExtFingerOrientation" value="DirA"/>
      <static slot="PalmOrientation" value="DirR"/>
      <static slot="HandLocation" value="LocShoulder LocCenterLeft LocStretched"/>
     </parallel>
    </constraints>
  </gesture>
 </behaviorspec>
</utterance></definition>
```

3.1 Utterance Specification

Within the ACE framework, utterance specifications can be described in two different ways using the Multimodal Utterance Representation Markup Language (MURML [20]). Firstly, verbal utterances together with co-verbal gestures can be specified as feature-based descriptions in which the outer form features of a gesture (i.e., the posture of the gesture stroke) are explicitly described. Gesture affiliation to dedicated linguistic elements is determined by matching time identifiers. Figure 3 illustrates an example of a feature-based MURML specification for speech-gesture production. Secondly, gestures can be specified as key-frame animations in which each key-frame specifies a part of the overall gesture movement pattern describing the current state of each joint. Speed information for the interpolation between every two key-frames and the corresponding affiliation to parts of speech is obtained from assigned time identifiers. Key-frame animations in ACE can be defined either manually or derived from motion capturing data from a human demonstrator, allowing the animation of virtual agents in real-time. In our present work we focus on the generation of feature-based utterance descriptions, although key-frame animations—and therewith captured human motion—can also be realized on the robot using the same interface.

3.2 Gesture Motor Control

Gesture motor control is realized hierarchically in ACE: during higher-level planning, the motor planner is provided with timed form features as annotated in the MURML specification. This information is then passed on to independent motor control modules. The idea behind this functional-anatomical decomposition of motor control is to break down the complex control problem into solvable sub-problems. ACE [17] provides specific motor planning modules for the arms, the wrists, and the hands which, in turn, instantiate local motor programs (LMPs). These are used to animate required sub-movements. LMPs operate within a limited set of DOF and over a designated period of time. For the motion of each limb, an abstract motor control program (MCP) coordinates and synchronizes the concurrently running LMPs, gearing towards an overall solution to the control problem. The top-level control of the ACE framework, however, does not attend to how such sub-movements are controlled. To ensure an effective interplay of the LMPs involved in a MCP, the planning modules arrange them in a controller network which defines their potential interdependencies for mutual (de-)activation. LMPs are able to transfer activation between themselves and their predecessors or successors to allow for context-dependent gesture transitions. Thus, they can activate or deactivate themselves at run-time depending on feedback information on current movement conditions.

3.3 Speech Synthesis

Speech output is generated using the open source text-to-speech synthesis system MARY (Modular Architecture for Research on speech sYnthesis) [39]. It features a modular design and an XML-based internal data representation. Numerous languages including English and German are supported. A number of settings allow for an adjustment of various voice features. For further details on MARY see [39].

3.4 On-line Scheduling of Multimodal Utterances

The concept underlying the multimodal production model acts on an empirically suggested assumption [24] referred to as a *segmentation hypothesis* [17]. It claims that the production of continuous speech and gesture is organized in successive segments. Each of these segments represents a single idea unit referred to as a *chunk* of speech-gesture production. A chunk, in turn, consists of an intonation phrase and a co-expressive gesture phrase, concertedly conveying a prominent concept. Levelt [21] defines intonation phrases to represent units over which the phonological structure of continuous speech is organized. With respect to gestures, Kendon [14] describes gesture phrases as units of gestural movement comprising one or more subsequent phases: preparation, stroke, retraction, hold.

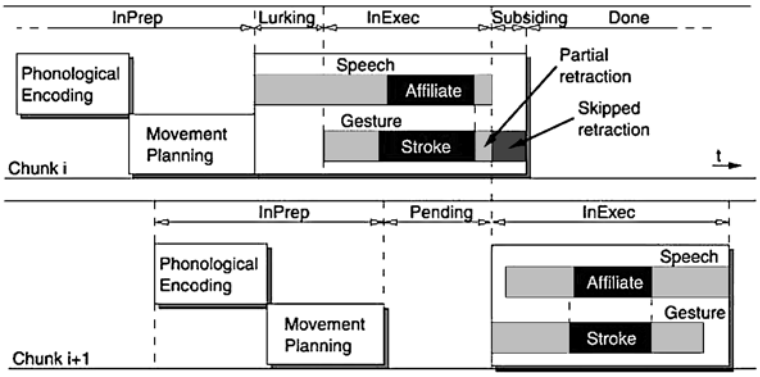

Fig. 4 Blackboards running through a sequence of processing states for incremental production of multimodal chunks [17]

Accordingly, in our model incremental production of successive coherent chunks is realized by processing each chunk on a separate 'blackboard' running through a sequence of states (Fig. 4). Timing of gestures is achieved on-line by the ACE engine as follows. Within a chunk, synchrony is generally achieved by adapting the gesture to structure and timing of speech. To do this, the ACE scheduler retrieves timing information about the synthetic speech at the millisecond level and defines the start and the end of the gesture stroke accordingly. These temporal constraints are automatically propagated down to each single gesture component. A more detailed overview of the internal planning process within ACE can be found in [17]. The second aspect of scheduling, namely, the decision to skip preparation or retraction phases, results from the interplay of motor programs at run-time. Motor programs monitor the body's current movement state and are autonomously activated to realize the planned gesture stroke as scheduled. Whenever the motor program of the following gesture takes over the control of the effectors from the preceding program, the retraction phase turns into a transition into the next gesture. Such on-line scheduling results in fluent and continuous multimodal behavior.

4 Robot Control Architecture

In an effort to enable a humanoid robot to flexibly produce speech and co-verbal gesture at run-time, a given robot control architecture needs to combine conceptual representation and planning provided by ACE with motor control primitives for speech and arm movements for the robot. This, however, poses a number of challenges including the capacity to adequately account for certain physical properties, e.g., motor states, maximum joint velocity, strict self-collision avoidance, and variation in DOF. In light of ACE being originally designed for a virtual rather than physical platform, these challenges must be met when transferring the ACE framework to the Honda humanoid robot, whose upper body comprises a torso with two 5DOF arms and 1DOF hands, as well as a 2DOF head [10].

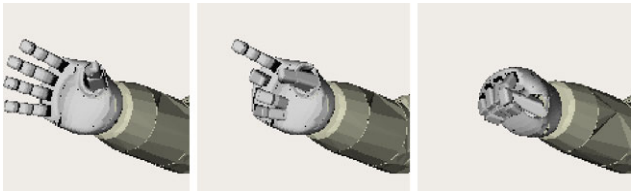

Fig. 5 Different hand shapes used for hand gesture generation on the Honda humanoid robot [36]

Although ACE provides movement descriptions in joint space to animate the body of a virtual agent, we only extract task space information when generating the corresponding robot trajectory. This allows us to circumvent the correspondence problem [29], which arises due to body dissimilarity when mapping movements from one agent's body to a different agent's body. The information obtained at the task space level includes the wrist position and orientation as well as the designated hand shape, which is forwarded to the robot motion control module to instantiate the actual robot movement. Problematically, given the small number of DOF in its hands, the humanoid robot is more limited in performing single finger movements than a virtual character. We counter this limitation by specifying three basic hand shapes that can be utilized by the robot. A variety of finger constellations derived from the ACE body model can then be mapped onto them. Hand gestures in which the hands are open or closed, and pointing gestures are directly transferable. Any hand gesture employing more than the index finger is modeled using an open hand shape. Figure 5 displays the three different hand shapes used for hand gesture generation on the Honda humanoid robot.

The problem of inverse kinematics (IK) of the arm is solved on the velocity level using the robot's whole body motion (WBM) controller framework [6]. The WBM framework allows to control all DOF of the humanoid robot based on given end-effector targets. It provides a flexible method to control upper body movement by only specifying relevant task dimensions selectively in real-time, yet while generating smooth and natural movement. Redundancies are optimized with regard to joint limit avoidance and self-collision

avoidance. For more details on WBM control for the Honda humanoid robot see [6].

After solving inverse kinematics for the internal body model provided for WBM control, the joint space description of the designated trajectory is applied to the real robot. A bi-directional interface using both efferent actuator control signals and afferent sensory feedback is used to monitor possible deviations of actual robot motor states from the kinematic body model provided by ACE. This is realized by a feedback loop that updates the internal model of the robot in the WBM controller as well as the kinematic body model coupled to ACE at a sample rate r. This process synchronizes two competing sample rates in order that successful integration can ensue: firstly, that of the ACE engine, and secondly, that of the WBM software controlling the robot. For this purpose, a number of alternative mapping rates could be employed:

1. sampling only at target positions: ACE sends only the end positions or orientations of movement segments and delegates the robot movement generation entirely to the robot's WBM controller;
2. sampling at each n-th frame: ACE sends control parameters at a fixed rate to the robot's WBM controller;
3. adaptive sampling rate: ACE "tethers" WBM using different sampling rates, ranging from one sample per frame to taking only the end positions, depending on the complexity of the trajectory.

If the trajectory is linear, then we can expect that strategy 1 above would serve as the best mechanism since only distance information would likely be required. If, on the other hand, the trajectory is complex, we can expect that strategy 2 would be optimal, since a sequence of small movement vectors would likely be required to guide the robot controller. If, however, the gesture is formed from different types of sub-movements as possible in our framework, e.g., a linear trajectory for gesture preparation with a curved trajectory for the stroke, we can expect that the combined approach of strategy 3 using an adaptive sampling rate would become optimal.

In our current set-up, we employ the second method with a maximal sampling rate, i.e., each successive frame of the movement trajectory is sampled and transmitted to the robot controller ($n = 1$). Given a frame rate of 20 frames per second (flexibly adjustable with ACE), this can result in a large number of sample points which, in turn, ensures that the robot closely follows the possibly complex trajectory planned by ACE. Results presented in the following section were obtained with this method. Alternatively, using the third strategy would allow for adjusting the sampling rate depending on the trajectory's complexity, which may well vary from simple straight movements (e.g., for gesture preparation) to complex curved shapes for the gesture stroke phase. Whether or not this strategy leads to improved results for the generation of robot gesture in combination with ACE is a point of future investigation.

A main advantage of our approach to robot control is the trajectory formulation in terms of effector targets and their respective orientations in task space. On this basis, it is fairly straightforward to derive a joint space description for the Honda humanoid robot by using the standard WBM controller. Alternatively, joint angle values could be extracted from ACE and directly mapped onto the robot body model. However, being a virtual agent application, ACE does not entirely account for physical restrictions such as collision avoidance, which may lead to joint states that are not feasible on the robot. Therefore, by solving IK using the robot's internally implemented WBM controller, we ensure a safer generation of robot posture. Furthermore, studies in which participants' gaze was eye-tracked while observing hand and arm movements provide evidence that humans mostly track the hand or end-point, even if the movement is performed with the entire arm [23]. Thus, the form and meaning of a gesture can be conveyed even with a deviation from original joint angles.

Having implemented an interface that couples ACE with the perceptuo-motor system of the Honda robot, the control architecture outlined in Fig. 6 is now used as the underlying action generation framework for the humanoid robot. It combines conceptual representation and planning with motor control primitives for speech as well as hand and arm movements of a physical robot body. Further details of the implementation are presented in [35] and [36].

5 Technical Results

Results were produced in a feedforward manner whereby commands indicating the wrist position and hand orientation of the ACE body model were transmitted in real-time to the robot at a sample rate of 20 frames per second. Figure 7 illustrates the multimodal output generated in our current framework using the MURML utterance presented in Fig. 3. The robot is shown next to a panel which displays the current state of the internal robot body model and ACE kinematic body model, respectively, at each time step. In addition, speech output is transcribed to illustrate the words spanning different segments of the gesture movement sequence, indicating temporal synchrony achieved between the two modalities. It is revealed that the physical robot is able to perform a generated gesture fairly accurately but with some inertial delay compared to the internal ACE model. This observation is supported by Fig. 8, in which each dimension of the wrist position for the ACE body model and the robot is plotted against time. Further results illustrating the difference in motion speed between the two platforms as observed during the performance of various gestures are presented in [35].

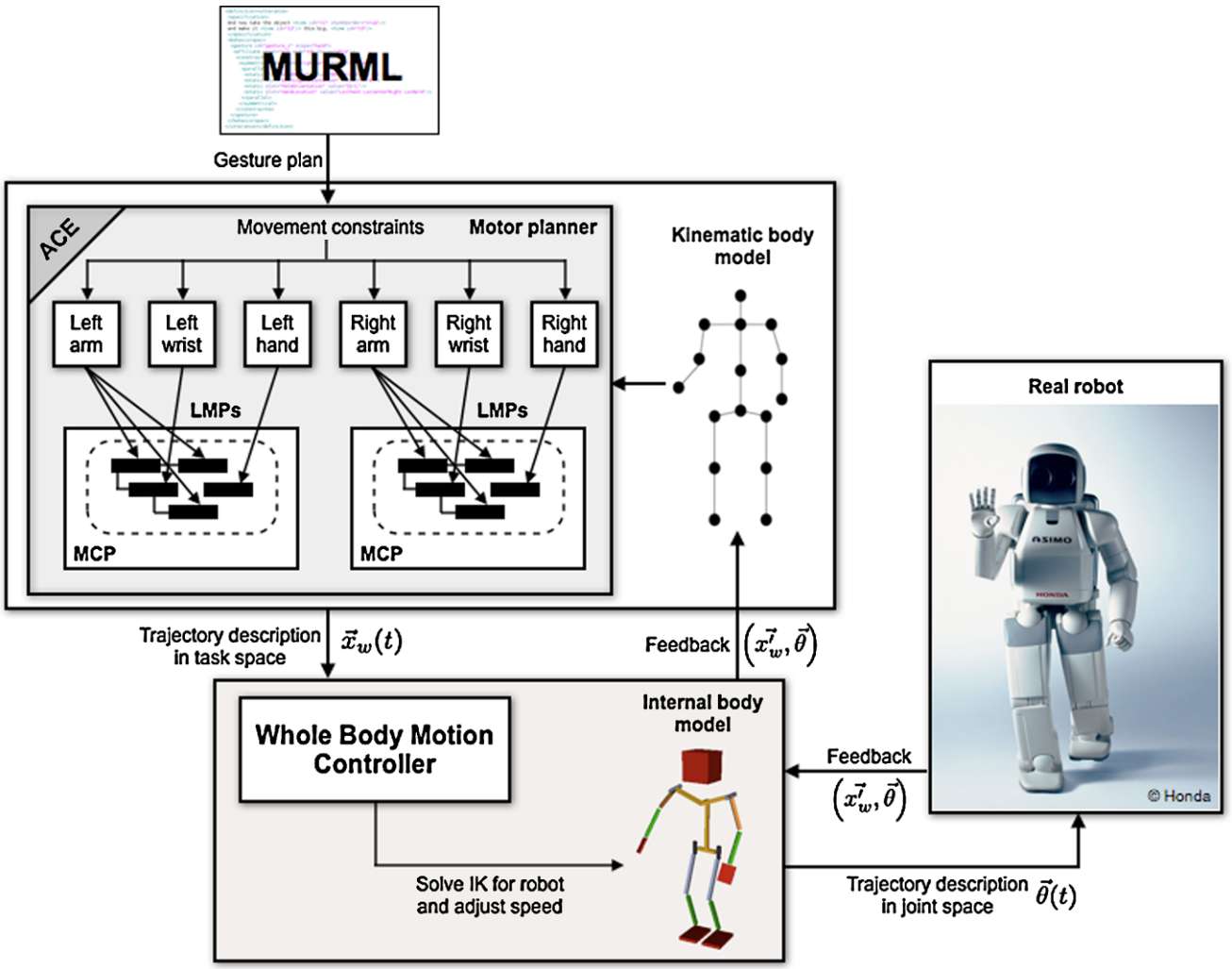

Fig. 6 Robot control architecture for the realization of multimodal behavior

Despite the general limitation in motion speed, these findings substantiate the feasibility of the proposed approach. Arbitrary MURML-based speech-gesture representations—as well as key-frame animation descriptions of gestures, optionally derived from human motion capturing data—can be realized using the current framework. Extensive tests with multiple various gesture representations (including both one-armed and two-armed movements) performed on the robot further revealed that neglecting joint angle information as generated in ACE does not impair the overall shape of a gesture. Hence, controlling the robot via task space commands turns out to be an adequate and safe way to generate arm movements for the robot.

Although Fig. 7 suggests acceptable temporal synchrony between both output modalities, synchronization of speech and gesture does not yet appear to be optimal. Tests using long sentences in speech as well as utterances with the speech affiliate situated at the beginning of the sentence revealed that movement generation tends to lag behind spoken language output. Consequently, we need to explore ways to handle the difference in time required by the robot's physically constrained body in comparison to the kinematic body model in ACE. Our idea for future work is to tackle this challenge by extending the cross-modal adaptation mechanisms provided by ACE with a more flexible multimodal utterance scheduler. This will allow for a finer mutual adaptation between robot gesture and speech. In the current implementation, the ACE engine achieves synchrony within a chunk mainly by gesture adaptation to structure and timing of speech, obtaining absolute gesture time information at the phoneme level. Improved synchronization requires the incorporation of a forward model to predict the estimated time needed by the robot for gesture preparation. Additionally, predicted values must be controlled at run-time and, if necessary, adjusted based on constantly updated feedback information on the robot state.

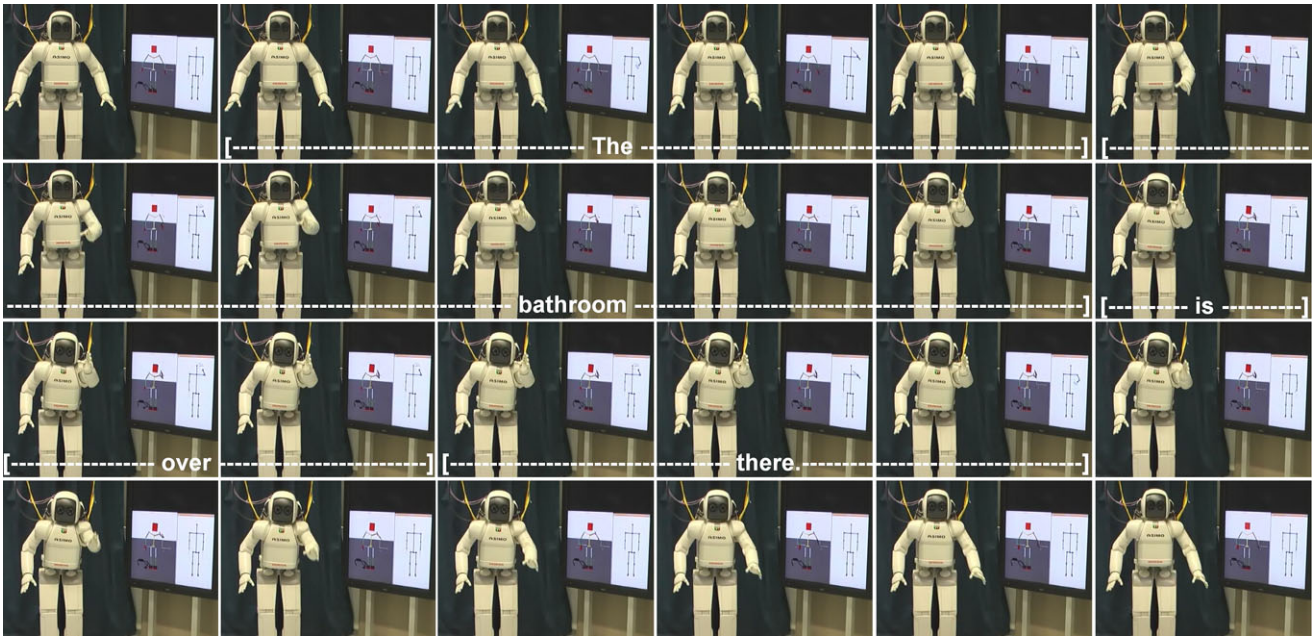

Fig. 7 Example of a multimodal utterance realized with the current framework from the specification given in Fig. 3; for comparison, the physical robot, internal robot body model, and the kinematic ACE body model are shown (*left to right*, *top-down*, sampled every four frames (0.16 s)) [36]

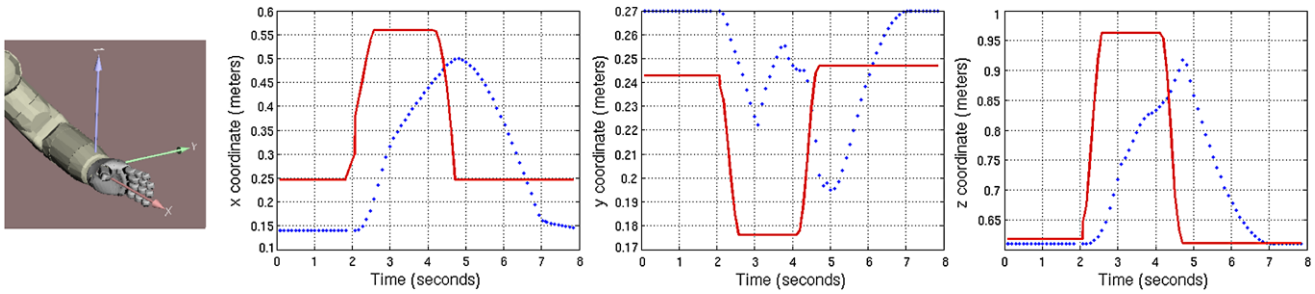

Fig. 8 Plots of x-, y- and z-coordinate respectively of the wrist positions of the ACE body model (*solid*) and the physical robot (*dotted*) during gesture execution [36]

6 Empirical Evaluation: Unimodal Versus Multimodal Robot Behavior in HRI

In order to gain a deeper understanding of how communicative robot gesture may impact and shape human experience in human-robot interaction (HRI), we conducted a between-subjects experimental study using the Honda humanoid robot. For this purpose, we designed a suitable scenario for gesture-based HRI and identified benchmarks to empirically evaluate the developed framework. The study scenario comprised a joint task that was to be performed by a human participant in collaboration with the humanoid robot. Our main motivation for choosing a task-based interaction was to realize a largely controllable yet meaningful interaction which would allow for a measurable comparison of participants' reported experiences. In the given task, the robot referred to various objects by utilizing either unimodal (speech only) or multimodal (speech and gesture) utterances, based on which the participant was expected to perceive, interpret and perform an according action.

6.1 Hypothesis

Based on findings resulting from gesture research in human-human as well as human-agent interaction we developed the following hypothesis for gesture-based human-robot interaction:

Participants who are presented with multimodal instructions by the robot (using speech and gesture) will evaluate the robot more positively than those who are presented with unimodal information by the robot (using only speech).

6.2 Experimental Design

The experiment was set in a kitchen environment in which the humanoid played the role of a household robot. Partici-

pants were told that they were helping a friend move house and were tasked with emptying a cardboard box of kitchen items, each of which had to be placed in its designated location. The box contained nine kitchen items whose storage placement is not typically known a priori (unlike plates, e.g., which are usually piled on top of each other). Specifically, they comprised a thermos flask, a sieve, a ladle, a vase, an eggcup, two differently shaped chopping boards and two differently sized bowls. The cardboard box containing the kitchen items used in the experiment is displayed in Fig. 9.

Fig. 9 Cardboard box containing kitchen items used in the experimental study

The objects were to be removed from the box and arranged in a pair of kitchen cupboards (upper and lower cupboard with two drawers). For this, the participant was allowed to move freely in the area in front of the robot, typically walking between the cardboard box with items and the kitchen cupboards. Given the participant's non-familiarity with the friend's kitchen environment, the robot was made to assist the human with the task by providing information on where each item belongs. A table situated beside the kitchen cupboard was provided for the case that the participant did not understand where the item had to be placed. A sketch of the experimental set-up is shown in Fig. 10.

Conditions We manipulated the robot's non-verbal behavior in three experimental conditions:

- In *Condition 1*, the *unimodal (speech-only)* condition, the robot presented the participant solely with a set of nine verbal instructions to explain where each object should be placed. The robot did not move its body during the whole interaction; no gesture or gaze behaviors were displayed.
- In *Condition 2*, the *congruent multimodal (speech-gesture)* condition, the robot presented the participant with the identical set of nine verbal instructions used in condition 1. In addition, they were accompanied by a total of 21 corresponding gestures explaining where each object should be placed. Speech and gesture were semantically matching, e.g., the robot said "put it up there" and pointed up. Simple gaze behavior supporting hand and arm gestures (e.g., looking right when pointing right) was displayed during the interaction.
- In *Condition 3*, the *incongruent multimodal (speech-gesture)* condition, the robot presented the participant

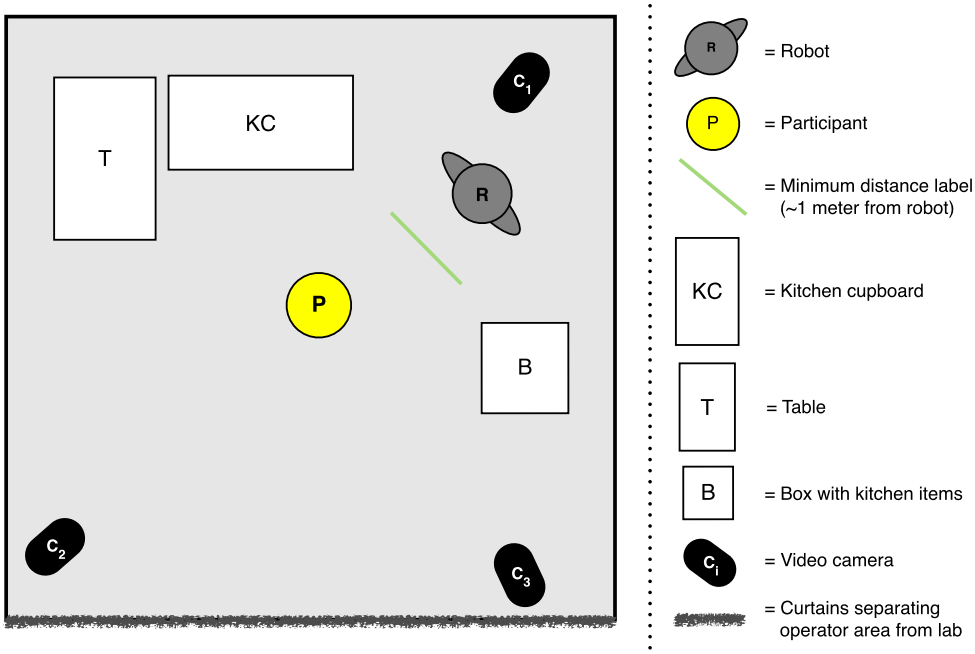

Fig. 10 Sketch of the experimental set-up in the lab

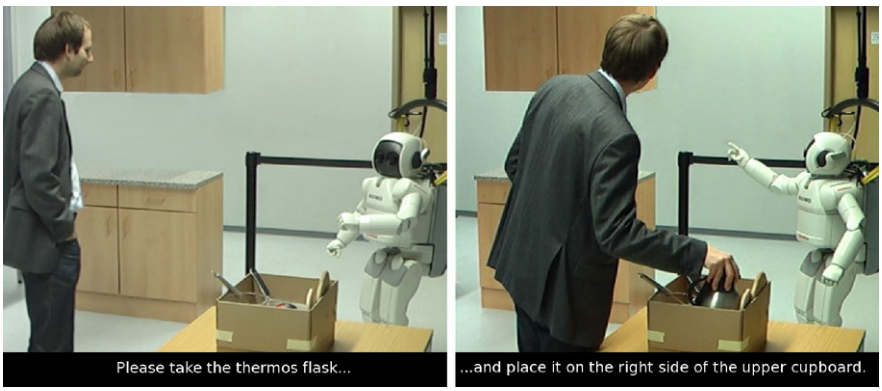

Fig. 11 Example of a multimodal *two-chunk utterance* delivered by the robot during interaction

Fig. 12 Example of a multimodal *three-chunk utterance* delivered by the robot during interaction. Three different types of gesture are used (*left to right*): *iconic gesture* illustrating the shape of the vase; *pantomimic gesture* conveying the act of opening the cupboard; *deictic gesture* pointing at designated position [37]

with the identical set of nine verbal instructions used in condition 1. Again, in addition, they were accompanied by a total of 21 gestures, out of which ten gestures (47.6%) semantically matched the verbal instruction, while the remaining eleven gestures (52.4%) were semantically non-matching, e.g., the robot occasionally said "put it up there" but pointed downwards. The reason for combining semantically non-matching gestures with matching ones in this condition was to avoid a complete loss of the robot's credibility after a few utterances. Simple gaze behavior supporting hand and arm gestures (e.g., looking right when pointing right) was displayed during the interaction.

Verbal Utterances In order to keep the task solvable under all three conditions, we decided to design the spoken utterances in a self-sufficient way, i.e., the gestures used in the multimodal conditions contained illustrative information that was not indispensable to solving the task. Each instruction presented by the robot typically consisted of two or three continuously connected *utterance chunks*. Based on the definition provided in [17], each *chunk* refers to a single idea unit represented by an intonation phrase and, optionally in a multimodal utterance, by an additional co-expressive gesture phrase. The verbal utterance chunks were based on the following syntax:

– *Two-chunk utterance*:
  `<Please take the [object]>`
  `<and place it [position+location].>`
  Example: *Please take the thermos flask and place it on the right side of the upper cupboard.*

– *Three-chunk utterance:*
  `<Please take the [object],>`
  `<then open the [location]>`
  `<and place it [position].>`
  Example: *Please take the eggcup, then open the right drawer and place it inside.*

Examples of a multimodal two-chunk and a three-chunk utterance delivered by the robot are illustrated in Figs. 11 and 12 respectively.

Gestures In the multimodal conditions, the robot used three different types of gesture along with speech to indicate the designated placement of each item:

- *Deictic gestures*, e.g., to indicate positions and locations
- *Iconic gestures*, e.g., to illustrate the shape/size of objects

Table 1 Dependent measures used to evaluate the *quality of presentation*

Measure	Questionnaire item	Scale
Gesture quantity	"The amount of gestures performed by the robot were…"	1 = too few, 5 = too many
Gesture speed	"The execution of gestures was…"	1 = too slow, 5 = too fast
Gesture fluidity	"The execution of hand and arm movements was fluid."	1 = not appropriate, 5 = very appropriate
Speech-gesture content	"The robot's speech and gesture were semantically matching (content)."	1 = not appropriate, 5 = very appropriate
Speech-gesture timing	"The robot's speech and gesture were well synchronized (timing)."	1 = not appropriate, 5 = very appropriate
Naturalness	"The combined use of speech and gesture appeared…"	1 = artificial, 5 = natural

- *Pantomimic gestures*, e.g., hand movement using a ladle or opening cupboard doors.

Examples of the three gesture types are displayed in Fig. 12.

Robot Control and Behavior During the study, the Honda humanoid robot was partly controlled using a Wizard-of-Oz technique to ensure minimal variability in the experimental procedure. The experiment room was partitioned with a curtain such that the robot and kitchen environment were located at one end and the wizard operating the control computer was located at the other end, outside the participant's field of view.

The robot's speech was identical across conditions. It was generated using the text-to-speech system MARY [39] set to a neutral voice. To avoid uncertainties, neither speech recognition nor active vision were used during the experiment. Instead, the experimenter initiated the robot's interaction behavior from a fixed sequence of pre-determined utterances. Once triggered, a given utterance was generated autonomously at run-time. The ordering and generation of this sequence remained identical across conditions and experimental runs.

The robot delivered each two-chunk or three-chunk instructional utterance as a singular one-shot expression without any significant breaks in the delivery process. Successive chunks indicating object, position and location were delivered contiguously in the manner of natural speech. Moreover, in the co-verbal gesture conditions, gestures became confluent with the utterance process. Participants were instructed to indicate when they had finished placing an item and were ready for the following item by saying "next".

6.3 Dependent Measures

Based on the participants' answers to a post-experiment questionnaire using a five-point Likert scale for each item, we investigated two main aspects of the reported interaction experience: firstly, the perceived *quality of presentation* was measured using six questionnaire items; secondly, the *perception of the robot* was assessed based on eight characteristics covered by additional questionnaire items. Tables 1 and 2 give an overview of the dependent measures, questionnaire items, and scales used, respectively, for each evaluation category.

Table 2 Dependent measures used to evaluate the *perception of the robot*

Measure	Questionnaire item	Scale
Sympathetic Competent Lively Active Engaged Friendly Communicative Fun-loving	"Please assess to which extent the following characteristics apply to the robot: […]"	1 = not appropriate, 5 = very appropriate

6.4 Participation

A total of 60 people (30 female, 30 male) participated in the experiment, ranging in age from 20 to 62 years ($M = 31.12$, $SD = 10.21$). All participants were native German speakers who were recruited at Bielefeld University and had never before participated in a study involving robots. Based on five-point Likert scale ratings (1 = very little, 5 = very much), participants were identified as having negligible experience with robots ($M = 1.22$, $SD = 0.45$), while their computer and technology know-how was moderate ($M = 3.72$, $SD = 0.90$). Participants were randomly assigned to one of the three different experimental conditions (i.e., 20 participants per condition), while maintaining gender- and age-balanced distributions.

6.5 Experimental Procedure

Participants were first given a brief written scenario and task description to read outside the experimental lab. They were then brought into the experiment room where the experimenter verbally reiterated the task description to ensure the participants' familiarity. Participants were given the opportunity to ask any clarifying questions. The experimenter then

Table 3 Mean values for the rating of presentation quality in the three conditions (standard deviations in parentheses)

	Cond. 1 Unimodal	Cond. 2 Congruent	Cond. 3 Incongruent
Gesture quantity	1.90 (.99)	2.80 (.62)	3.00 (.56)
Gesture speed		2.85 (.37)	2.95 (.22)
Gesture fluidity		3.25 (.97)	3.95 (1.05)
Speech-gesture content		3.65 (1.04)	3.30 (1.26)
Speech-gesture timing		3.90 (.79)	4.05 (1.10)
Naturalness		3.20 (1.06)	3.30 (1.13)

left the participant to begin the interaction with the robot. At the beginning of the experiment, the robot greeted the participant and gave a verbal introduction to the task. It then presented the participant with individual utterances as described in the experimental design, each of which was triggered by the experimenter sitting at a control terminal. The participant attempted to follow the uttered instructions by placing each item into its designated location. At the end of the interaction, the robot thanked the participant for helping and bid them farewell.

In the unimodal (speech-only) condition all utterances including the greeting and farewell were presented verbally; in the multimodal (speech-gesture) conditions, all utterances including the greeting and farewell were accompanied by co-verbal gestures.

After completing the task, participants filled out a post-experiment questionnaire that recorded their demographic background and, based on a five-point Likert scale, measured their affective state, evaluation of the task and interaction, and perception of the robot. Upon completion of the questionnaire, the participants were de-briefed and received a chocolate bar as a thank-you. The questionnaire data was collated and analyzed, the results are presented and discussed in the following.

6.6 Results and Discussion

Questionnaire data was analyzed regarding the effect of experimental conditions on assessment of presentation quality and robot perception.

Quality of Presentation We investigated the perceived quality of presentation with regard to gesture, speech, and content. Mean values and standard deviations are summarized in Table 3. Note that for condition 1 (*unimodal*) only gesture quantity was measured, since participants in this condition were not presented with any non-verbal behavior by the robot and thus could not rate the quality of the robot's gestures.

With regard to gesture quantity, the overall mean value for the two gesture conditions was $M = 2.90$ ($SD = 0.59$). This means, participants were quite satisfied with the gesture rate. For the unimodal condition, participants rated gesture quantity as rather low ($M = 1.90$, $SD = 0.99$), which can be attributed to the lack of non-verbal behavior displayed by the robot.

Table 4 Mean values for the rating of robot perception in the three conditions based on a 5-point Likert scale (standard deviations in parentheses); $^{+} = p \leq 0.10$, $^{*} = p \leq 0.05$, $^{**} = p \leq 0.01$, $^{***} = p \leq 0.001$

	Cond. 1 Unimodal	Cond. 2 Congruent	Cond. 3 Incongruent
Sympathetic	3.60 (1.05)	4.20 (.95)$^{+}$	4.15 (1.09)
Competent	3.85 (.93)	4.26 (.87)	3.75 (1.16)
Lively	2.52 (.84)	3.12 (.97)*	3.32 (.76)**
Active	2.35 (.88)	3.20 (1.11)**	3.45 (.76)***
Engaged	3.25 (1.29)	3.60 (1.35)	4.15 (.88)*
Friendly	4.15 (1.04)	4.35 (1.31)	4.60 (.68)
Communicative	3.00 (1.08)	3.15 (1.31)	3.60 (1.05)$^{+}$
Fun-loving	1.95 (.83)	2.65 (1.23)*	2.70 (1.30)*

For the multimodal conditions, gesture quality was further measured based on five attributes (overall mean value and standard deviation for the two gesture conditions in parentheses): gesture speed ($M = 2.90$, $SD = 0.30$), gesture fluidity ($M = 3.60$, $SD = 1.06$), semantic matching of speech and gesture ($M = 3.48$, $SD = 1.14$), temporal matching of speech and gesture ($M = 3.97$, $SD = 0.95$), and naturalness ($M = 3.25$, $SD = 1.08$). In both gesture conditions, the five quality attributes were rated with mean values between 2.8 and 4.1 on five-point Likert-scales, indicating that participants were generally satisfied with the quality of gestures performed by the robot.

Perception of the Robot We assessed how the humanoid robot was perceived by participants using eight characteristics. To test our hypothesis we conducted independent-samples t-tests with 95% confidence intervals as follows: first, we compared questionnaire data from condition 1 with condition 2 (*unimodal vs. congruent multimodal*); second, we compared data from condition 1 with condition 3 (*unimodal vs. incongruent multimodal*). Mean values for the robot's perception scales in the three different conditions are listed together with their standard deviation values in Table 4 and are visualized in Fig 13. Items showing statistically significant effects in multimodal gesture conditions compared to the unimodal speech-only condition are marked with asterisks (*).

On average, all qualities were rated higher, i.e., more positively, in the multimodal gesture conditions, with a lower mean value for 'competent' in the incongruent speech-gesture condition being the exception. These results support

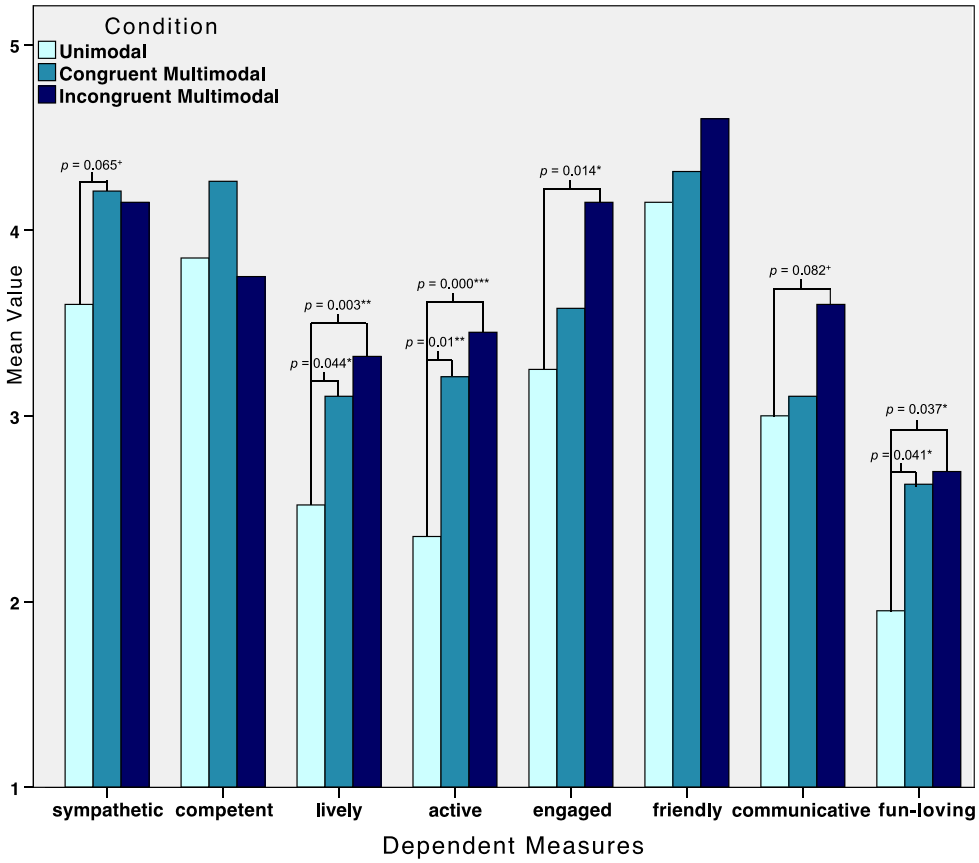

Fig. 13 Mean values of the dependent measures rating participants' perception of the robot

our hypothesis and suggest that the inclusion of gestural behavior casts the robot in a more positive light than in the speech-only condition.

Comparing condition 1 (*unimodal*) with condition 2 (*congruent multimodal*), the three characteristics 'lively' ($t(38) = -2.09$, $p = 0.044$), 'active' ($t(38) = -2.70$, $p = 0.01$) and 'fun-loving' ($t(38) = -2.12$, $p = 0.041$) are observed to be significantly higher in the congruent with-gesture condition than in the unimodal condition using speech only. In addition, a comparison of the characteristic 'sympathetic' between conditions 1 and 2 is shown to be significant at the 10% level ($t(38) = -1.90$, $p = 0.065$), with higher mean values in the congruent multimodal condition.

When comparing condition 1 (*unimodal*) with condition 3 (*incongruent multimodal*), the four characteristics 'lively' ($t(38) = -3.17$, $p = 0.003$), 'active' ($t(38) = -4.25$, $p = 0.000$), 'engaged' ($t(38) = -2.58$, $p = 0.014$) and 'fun-loving' ($t(32.16) = -2.18$, $p = 0.037$) are found to be rated significantly higher in the multimodal condition. In addition, comparing the characteristic 'communicative' between condition 1 and 3 shows a significant effect at the 10% level ($t(38) = -1.79$, $p = 0.082$), with higher mean values in the incongruent multimodal condition.

An additional comparison of data from condition 2 with condition 3 (*congruent vs. incongruent multimodal*) showed no significant effect of experimental conditions. However, with the exception of dependent measures 'sympathetic' and 'competent', our analyses indicated a trend towards higher mean values in the incongruent multimodal condition.

The significantly higher rating of 'lively' and 'active' in the two multimodal conditions can be attributed to the robot's gestural movements, since the robot appears comparatively stiff in the speech-only condition. The ratings of the characteristics 'fun-loving', 'engaged', 'sympathetic' and 'communicative' suggest that human-like non-verbal behaviors including gestures actually trigger a more positive response within the human participant. The results further reveal that even a robot that occasionally makes incorrect gestures is still more favorable than one that performs no hand and arm gestures at all. In fact, on average the robot is evaluated as more lively, active, engaged, friendly, communicative and fun-loving in the incongruent speech-gesture condition compared with the congruent condition. This suggests that a robot's non-verbal communicative behavior can even trigger a stronger positive response within the human participant when it is not 'perfect'. Overall, the results demonstrate that co-verbal gestures performed by a humanoid robot lead to an enhanced human-robot interaction experience, i.e., the robot is generally rated more positively when it displays non-verbal behaviors. These findings support our approach to endow social robots with communicative gestural behavior.

7 Conclusion and Future Work

We presented a robot control architecture which enables the Honda humanoid robot to generate gestures and synchronized speech at run-time, while not being limited to a predefined repertoire of motor actions. The present framework builds upon a speech and gesture production model for virtual human agents. Representing a sophisticated multimodal scheduler, the Articulated Communicator Engine (ACE) allows for an on-line production of flexibly planned behavior representations. Our framework combines conceptual, XML-based representation and planning with motor control primitives for speech and arm movements.

Meeting strict temporal synchrony constraints will present a main challenge to our framework in the future. Evidently, the generation of finely synchronized multimodal 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 must be produced at run-time. Currently, the ACE engine achieves synchrony mainly by gesture adaptation to structure and timing of speech, obtaining absolute time information at phoneme level. To tackle this new dimension of requirements, however, the cross-modal adaptation mechanisms applied in ACE have to be extended to allow for a finer mutual adaptation between robot gesture and speech. For this, afferent feedback provided by our robot control architecture needs to be integrated into a more sophisticated scheduler.

In order to investigate how humans perceive representational hand and arm gestures performed by the robot during a task-related interaction, we evaluated our technical framework in an experimental study using the Honda humanoid robot. Our findings reveal that the perception and evaluation of the robot is rated more positively when it displays non-verbal behaviors in the form of co-verbal gestures along with speech. This is also true for hand and arm gestures that do not semantically match the information content conveyed via speech, suggesting that a humanoid robot that generates gestures—even if in part they are semantically 'incorrect'—is still more favorable than one that performs no gestures at all. In fact, on average the robot is evaluated as more lively, active, engaged, friendly, communicative and fun-loving in the incongruent speech-gesture condition compared with the congruent condition. This suggests that the robot's non-verbal communicative behavior triggers a stronger positive response within the human participant when it is not 'perfect' and thus potentially less predictable. These implications should be further elucidated in subsequent studies to point out the direction for future social robotics research that is dedicated to the design of acceptable behaviors for artificial communicators.

In the study presented, the robot's gaze behavior was modeled in a very simplistic way in the multimodal conditions; robot gaze in the speech-only condition was static throughout the interaction. These design choices were made on purpose to direct the participants' attention to the hand and arm movements performed by the robot in the speech-gesture conditions. As a consequence, however, the robot's gazing behavior did not appear very natural during the interaction, since the robot did not follow the human interaction partner with its gaze. In future studies, it will be desirable to investigate the impact and interaction of the robot's gaze in combination with gestural hand and arm movements.

Despite some limitations, our results do nonetheless suggest that a robot presenting social cues in the form of co-verbal hand and arm gestures, as generated with our framework, is perceived in a more positive way than a robot whose sole means of communication is limited to a single modality, namely speech. These findings contribute to an advancement in human-robot interaction and give new insights into human perception and understanding of gestural machine behaviors. Specifically, they shed light on how humans perceive and interpret utterances in relation to different communication modalities. Our findings suggest that human-like behavior in a humanoid robot has a positive impact on the way humans perceive the robot in an interaction. Ultimately, these results will allow us to design and build better artificial communicators in the future.

Acknowledgements The work described was supported by the Honda Research Institute Europe.

References

1. Bennewitz M, Faber F, Joho D, Behnke S (2007) Fritz—a humanoid communication robot. In: RO-MAN 07: Proc of the 16th IEEE international symposium on robot and human interactive communication
2. Bergmann K, Kopp S, Eyssel F (2010) Individualized gesturing outperforms average gesturing—evaluating gesture production in virtual humans. In: Proceedings of the 10th conference on intelligent virtual agents. Springer, Berlin, pp 104–117
3. Billard A, Calinon S, Dillmann R, Schaal S (2008) Robot programming by demonstration. In: Siciliano B, Khatib O (eds) Handbook of robotics. Springer, New York, pp 1371–1394
4. Calinon S, Billard A (2007) Learning of gestures by imitation in a humanoid robot. In: Dautenhahn K, Nehaniv C (eds) Imitation and social learning in robots, humans and animals: behavioural, social and communicative dimensions. Cambridge University Press, Cambridge, pp 153–177
5. Cassell J, Bickmore T, Campbell L, Vilhjálmsson H, Yan H (2000) Human conversation as a system framework: designing embodied conversational agents. In: Embodied conversational agents. MIT Press, Cambridge, pp 29–63
6. Gienger M, Janßen H, Goerick S (2005) Task-oriented whole body motion for humanoid robots. In: Proceedings of the IEEE-RAS international conference on humanoid robots, Tsukuba, Japan

7. Goldin-Meadow S (1999) The role of gesture in communication and thinking. Trends Cogn Sci 3:419–429
8. Gorostiza J, Barber R, Khamis A, Malfaz M, Pacheco R, Rivas R, Corrales A, Delgado E, Salichs M (2006) 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
9. Hartmann B, Mancini M, Pelachaud C (2005) Implementing expressive gesture synthesis for embodied conversational agents. In: Gesture in human-computer interaction and simulation
10. Honda Motor Co, L (2000) The Honda humanoid robot Asimo, year 2000 model. http://world.honda.com/ASIMO/technology/spec.html
11. Hostetter AB (2011) When do gestures communicate? A meta-analysis. Psychol Bull 137(2):297–315
12. Hostetter AB, Alibali MW (2008) Visible embodiment: gestures as simulated action. Psychon Bull Rev 15(3):495–514
13. Itoh K, Matsumoto H, Zecca M, Takanobu H, Roccella S, Carrozza M, Dario P, Takanishi A (2004) Various emotional expressions with emotion expression humanoid robot we-4rii. In: Proceedings of the 1st IEEE technical exhibition based conference on robotics and automation proceedings TExCRA 2004, pp 35–36
14. Kendon A (1980) Gesticulation and speech: two aspects of the process of utterance. In: The relationship of verbal and non-verbal communication, pp 207–227
15. Kendon A (2004) Gesture: visible action as utterance. Gesture 6(1):119–144
16. Kopp S, Wachsmuth I (2000) A knowledge-based approach for lifelike gesture animation. In: Horn W (ed) ECAI 2000—Proceedings of the 14th European conference on artificial intelligence. IOS Press, Amsterdam, pp 663–667
17. Kopp S, Wachsmuth I (2004) Synthesizing multimodal utterances for conversational agents. Comput Animat Virtual Worlds 15(1):39–52
18. Kopp S, Bergmann K, Wachsmuth I (2008) Multimodal communication from multimodal thinking—towards an integrated model of speech and gesture production. Semant Comput 2(1):115–136
19. Kramer N, Simons N, Kopp S (2007) The effects of an embodied conversational agent's nonverbal behavior on user's evaluation and behavioral mimicry. In: Proc of intelligent virtual agents (IVA 2007), vol 4722. Springer, Berlin, pp 238–251
20. Kranstedt A, Kopp S, Wachsmuth I (2002) 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
21. Levelt W (1989) Speaking. MIT Press, Cambridge
22. Macdorman K, Ishiguro H (2006) The uncanny advantage of using androids in cognitive and social science research. Interact Stud 7(3):297–337
23. Mataric MJ, Pomplun M (1998) Fixation behavior in observation and imitation of human movement. Cogn Brain Res 7(2):191–202
24. McNeill D (1992) Hand and mind: what gestures reveal about thought. University of Chicago Press, Chicago
25. McNeill D (2005) Gesture and thought. University of Chicago Press, Chicago
26. Minato T, Shimada M, Ishiguro H, Itakura S (2004) Development of an android robot for studying human-robot interaction. In: Innovations in applied artificial intelligence, pp 424–434
27. Miyashita T, Shinozawa K, Hagita N (2006) Gesture translation for heterogeneous robots. In: Proceedings of 6th IEEE-RAS international conference on humanoid robots, pp 462–467
28. Mori M (1970) The uncanny valley. Energy 7(4):33–35 (KF MacDorman and T Minato, Trans)
29. Nehaniv CL, Dautenhahn K (1998) The correspondence problem
30. Ng-Thow-Hing V, Luo P, Okita S (2010) Synchronized gesture and speech production for humanoid robots. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, pp 4617–4624
31. Niewiadomski R, Bevacqua E, Mancini M, Pelachaud C (2009) Greta: an interactive expressive ECA system. In: Proceedings of 8th int conf on autonomous agents and multiagent systems (AAMAS2009), pp 1399–1400
32. Okuno Y, Kanda T, Imai M, Ishiguro H, Hagita N (2009) Providing route directions: design of robot's utterance, gesture, and timing. In: Proceedings of the 4th ACM/IEEE international conference on human robot interaction, HRI'09. ACM, New York, pp 53–60
33. Pollard N, Hodgins J, Riley M, Atkeson C (2002) Adapting human motion for the control of a humanoid robot. In: Proceedings of international conference on robotics and automation, pp 1390–1397
34. Reiter E, Dale R (2000) Building natural language generation systems. Cambridge Univ Press, Cambridge
35. Salem M, Kopp S, Wachsmuth I, Joublin F (2010) Generating robot gesture using a virtual agent framework. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems
36. Salem M, Kopp S, Wachsmuth I, Joublin F (2010) Towards an integrated model of speech and gesture production for multi-modal robot behavior. In: Proceedings of the IEEE international symposium on robot and human interactive communication
37. Salem M, Rohlfing K, Kopp S, Joublin F (2011) A friendly gesture: investigating the effect of multimodal robot behavior in human-robot interaction. In: Proceedings of the IEEE international symposium on robot and human interactive communication
38. Saygin A, Chaminade T, Ishiguro H, Driver J, Frith C (2011) The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. Soc Cogn Affect Neurosci. doi:10.1093/scan/nsr025
39. Schröder M, Trouvain J (2003) The German text-to-speech synthesis system MARY: a tool for research, development and teaching. Int J Speech Technol 6:365–377
40. Sidner C, Lee C, Lesh N (2003) The role of dialog in human robot interaction. In: International workshop on language understanding and agents for real world interaction
41. Sugiyama O, Kanda T, Imai M, Ishiguro H, Hagita N (2007) Natural deictic communication with humanoid robots. In: Proceedings of the IEEE international conference on intelligent robots and systems, pp 1441–1448
42. Wachsmuth I, Kopp S (2002) Lifelike gesture synthesis and timing for conversational agents. In: Wachsmuth I, Sowa T (eds) Gesture and sign language in human-computer interaction. LNAI, vol 2298. Springer, Berlin, pp 120–133

Maha Salem is a Ph.D. student at the Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University, Germany. Her research interests lie in the field of social human-robot interaction with special focus on multimodal interaction and non-verbal expressiveness in humanoid robots.

Stefan Kopp is director of the Sociable Agents Group of the Center of Excellence 'Cognitive Interaction Technology' (CITEC) at Bielefeld University, Germany. He is heading projects in the Collaborative Research Center 'Alignment in Communication' (CRC 673) as well as the CoR-Lab.

Ipke Wachsmuth is a professor of Artificial Intelligence at Bielefeld University, Germany. He is co-initiator and coordinator of the Collaborative Research Center 'Alignment in Communication' (CRC 673) and adjoint member of the CoR-Lab.

Katharina Rohlfing received her Ph.D. in Linguistics from Bielefeld University, Germany, in 2002. Since 2008, she is head of the Emergentist Semantics Group within the Center of Excellence 'Cognitive Interaction Technology' (CITEC), Bielefeld University, Germany. She is interested in multimodal learning processes.

Frank Joublin is Principal Scientist at the Honda Research Institute Europe, Offenbach, Germany, and since 2008 a board member of the CoR-Lab Graduate School, Bielefeld University, Germany.