Generating Multi-Modal Robot Behavior Based on a Virtual Agent Framework

Maha Salem, Stefan Kopp, Ipke Wachsmuth, Frank Joublin

***Abstract*— One of the crucial steps in the attempt to build sociable, communicative humanoid robots is to endow them with expressive non-verbal behaviors along with speech. One such behavior is gesture, frequently used by human speakers to emphasize, supplement, or even complement what they express in speech. The generation of speech-accompanying robot gesture together with an evaluation of the effects of this multi-modal behavior is still largely unexplored. We present an approach to systematically address this issue by enabling the humanoid Honda robot to flexibly produce synthetic speech and expressive gesture from conceptual representations at run-time, while not being limited to a predefined repertoire of motor actions in this. Since this research challenge has already been tackled in various ways within the domain of virtual conversational agents, we build upon experiences gained with speech-gesture production models for virtual humans.**

I. INTRODUCTION

Humanoid robot companions that are intended to engage in natural and fluent human-robot interaction in rich environmental settings must be able to produce speech-accompanying, non-verbal behaviors. Forming an integral part of human communication, hand and arm gestures are primary candidates for extending the communicative capabilities of social robots. This, however, poses a number of research challenges, especially with regard to a motor control for arbitrary, expressive hand-arm movement and its coordination with other interaction modalities such as speech. The generation of co-verbal gestures for artificial humanoid bodies demands a high degree of control and flexibility concerning shape and time properties of the gesture, while ensuring a natural appearance of the movement. Ideally, if such non-verbal behaviors are to be realized, they have to be derived from conceptual, to-be-communicated information.

Since the challenge of multi-modal behavior realization has already been explored in various ways within the domain of virtual conversational agents, our approach builds upon the experiences gained from the development of a speech and gesture production model used for the virtual human *Max* [2]. 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. Employing it as an underlying action generation architecture for the Honda humanoid robot, ACE draws upon a tight, bi-directional coupling of the robots perceptuo-motor system with multi-modal scheduling via both efferent control signals and afferent feedback.

The work described is supported by the Honda Research Institute Europe.

M. Salem is at the Research Institute for Cognition and Robotics, Bielefeld, Germany msalem@cor-lab.uni-bielefeld.de

S. Kopp is at the Sociable Agents Group, Bielefeld University, Germany skopp@techfak.uni-bielefeld.de

I. Wachsmuth is at the Artificial Intelligence Group, Bielefeld University, Germany ipke@techfak.uni-bielefeld.de

F. Joublin is at the Honda Research Institute Europe, Offenbach, Germany frank.joublin@honda-ri.de

II. SPEECH-GESTURE PRODUCTION MODEL FOR A HUMANOID ROBOT

Within the ACE framework, there are two different ways to describe gesture representations using the XML-based Multi-modal Utterance Representation Markup Language (MURML [3]). Firstly, verbal utterances in combination with co-verbal gestures can be specified with feature-based descriptions. In such MURML utterances, the outer form features of a gesture (i.e., the posture designated for the gesture stroke) are explicitly described. Their affiliation to dedicated linguistic elements is determined by matching time identifiers. Fig. 1 illustrates an example of a feature-based MURML specification that can be used as input for speech-gesture production. Secondly, gestures can be specified as keyframe animations whereby each keyframe specifies a 'key posture', a part of the overall gesture movement pattern describing the current state of each joint. Speed information for the interpolation between every two key postures and the corresponding affiliation to parts of speech is obtained from assigned time identifiers. Keyframe animations in ACE can be either defined manually or, alternatively, derived from motion capturing data from a human demonstrator, allowing the animation of virtual agents in real-time.

A. On-line Scheduling of Multi-Modal Utterances

In a given multi-modal utterance, each intonation phrase together with a co-expressive gesture phrase represents a

```
<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>
```

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

single idea unit which is referred to as a *chunk* of speech-gesture production [2]. Incremental production of successive coherent chunks is realized by processing each chunk on a separate 'blackboard' running through a sequence of states (Fig. 2). Speech-gesture synchronization within a chunk is

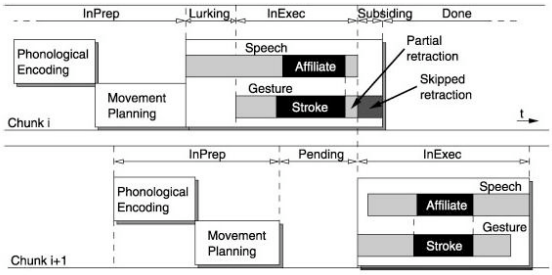

Fig. 2. Blackboards running through a sequence of processing states for incremental production of multi-modal chunks [2].

achieved on-line by the ACE engine 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 [2].

B. Speech Synthesis

Spoken utterances are generated using the open source text-to-speech synthesis system MARY (Modular Architecture for Research on speech sYnthesis) [4]. Its main features are a modular design and an XML-based internal data representation. Several languages are supported including English and German. For further details on MARY see [4].

C. Robot Control Architecture

In order to enable the humanoid robot to flexibly produce speech and co-verbal gesture at run-time, a robot control architecture is required which combines conceptual representation and planning provided by ACE with motor control primitives for speech and arm movements for the robot. This endeavor poses a number of interesting challenges including a failure to adequately account for certain physical properties – motor states, maximum velocity, strict self collision avoidance, variation in DOFs, etc. This is in light of ACE being originally designed for a virtual rather than physical platform. Hence, when transferring the ACE framework to the physical robot these challenged must be met.

Since gesture generation with ACE is based on external form features as annotated in the MURML specification, our robot control architecture suggests that arm movement trajectories are described directly in task space. The information obtained at the task space level including wrist orientation and designated hand shape is forwarded to the robot motion control module which instantiates the actual robot movement. Inverse kinematics (IK) of the arm is then solved on the velocity level using the whole body motion (WBM) controller framework [1]. The WBM framework allows to control all DOFs of the Honda humanoid robot based on given end-effector targets, providing 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 research robot see [1].

After 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 physical robot. A bi-directional interface using both efferent actuator control signals and afferent sensory feedback monitors possible deviations of actual robot motor states from the kinematic body model provided by ACE. It 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*. Fig. 3 illustrates our robot control architecture embedding the ACE framework.

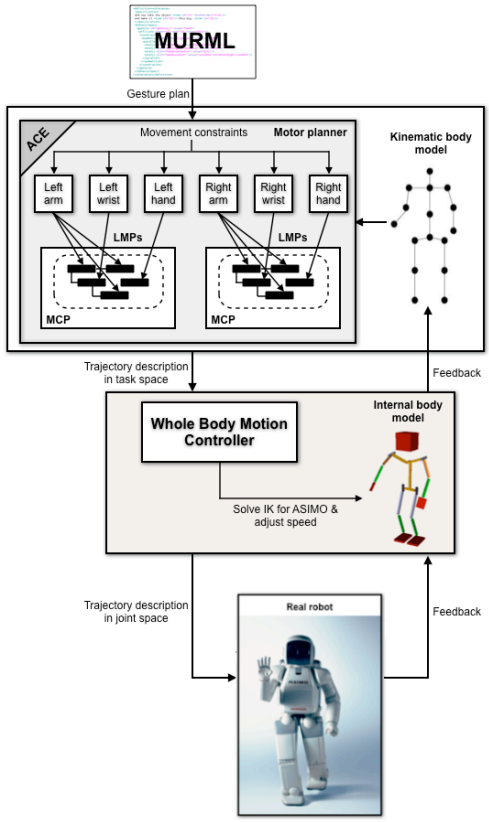

Fig. 3. Robot control architecture for the generation of gesture behavior.

III. Results and Discussion

Results were produced in a feed-forward manner whereby commands indicating the wrist position and hand orientation of the ACE body model were constantly transmitted to the robot at a sample rate of 20 frames per second. IK was solved using the provided whole body motion controller.

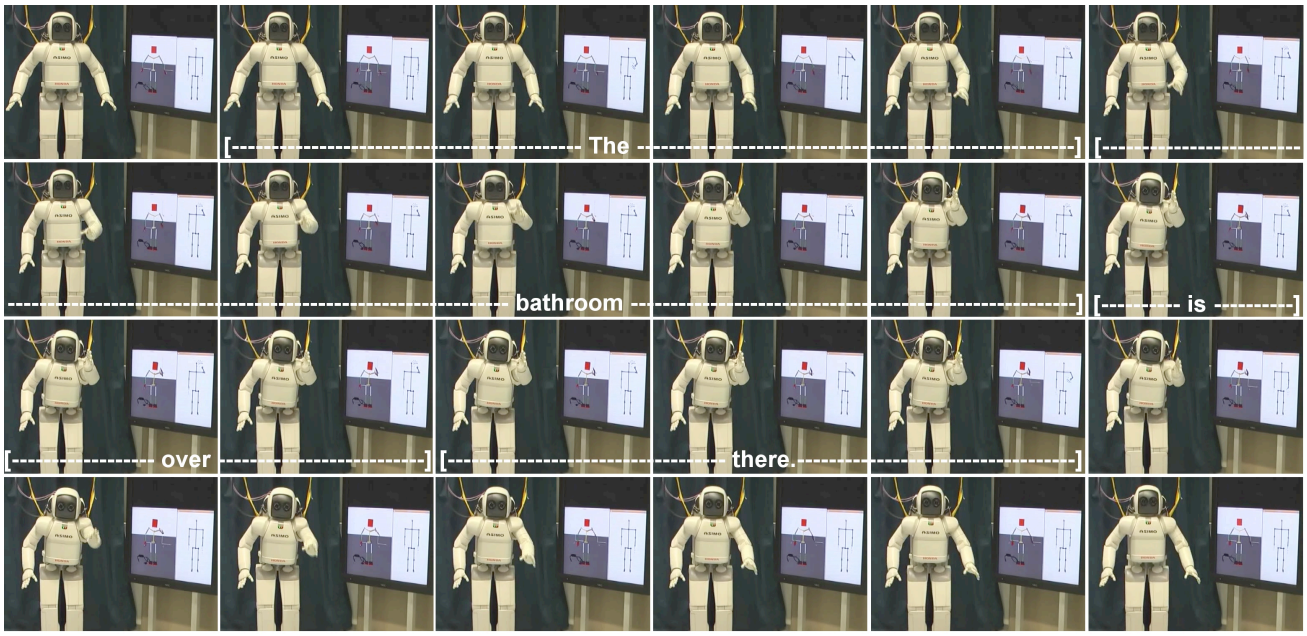

Fig. 4. Example of a multi-modal utterance realized in the current framework, allowing comparison of the physical robot, the internal robot body model and the ACE kinematic body model (left to right, top-down, sampled every four frames (0.16sec)).

Speech output was synthesized using the MARY text-to-speech system based on the multi-modal utterance scheduler in ACE. Fig. 4 illustrates the multi-modal output generated in our current framework using the MURML utterance presented in Fig. 1. 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. Despite the general limitation in motion speed, these findings substantiate the feasibility of the proposed approach. Arbitrary MURML-based speech-gesture representations can be realized using the current framework. Synchronization of speech and gesture, however, does not appear to be optimal yet. Although Fig. 4 suggests acceptable temporal synchrony between both output modalities, tests using long sentences in speech 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 is to tackle this challenge by extending the cross-modal adaptation mechanisms provided by ACE with a more flexible multi-modal utterance scheduler which will allow for a finer mutual adaptation between robot gesture and speech.

IV. Conclusion and Future Work

We presented a robot control architecture which enables the Honda humanoid research robot to generate gestures and synchronized speech at run-time. Meeting strict temporal synchrony constraints will present a main challenge to our framework in the future. Evidently, 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. 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.

Our results help to shed light on conceptual motorics in robotic agents. Essentially, they substantiate the feasibility of our approach while pointing out the direction for our future research. Once our robot control architecture has been extended to account for a finer synchronization of gesture and speech, it will be assessed in human-robot interaction studies, providing new insights into human perception and understanding of gestural machine behaviors and how these can be used to design more natural communication in robots.

References

[1] 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.

[2] S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52, 2004.

[3] 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.

[4] M. Schröder and Jürgen Trouvain. The german text-to-speech synthesis system mary: A tool for research, development and teaching. In *International Journal of Speech Technology*, pages 365–377, 2003.