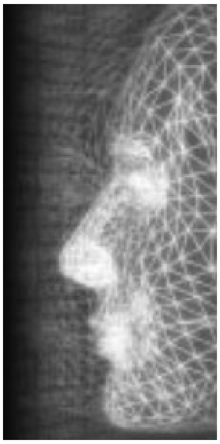
COMPUTER ANIMATION AND VIRTUAL WORLDS
Comp. Anim. Virtual Worlds 2004; **15:** 39–52 (DOI: 10.1002/cav.6)

Synthesizing multimodal utterances for conversational agents



By Stefan Kopp and Ipke Wachsmuth*

Conversational agents are supposed to combine speech with non-verbal modalities for intelligible multimodal utterances. In this paper, we focus on the generation of gesture and speech from XML-based descriptions of their overt form. An incremental production model is presented that combines the synthesis of synchronized gestural, verbal, and facial behaviors with mechanisms for linking them in fluent utterances with natural co-articulation and transition effects. In particular, an efficient kinematic approach for animating hand gestures from shape specifications is presented, which provides fine adaptation to temporal constraints that are imposed by cross-modal synchrony. Copyright © 2004 John Wiley & Sons, Ltd.

Received: 7 March 2003; Revised: 30 June 2003

KEY WORDS: multimodal conversational agents; gesture animation; model-based computer animation; motion control

Introduction

Techniques from artificial intelligence, computer animation, and human–computer interaction are increasingly converging in the field of embodied conversational agents.[1] Such agents are envisioned to have similar properties to humans in face-to-face communication, including the ability to generate simultaneous verbal and non-verbal behaviours. This includes co-verbal gestures that humans frequently produce during speech to emphasize, clarify or even complement the conveyance of central parts of an utterance.

Current conversational agents, e.g. the real estate agent *REA*,[2] the pedagogical agent *Steve*[3] or in the *BEAT* system,[4] generate their multimodal utterances by, first, planning the communicative acts to be performed and, secondly, synthesizing verbal and non-verbal behaviors. The latter stage involves the generation of appropriate, intelligible verbal and gestural acts per se as well as their combination in a seamless, human-like flow of multimodal behaviour. At the same time, verbal and non-verbal behaviours have to be finely synchronized at distinct points of time to ensure coherence of the resulting utterances. For example, the co-expressive elements in speech and co-verbal gesture appear in semantically and pragmatically coordinated form[5] and—vitally important—in temporal synchrony even at the level of single syllables. Meeting these demands for synchrony, continuity, and lifelikeness simultaneously poses continuous problems for the automatic synthesis of multimodal utterances in conversational agents.

In our lab, the anthropomorphic agent *Max* is under development. Max acts as an assembly expert in an immersive 3D virtual environment (see Figure 1 for the overall scenario). The agent demonstrates assembly procedures to the user by combining facial and upper limb gestures with spoken utterances. An important aspect in synthesizing Max's utterances is the real-time creation of synchronized gestural and verbal behaviors from application-independent descriptions of their outer form. Such descriptions are supposed to be created during high-level utterance planning and to be specified in MURML, an XML-based representation language.[6] In this paper, we present the utterance production model employed in Max with a focus on the gesture animation process. After discussing related work in the next section, we describe a production model that employs natural mechanisms of crossmodal adaptation to incrementally create fluent and coherent utterances of multiple verbal and gestural parts. The model combines a system for synthesizing accented speech with a hierarchical approach to planning and controlling upper-limb movements of an articulated figure, the high-level planning stages being

*Correspondence to: Stefan Kopp, Artificial Intelligence Group, Faculty of Technology, University of Bielefeld, D-33594 Bielefeld, Germany.
E-mail: skopp@techfak.uni-bielefeld.de

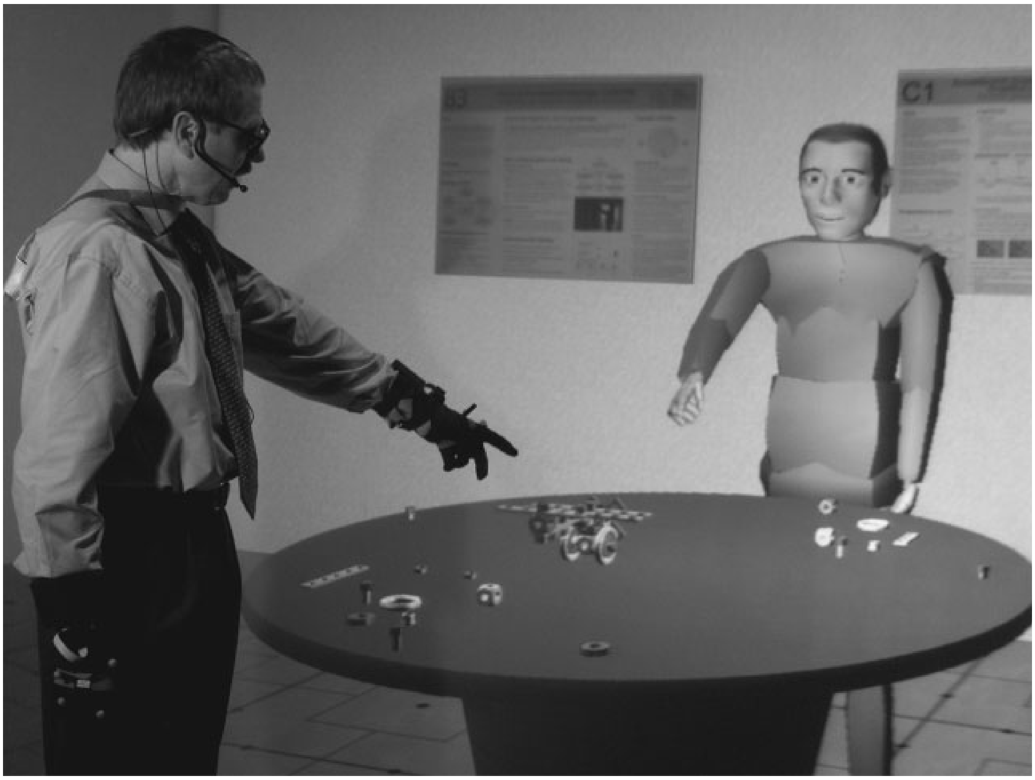


Figure 1. Multimodal interaction with Max.

described by Kopp and Wachsmuth.[7] Then, we focus on the problem of creating adequate gesture animations in real time and present a kinematic approach that emphasizes the accurate and reliable reproduction of given spatio-temporal gesture features.

Related Work

In current conversational agents, co-verbal gestures are usually created for the rhematic elements in speech by mapping communicative acts onto gestural behaviours drawn from static libraries.[2,3] Due to the demand for realism and real-time capability, such behaviours are associated with animations that are either captured from real humans or manually predefined to a large extent, sometimes being parameterizable or combinable to more complex movements. In the Animated Conversation[8] and REA system[2] as well as in the recent BEAT system,[4] Cassell *et al*. succeeded in predicting the timing of gesture animations from synthesized speech such that the expressive phase coincides with the most prominent syllable in speech. Yet, the employed techniques suffer from limited flexibility when it comes to adjusting a gesture's timing accordingly,[2,9] or concatenating them to continuous motion. Cassell[10] states that the problem of creating gesture animations and synchronizing them with speech has not been solved so far, 'due in part to the difficulty of reconciling the demands of graphics and speech synthesis software' (p. 16). This can be ascribed, first, to the lack of sufficient means of modulating, e.g. shrinking or stretching, single gesture phases[4] and, secondly, to a behavior execution that runs 'ballistical', i.e. without the possibility to exert influence, in an animation system whose reliability is sometimes hard to predict.

A fully automatic creation of upper limb movements by means of applying control models was targeted by only few researchers. Koga *et al*.[11] proposed a purely kinematic model for simulating pre-planned arm movements for grasping and manipulating objects. In particular, this work succeeded in applying findings from neurophysiology to create natural arm postures. Approaches based on control algorithms in dynamic simulations or optimization criteria provide a high level of control and may lead to physically realistic movements. However, these techniques suffer from difficulties in formulating control schemes for highly articulated figures and immense computational cost. Matarić *et al.*[12] stress the problem of determining appropriate control strategies and propose the combined application of different controllers for simulating natural upper limb movements. Gibet *et al.*[13] apply generic error-correcting controllers for generating sign language from script-like specifications. Their models succeeded in simulating natural movement characteristics to some extent but did not focus on how to meet various timing constraints as required in co-verbal gesture.

In summary, the problem of synchronizing gesture animations with spoken utterances has not been solved beyond bringing single points in more or less atomic behaviours to coincidence. The current state is particularly insufficient for virtual agents that shall be able to produce more extensive, coherent multimodal utterances in a smooth and lifelike fashion.

An Incremental Model of Speech–Gesture Production

Our approach to synthesizing multimodal utterances starts from straightforward descriptions of their desired outer form, which are supposed to be generated at higher levels of utterance planning and to be specified in MURML, an XML-based representation language.[6] Such descriptions contain the verbal utterance, augmented with co-verbal gestures—explicitly stated in terms of form features—by defining only their affiliation to certain linguistic elements. An example is shown in Figure 2. Taking MURML specifications as input, our production model aims at creating synchronized verbal and non-verbal behaviours in a human-like flow of multimodal behaviour.

The Segmentation Hypothesis

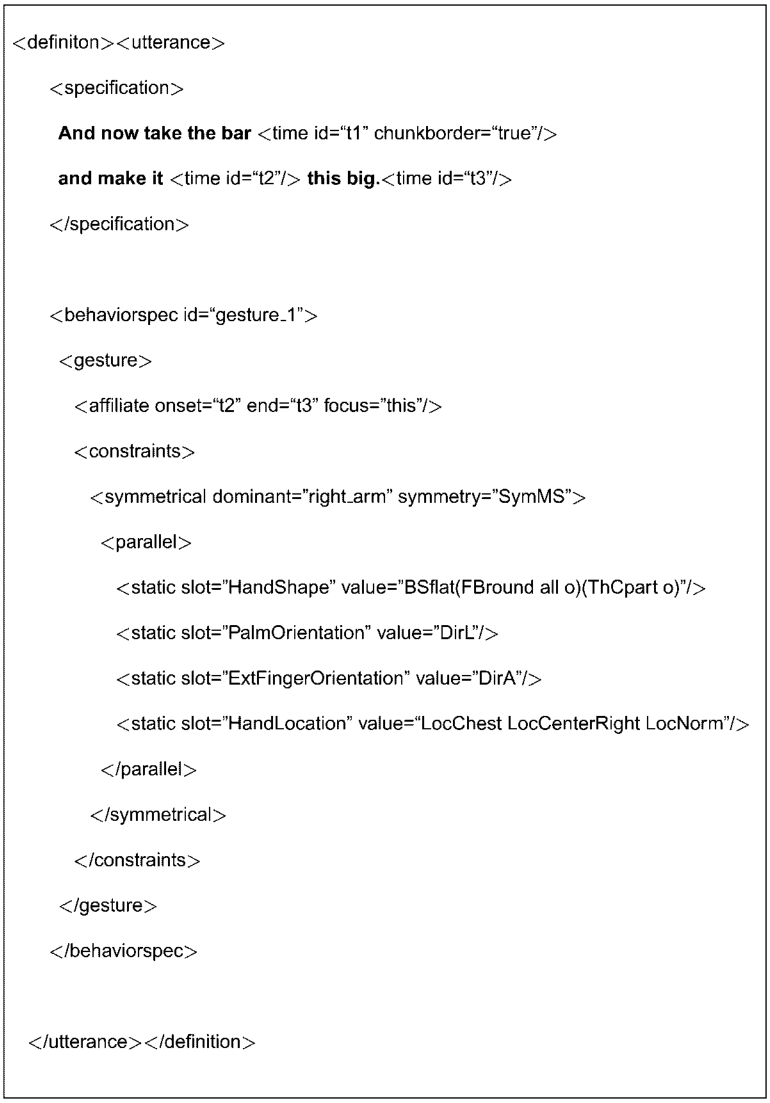
In order to organize the production of gesture and speech over multiple sequential behaviours, we adopt an empirically suggested assumption[5] as a *segmentation hypothesis*: continuous speech and gesture are co-produced in successive segments each expressing a single idea unit. The inherent segmentation of speech–gesture production in humans is reflected in the hierarchical structures of overt gesture and speech and their cross-modal correspondences.[5,14]. Kendon[14] defined units of gestural movement to consist of *gesture phrases* which comprise one or more subsequent movement phases, notably *preparation*, *stroke* (the expressive phase), *retraction* and *holds*. Similarly, the phonological structure of connected speech in intonation languages such as English and German is organized over *intonation phrases*.[15] Such phrases are separated by significant pauses, they follow the syntactical phrase structure, and display a meaningful pitch contour with exactly one primary pitch accent (*nucleus*).

We define *chunks* of speech–gesture production to be pairs of an intonation phrase and a co-expressive gesture phrase, i.e. complex utterances with multiple gestures are considered to consist of several chunks. Within each chunk, the prominent concept is concertedly conveyed by a gesture and an affiliated word or subphrase (in short, affiliate). The coexpressivity is evidenced by a general temporal synchrony: gestural movements are timed such that the meaning-bearing stroke phase starts before the affiliate and frequently spans it, optionally by inserting dedicated hold phases in the flow of movement. This coupling is refined if one of the affiliated words is prosodically focused, e.g. for emphasizing or contrasting purposes, and hence carries the nucleus of the phrase. In this case, the gesture stroke starts with the nucleus at the latest and is not finished before it.[5,16,17]

Mechanisms of Cross-Modal Coordination

In humans, the synchrony of gesture and speech is accomplished by means of cross-modal adaptation. The segmentation hypothesis enables us to treat the effective mechanisms on different levels of the utterance and to organize the overall production process in stages.

Producing a Chunk. Within a chunk the synchrony between the affiliate (or nucleus) and the stroke is mainly accomplished by the gesture adapting to the structure and timing of running speech. In producing a single chunk, the intonation phrase can therefore be synthesized in advance, potentially augmented with a strong pitch accent for narrow focus. As in previous systems (e.g. BEAT[4]), absolute time information at the phoneme level is then employed to set up timing constraints for co-verbal gestural or facial behaviors. The gesture stroke is either set to precede the affiliate's onset by a given offset (per default one syllable's approximate duration of 0.3 s) or to start exactly at the nucleus if a narrow focus has been defined. In any case, the stroke is set to span the whole affiliate before retraction starts. This may be achieved for dynamic strokes with a post-stroke hold or additional repetitions, both strategies observable in humans.[5]



```
<definiton><utterance>
   <specification>
    And now take the bar <time id=“t1” chunkborder=“true”/>
    and make it <time id=“t2”/> this big.<time id=“t3”/>
   </specification>

   <behaviorspec id=“gesture_1”>
    <gesture>
      <affiliate onset=“t2” end=“t3” focus=”this”/>
      <constraints>
        <symmetrical dominant=”right_arm” symmetry=”SymMS”>
          <parallel>
            <static slot=”HandShape” value=”BSflat(FBround all o)(ThCpart o)”/>
            <static slot=”PalmOrientation” value=”DirL”/>
            <static slot=”ExtFingerOrientation” value=”DirA”/>
            <static slot=”HandLocation” value=“LocChest LocCenterRight LocNorm”/>
          </parallel>
        </symmetrical>
      </constraints>
    </gesture>
   </behaviorspec>

 </utterance></definition>
```

Figure 2. Sample XML specification of a multimodal utterance.

Combining Successive Chunks. Humans appear to anticipate the synchrony of the forthcoming affiliate (or nucleus) and stroke already at the boundary of successive chunks since it is prepared there in both modalities: the onset of the gesture phrase co-varies with the position of the nucleus, and the onset of the intonation phrase co-varies with the stroke onset.[5,16,17] In consequence, movement between two strokes depends on the timing of the successive strokes and may range from the adoption of intermediate rest positions to direct transitional movements (co-articulation). Likewise, the duration of the silent pause between the intonation phrases may vary according to the required duration of gesture preparation. Simulating these mechanisms is highly context-dependent for it has to take into account properties of the subsequent stroke (form, location, timing constraints) as well as current movement conditions when the previous chunk can be relieved, i.e. when its intonation phrase and its gesture stroke are completed.

The Production Process

Our production model combines the aforementioned coordination mechanisms to create, as seamlessly as possible, a natural flow of speech and gesture across successive coherent chunks. To this end, the classical

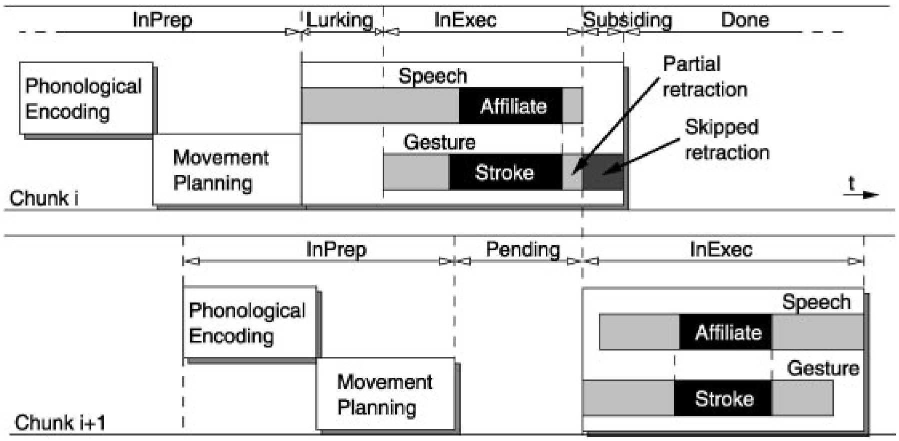


Figure 3. Incremental production of multimodal chunks.

two-phase, planning–execution procedure is extended for each chunk by additional phases in which the production processes of subsequent chunks can relieve one another. Each chunk is processed on a separate blackboard running through a series of processing states (see Figure 3).

1. *InPrep*: Separate modules for speech synthesis, high-level gesture planning and facial animation contribute to a chunk's blackboard during the overall planning process. The text-to-speech system synthesizes the intonation phrase and controls prosodic parameters like speech rate and intonation to create natural pitch accents. Concurrently, the gesture planner defines the expressive gesture phase in terms of movement constraints by selecting a lexicalized gesture template in MURML, allocating body parts, expanding abstract movements constraints and resolving deictic references (as described by Kopp and Wachsmuth[7]). At this stage, connecting effects are created when a subsequent chunk is anticipated: the pitch level in speech is maintained and gesture retraction is planned to lead into an interim rest position. Once timing information about speech has arrived on the blackboard, the face module prepares a lip-synchronous speech animation using simple viseme interpolation, augmented with eyebrow raises on accented syllables and emotional expression.

2. *Pending*: Once chunk planning has been completed, the state is set to *Pending*.

3. *Lurking*: A global scheduler monitors the production processes of successive chunks. If a chunk can be uttered, i.e. the preceding chunk is *Subsiding* (see below), the scheduler defines the intra-chunk synchrony as aforementioned and reconciles it with the onsets of the intonation and gesture phrases. In case the affiliate is located early in the intonation phrase, the scheduler lets the gesture's preparation precede speech. Due to predefined movement velocity—movement duration is estimated from its amplitude using a logarithmic law (see section on 'Gesture Motor Control')—the vocal pause between subsequent intonation phrases may thus be stretched, depending on the time consumption of the preparation phase (see Figure 3). Besides this possible adaptation to gesture, intonation phrases are articulated ballistically as prepared by the text-to-speech system. Finally, the scheduler passes control over to the successive chunk.

At this point, the motor layer is responsible for, first, planning on-the-fly upper-limb animations of the agent that exactly satisfy the given movement and timing constraints. Secondly, gesture animations must be blended autonomously according to the given timing constraints as well as the current movement conditions. For example, a gesture whose form properties require—under current movement conditions—a more extensive preparation has to start earlier to naturally meet the mandatory time of stroke onset. Since at this point the preceding gesture may have not been fully retracted, fluent gesture transitions should emerge depending on the placement of the affiliate within the verbal phrase (see Figure 3). We describe such a motor control layer for Max in the next section.

4–6. *InExec, Subsiding, Done*: Depending on feedback information from behaviour executions, which is collected on the blackboard, the chunk state then switches to *InExec*. Eventually, once the intonation phrase, the facial animation and the gesture stroke have been completed, the chunk is *Subsiding* if the gesture is still retracting or *Done* otherwise.

Gesture Motor Control

The animation of co-verbal gesture requires a high degree of control and flexibility with respect to shape and time properties while at the same time ensuring naturalness of movement. We therefore conceive a motor planner that receives timed form features of the gesture stroke and that seeks a solution to drive Max's articulated structure for complete gestural movements. Max is based on an H-Anim compatible kinematic skeleton that comprises 103 DOF in 57 joints all subject to realistic limits. For the shoulder and wrist joint, we apply the approach by Wilhelms and Van Gelder[18] to define the joint limits geometrically in terms of *reach cones* with varying twist limits.

We adopt a biologically motivated, functional–anatomical decomposition of motor control to break down this control problem in solvable subproblems, as first proposed by Zeltzer[19] for gait animation. As depicted in Figure 4, specialized motor planning modules for the hands, the wrists, the arms, as well as the neck and the eyes instantiate local motor programs (LMPs) for animating submovements, i.e. within a limited set of DOFs and over a designated period of time. LMPs may differ in the employed motion generation method, working either in external coordinates or in joint angles. To recombine the LMPs for a solution to the overall control problem, LMPs run concurrently and synchronized in an abstract motor control program (MCP) for each limb's motion (see Figure 4). The effective interplay of the LMPs within the MCP is defined by the planning modules arranging them in a controller network that lays down their potential interdependencies for mutual (de-)activation. That way, different movement phases can be created by different motion generators. For example, goal-directed wrist movements like preparation and stroke are controlled externally whereas wrist retraction is created in joint angle space (see below).

The overall motor control framework does not make any provision as to how single sub-movements are controlled themselves. Since human arm movement primarily exhibits external *kinematic* regularities (see below) and due to real-time as well as reliability requirements, we decided to control the arm, wrist and

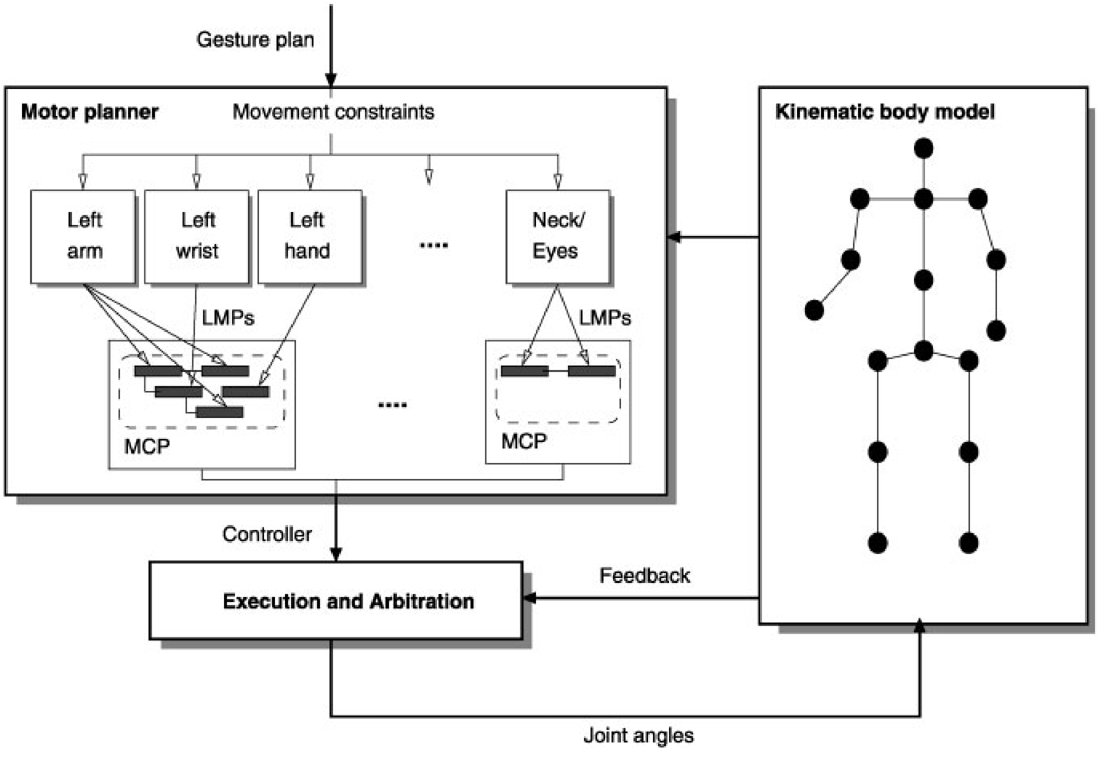


Figure 4. Overview of the motor planning and execution model.

hand movements kinematically in a feedforward way. Applied control methods—in the respective LMPs—include parametric keyframing (for hand movement, wrist retraction and swivel movement) and quaternion interpolation for wrist preparation and stroke. Arm movement can either be created in joint angle space or by defining the wrist trajectory in Cartesian coordinates. Since gestures have to reproduce external form properties—as given in the XML specification—arm movement trajectories are created directly in working space, which, in addition, allows to more easily detect and avoid body collisions. The section on 'Formation of Arm Trajectories' describes our method of constructing parametric curves that meet given position and timing constraints while resembling natural trajectories of unrestrained arm movement.

As described in the previous section, the motor control layer of Max is in charge of autonomously creating context-dependent gesture transitions. In addition to the capability of transferring activation between each other, LMPs are therefore able to (de-)activate themselves at run time on the basis of feedback about current movement conditions. To ensure fluent, at least C^1-continuous connection to the given boundary conditions, a kinematic feedforward controller cannot be created until the moment of activation of its LMP. The LMPs are therefore created in the motor planning modules with preliminary representations of the target movement (e.g. see 'Formation of Arm Trajectories' for arm movement) and are responsible for setting up their controllers when becoming active. The incessant application of active LMPs to the kinematic skeleton is coordinated by the their MCPs. For each frame, the externally formulated LMPs for wrist position, preparation/stroke of wrist flexion, and swivel movement are invoked first. Then, the inverse kinematics of the 7-DOF anthropomorphic arm is solved using an analytical algorithm from the *IKAN* package.[20] The arm's redundancy is interpreted as 'swivel' angle φ_e of the elbow about the shoulder–wrist axis and is either controlled by a dedicated LMP or is heuristically determined from the target wrist position $\vec{d}$ and longitudinal axis $\hat{z}$ of the hand (see Equation 1). The swivel is estimated by combining the sensorimotor transformation proposed by Soechting and Flanders[21] with a tendency to minimize wrist bending and damping for low arm elevations.

$$\varphi_e = a \cdot \lambda(h) \cdot (\varphi_e^S(\vec{d}) + b \cdot \varphi_e^G(\hat{z})) \tag{1}$$

where
$\varphi_e^S(\vec{d})$ is the swivel estimated from sensorimotor transformation;[21]
$\varphi_e^G(\hat{z})$ is the swivel that minimizes wrist bending;
$\lambda(h)$ is the damping factor at low arm elevations ($h < d_{\max}$), i.e.

$$\lambda(h) = \begin{cases} 0 & \text{for } h < d_{\min} \text{ (no elevation)}, \\ \sqrt{\frac{h-d_{\min}}{d_{\max}-d_{\min}}} & \text{for } d_{\min} \le h < d_{\max}, \\ 1 & \text{otherwise} \end{cases}$$

The solution arm configuration is calculated by selecting, for the 3-DOF shoulder and wrist joint, an Euler angle set that satisfies the twist limits at current joint altitude and elevation. Finally, LMPs that directly affect joint angles (neck and hand motion, wrist retraction) influence the posture by overriding the respective set of angles.

Formation of Arm Trajectories

Our approach to forming wrist trajectories relies on the well-known observation that complex arm movements consist of subsequently and ballistically performed segments with the following *kinematic* regularities of the effector trajectory:[22]

- short targeted segments are straight or curvilinear (either C- or S-shaped) and always planar;
- they exhibit a symmetrical bell-shaped velocity profile;
- a quasi-linear relation between amplitude and peak velocity, as well as an approximate logarithmic relation between amplitude and movement duration, holds;
- at any point except points of extreme bending, the movement speed v can be estimated from the radius r of the trajectory by the 'law of $\frac{2}{3}$': $v = k\, r^{\frac{1}{3}}$, where k is a constant *velocity gain factor* for each segment and assumed to be a parameter of motor control.

The segmentation of complex movements corresponds to points of maximum curvature or the change of the plane of movement. At these segmentation points, movement speed v drops and we call the associated times *breakpoints*.

Relying on these assumptions, path and kinematics of arm movements can be reproduced based on the local behavior of the trajectory at segmentation points. To this end, an intermediate representation is formed in the arm motor control module for each continuous movement phase, i.e. without any rest periods in between. Building on Morasso and Mussa Ivaldi,[23] this representation consists of a sequence of linear or curvilinear *guiding strokes*, concatenated to form the desired trajectory. Each guiding stroke bridges from one segmentation point to the next by stating the position, the time, the velocity and the velocity gain factor of the effector movement at its end point. For curvilinear guiding strokes, the normal vector of the movement plane as well as the overall form (left/right C, left/right S) must be defined in the MURML specification and are transferred to the respective guiding stroke. In addition, the form of the curvilinear segment can be optionally specified by the degree of curvature (from nearly straight to semicircle), the roundness (from nearly rectangular to nearly triangular) and the skewness (flattened at the beginning or the end).

From the sequence of guiding strokes, an LMP is created that continually estimates the duration of a hypothetical, goal-directed preparation from current hand location $\vec{x}$ to the required gesture start position $\vec{x}_s$ from the logarithmic law $T = c \cdot \log(|| \vec{x} - \vec{x}_s || + 1)$ and decides whether to activate itself. Once activated, the LMP completes the trajectory formation by (1) inserting a preparatory guiding stroke, (2) checking each guiding stroke for collisions with the torso (approximated by a bounding box) and replacing it, if necessary, with circumventing guiding strokes, (3) setting up all position constraints, (4) estimating the velocities at interior segmentation points and (5) constructing a parametric curve. The last three steps are crucial for mathematically forming a curve that satisfies the spatio-temporal gesture features while reproducing a naturally segmented velocity profile.

Setting up Position Constraints. In addition to the overall start point, the end point of each linear guiding stroke is to be interpolated at the corresponding breakpoint (end time). Curvilinear guiding strokes are approximated by three linear components; i.e. they give two additional inner data points, which can realize both curved or S-shaped segments. The positions of these data points are derived directly from the given shape properties of the segment. Their associated time points are set close to the midpoint of the segment's time interval ($t_1 = t_s + [2(t_e - t_s)/5]$ and $t_2 = t_s + [3(t_e - t_s)/5]$) to ensure a single bell-shaped velocity profile of the complete segment. As a result, a trajectory representation of g linear and c curvilinear guiding strokes gives $3c + g + 1$ position constraints with associated times, of which a total of $c + g + 1$ are breakpoints.

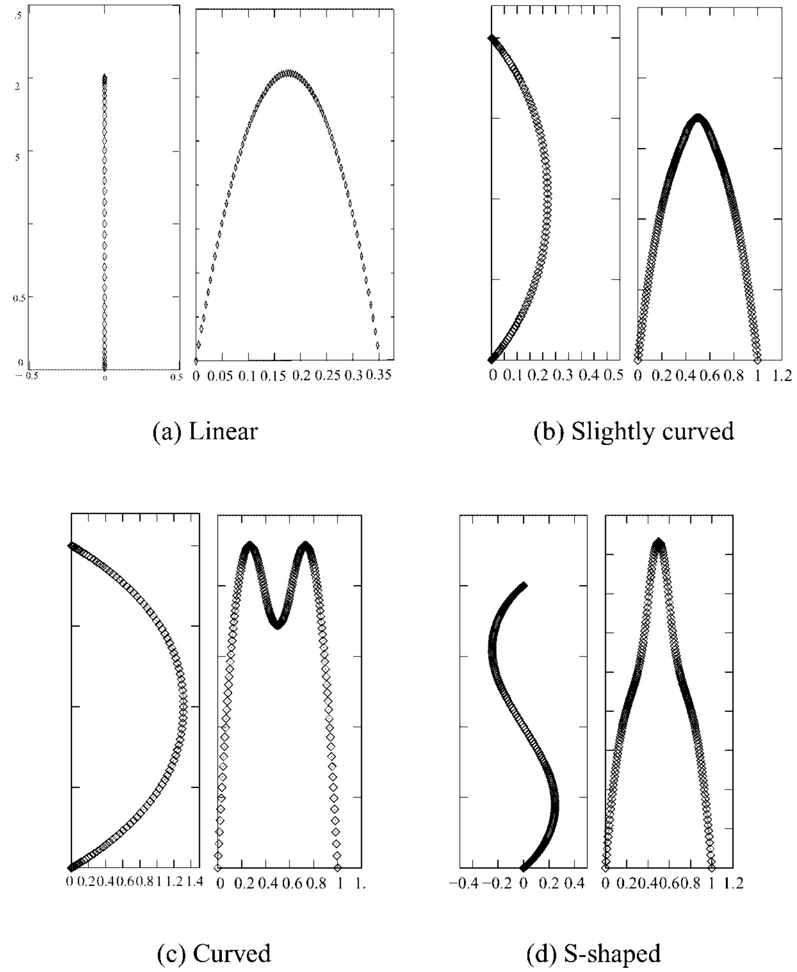
Velocity Estimation at Segmentation Points. At any interior segmentation point, two successive guiding strokes meet with an incoming ($\vec{d}_{i-1} \rightarrow \vec{d}_i$) and outgoing direction ($\vec{d}_i \rightarrow \vec{d}_{i+1}$). To estimate the velocity of the corresponding arm movement, we assume the trajectory tangent to be parallel to the overall movement direction given by the chord through $\vec{d}_{i+1} - \vec{d}_{i-1}$. The movement speed, first, generally depends on the spans of the adjacent guiding strokes and the time intervals between their breakpoints u_j. Therefore, the average movement speed ν is estimated (Equation 2). Secondly, movement speed is directly correlated to the trajectory's radius according to the 'law of $\frac{2}{3}$', except for extreme bendings. Since the trajectory at u_i tends to turn the incoming direction onto the outgoing one, we estimate its radius from the angle α between these vectors and normalize it (Equation 3). Finally, the speed is derived as shown in Equation 4. The constant k is defined as the mean of the velocity gain factors of the adjacent guiding strokes, which thus can be used to modify the style of movement. The default values being 1, lower values cause rather sharp bendings and lower velocities at breakpoints, whereas greater values lead to a more continuous and smoother movement. Yet, the factors must not be assigned a value too large to ensure naturally segmented movement kinematics with velocity drops at the segmentation points.

$$\nu = \frac{1}{2}\left(\frac{|| \vec{d}_i - \vec{d}_{i-1} ||}{u_i - u_{i-1}} + \frac{||\vec{d}_{i+1} - \vec{d}_i||}{u_{i+1} - u_i}\right) \tag{2}$$

$$\alpha = \angle(\vec{d}_i - \vec{d}_{i-1},\ \vec{d}_{i+1} - \vec{d}_i),\ r = 1 - \frac{\alpha}{\pi} \tag{3}$$

$$\vec{v} = \hat{v} \cdot k \cdot \nu \cdot r^{\frac{1}{3}} \tag{4}$$

Calculating a Parametric Curve. From the previous steps, a total of $3c + g + 1$ position constraints and $c + g + 1$ velocity constraints are given for a trajectory representation of g linear and c curvilinear guiding strokes. To construct a smooth trajectory that covers all strokes, each with a bell-shaped velocity profile, we apply a non-uniform cubic B-spline curve with $n = 4c + 2_g + 1$ control points $\vec{p}_0, \cdots, \vec{p}_n$ as first proposed by Morasso and Mussa Ivaldi.[23] Using a nonperiodic knot vector (first and last four knots equal: $t_0 = \cdots = t_3 = u_0$ and $t_{n+1} = \cdots = t_{n+4} = u_{3c+g}$), the boundary control points are automatically interpolated and the start and end tangents are parallel to $\vec{p}_{1+1} - \vec{p}_i$. The remaining $n - 3$ interior knots are distributed among the breakpoints (double multiplicity, equal to the corresponding times) and inner data points of the curvilinear guiding stroke (single knots equal to the corresponding times). The douple multiplicity narrows



(a) Linear

(b) Slightly curved

(c) Curved

(d) S-shaped

Figure 5. Generated paths and velocity profiles for different guiding strokes.

the influence of the control points, intended from the fact that the movement should be determined by local properties at the breakpoints. Consequently, interior velocities become a major means of controlling the trajectory. Finally, the control points are calculated directly from the position and velocity constraints. The resulting C^1-continuous spline gives a smooth trajectory and reproduces symmetrical bell-shaped velocity profiles for each ballistical movement unit. Furthermore, the quasi-constant relationship between amplitude and maximum speed for human movements of constant duration is accounted for. Figure 5 shows example trajectories and velocity profiles for single linear (a) and various curvilinear (b–d) segments. For comparison, experimentally observed human arm trajectories[23] are shown in Figure 6. A more complex gesture example, synthesized in real time from a MURML gesture form specification, is depicted in Figure 7. The corresponding velocity profile of the arm movement is shown in Figure 8. Animation examples can be found at the Max web page.*

Example

In experiments conducted in our lab, sensory data as well as video recordings of human multimodal utterances were acquired. To evaluate the presented model, a sample utterance was manually transcripted in MURML and automatically reproduced by Max. Figure 9 shows the original utterance that is divided into three chunks. The resulting synthetic utterance is

*http://www.techfak.uni-bielefeld.de/~skopp/max.html

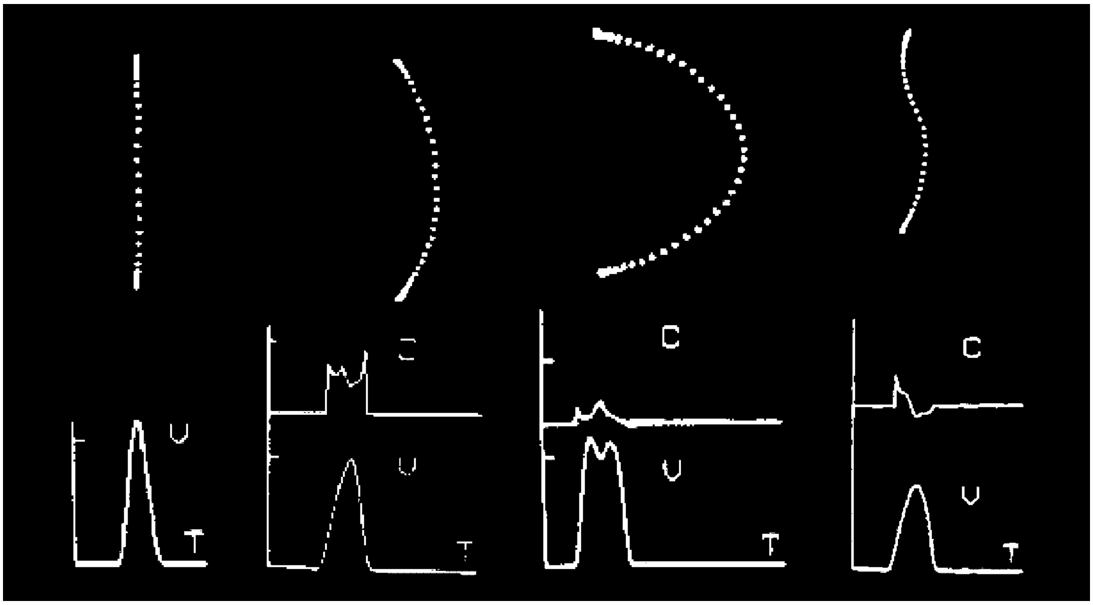


Figure 6. Experimentally observed trajectories (top) and velocity profiles (bottom) of goal-directed arm movement (Morasso and Mussa Ivaldi,[23] p. 136).

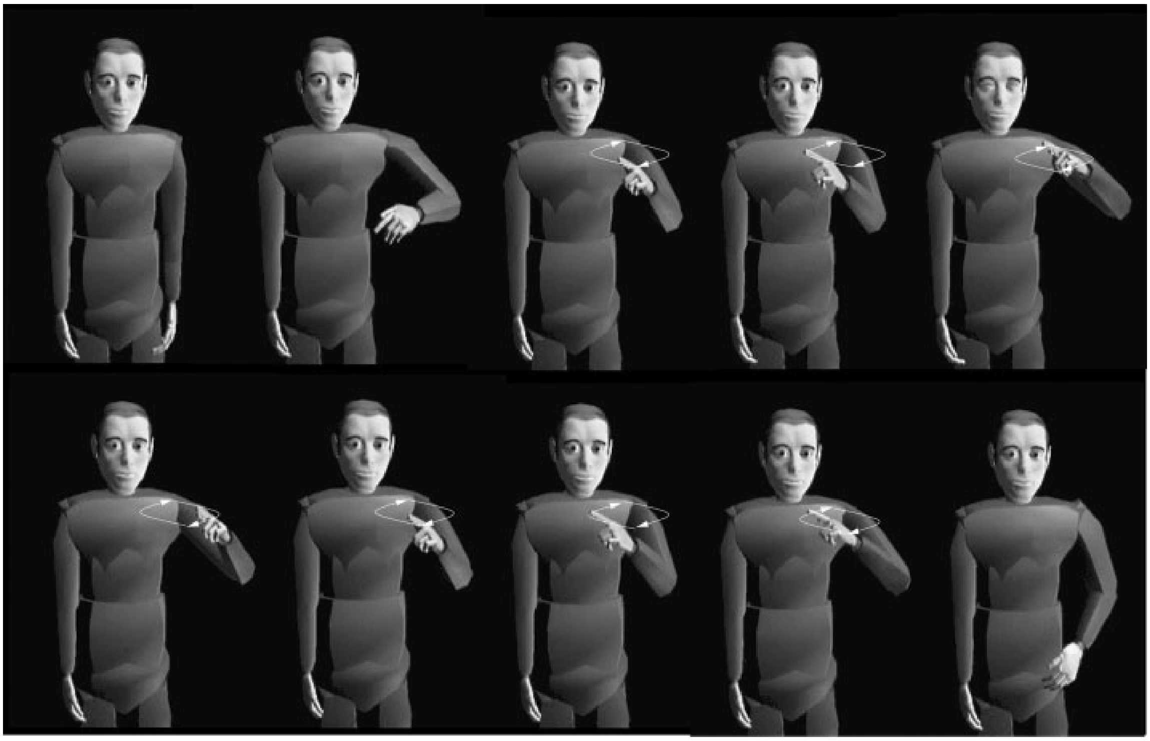


Figure 7. Sample gesture with an expressive phase out of three curvilinear guiding strokes forming a circular movement (indicated by the arrows).

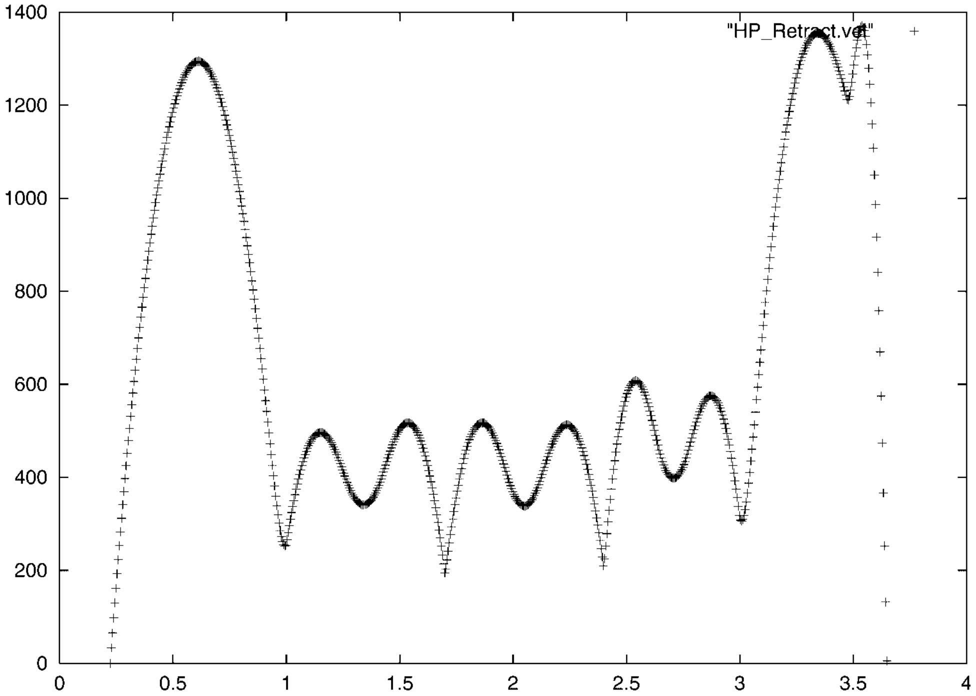


Figure 8. Velocity profile of the gesture shown in Figure 7. The slight final peak is caused by a linear guiding stroke that has been inserted to avoid a body collision during gesture retraction.

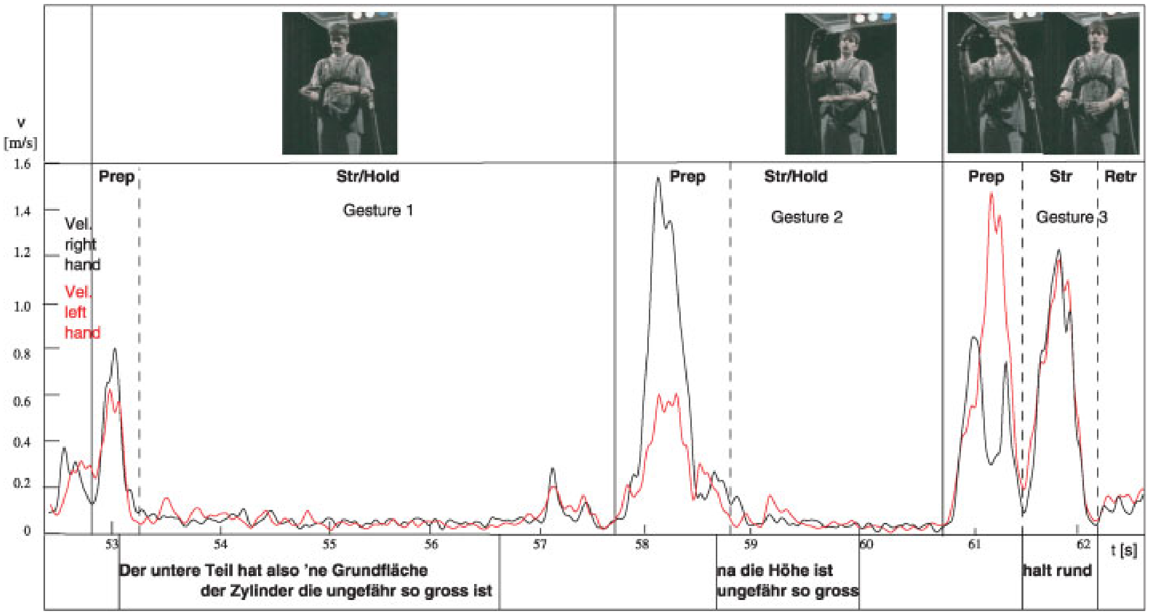


Figure 9. Experimentally observed utterance in which a subject multimodally describes a geometric shape. The bottom shows the verbal channel with three German phrases and silent pauses in between. Above, the velocity profiles of the arm movements are shown during three two-handed co-verbal gestures, the first two being static, the last one dynamic.

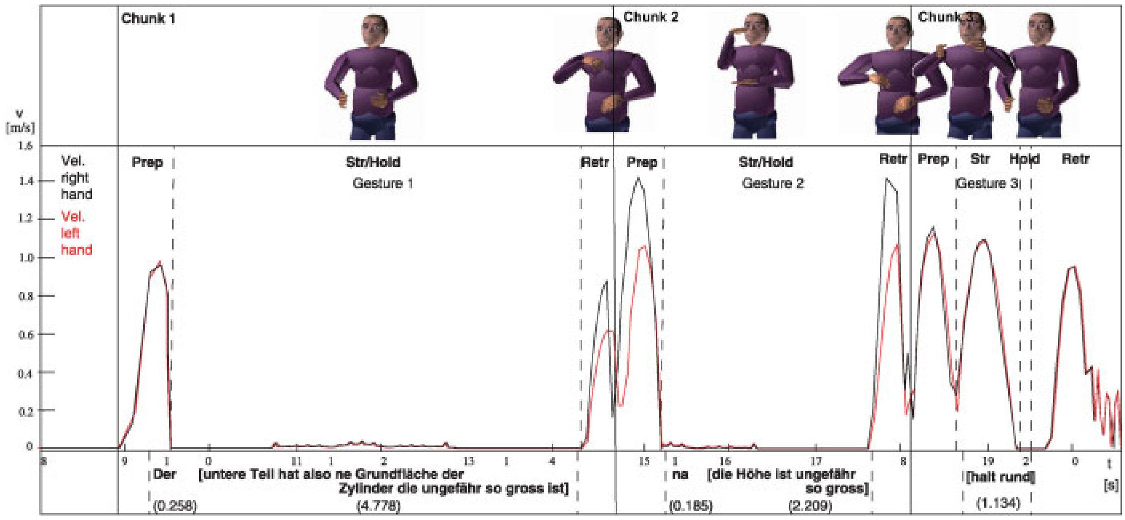


Figure 10. The utterance in Figure 9, reproduced by Max in real-time from manually created MURML transcriptions. In the verbal channel (bottom), the specified affiliates of the gestures are parenthesized.

depicted in Figure 10. Comparing both, we can state that, although the text-to-speech system dictated a slower speech rate, the internal temporal structure of the natural utterance was successfully reproduced. Due to the rule-based definition of gesture precedence all three chunks start with the gesture preparation which is consistent with the timing produced by the subject. The same holds for the verbal pauses inserted by Max to compensate for time-consuming preparation movements. This indicates that the model-based gesture animation does not produce unnatural velocities or movement durations. However, in contrast to the subject who maintains post-stroke holds until the next preparation sets in, Max starts with gesture retraction immediately after the affiliate's end but fluently blends it into the subsequent preparation phase. Furthermore, the subject exhibits many more secondary movements than Max, which in comparison makes Max appear a bit stiff. This problem can be tackled by permanently overlaying stochastic noise to single DOF as proposed by Perlin.[24]

Conclusion

Lifelike multimodal utterances of a wide variety are highly desirable for conversational agents. Instead of drawing predefined behaviours from fixed libraries as in most existing systems, all verbal and non-verbal utterances in our system are created on-the-fly from XML specifications of their overt form. To this end, a kinematic model of motion control was developed for generating gesture animations in real time. It comprises a model-based method for forming a parametric curve that achieves natural path *and* kinematics for a required wrist movement in space and provides satisfactory quality. Using a model in creating gesture animations from the scratch allows, first, to finely adapt gestural movements to accompanying speech and, secondly, to create context-dependent transition effects between successive gestures. These features are exploited in an incremental production model that combines the synthesis of synchronized gestural, verbal and facial behaviours with mechanisms for linking them to form fluent utterances. The resulting synthetic utterances achieve cross-modal synchrony even at the syllable level while reproducing natural co-articulation and transition effects. Our methods are demonstrable with the Max system, which exceeds the ability of current multimodal agents, in which synchronization of synthetic gestural and spoken utterances is accomplished by just bringing single points of behaviours to coincide.

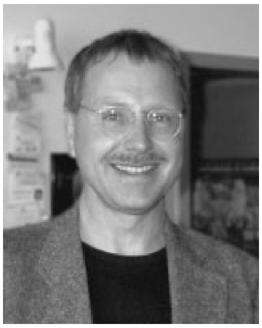
Concerning future work, it appears natural to further exploit the flexibility and generality of our synthesis model for the automatic planning of multimodal utterances of a wide variety. Ongoing work aims at enabling Max to imitate iconic gestures not only by reproducing their shape but also by understanding the underlying imagistic content and re-expressing it in an alternative gestural form. Another challenging aspect is the synchronization of the moments of stress in both modalities, which is often supposed to coordinate speech and gesture beyond the level of gesture stroke and affiliate. We expect this to yield a coordinated accentuation, e.g. according to an underlying rhythmic pulse, and to include the timing of velocity peaks of single movement phases which can be taken into account in our approach.[6] Furthermore, the gesture animation model will be further explored with respect to variations of the parameters, e.g. influencing the relationship between trajectory curvature and velocity. Modulating these variables systematically within reasonable limits may enable a fine-grained modulation of the agent's style of movement.

References

1. Cassell J, Sullivan J, Prevost S, Churchill E (eds). *Embodied Conversational Agents*. MIT Press: Cambridge, MA, 2000.
2. Cassell J, Bickmore T, Campbell L, Vilhjalmsson H, Yan H. Human conversation as a system framework: designing embodied conversational agents. In *Embodied Conversational Agents*, Cassell J *et al*. (eds). MIT Press: Cambridge, MA, 2000; 29–63.
3. Rickel J, Johnson WL. Animated agents for procedural training in virtual reality: perception, cognition, and motor control. *Applied Artificial Intelligence* 1999; **13**: 343–382.
4. Cassell J, Vilhjalmsson H, Bickmore T. BEAT: the behavior expression animation toolkit. In *Proceedings of SIGGRAPH 2001*, 2001; pp 477–486.
5. McNeill D. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press: Chicago, 1992.
6. Kopp S, Wachsmuth I. Model-based animation of coverbal gesture. In *Proceedings of Computer Animation 2002*. IEEE Computer Society Press: Los Alamitos, CA, 2002; pp 252–257.
7. Kopp S, Wachsmuth I. A knowledge-based approach for lifelike gesture animation. In *ECAI 2000 Proceedings of the 14th European Conference on Artificial Intelligence*, Horn W (ed.). IOS Press: Amsterdam, 2000; pp 661–667.
8. Cassell J, Pelachaud C, Badler N, Steedman M, Achorn B, Becket T, Douville B, Prevost S, Stone M. Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proceedings of SIGGRAPH'94*, 1994.
9. Churchill EF, Cook L, Hodgson P, Prevost S, Sullivan JW. 'May I help you?': designing embodied conversational agent allies. In *Embodied Conversational Agents*, Cassell J *et al*. (eds). MIT Press: Cambridge, MA, 2000; 64–94.
10. Cassell J. Nudge nudge wink wink: elements of face-to-face conversation for embodied conversational agents. In *Embodied Conversational Agents*, Cassell J *et al*. (eds). MIT Press: Cambridge, MA, 2000; 64–94.
11. Koga Y, Kondo K, Kuffner J, Latombe J-C. Planning motions with intentions. In *Proceedings of the 21st Annual Conference on Computer Graphics*, 1994; pp 395–408.
12. Matarić MJ, Zordan VB, Williamson MM. Making complex articulated agents dance. *Autonomous Agents and Multi-Agent Systems* 1999; **2**(1): 23–44.
13. Gibet S, Lebourque T, Marteau P-F. High-level specification and animation of communicative gestures. *Journal of Visual Languages and Computing* 2001; **12**(6): 657–687.
14. Kendon A. Gesticulation and speech: two aspects of the process of utterance. In *The Relationship of Verbal and Nonverbal Communication*, Key MR (ed.). Mouton: The Hague, 1980; 207–227.
15. Levelt WJ. *Speaking*. MIT Press: Cambridge, MA, 1989.
16. de Ruiter JP. Gesture and speech production. PhD thesis, University of Nijmegen, 1998. MPI Series in Psycholinguistics.
17. Nobe S. Where do *most* spontaneous representational gestures actually occur with respect to speech? In *Language and Gesture*, McNeill D (ed.). Cambridge University Press: Cambridge, UK, 2000.
18. Wilhelms J, Van Gelder A. Fast and easy reach-cone joint limits. *Journal of Graphics Tools* 2001; **2**(6): 27–41.
19. Zeltzer D. Motor control techniques for figure animation. *IEEE Computer Graphics and Applications* 1982; **2**(9): 53–59.
20. Tolani D, Goswami A, Badler NI. Real-time inverse kinematics techniques for anthropomorphic limbs. *Graphical Models* 2000; **62**: 353–388.
21. Soechting JF, Flanders M. Sensorimotor representations for pointing to targets in three dimensional space. *Journal of Neurophysiology* 1989; **62**(2): 582–594.
22. Latash ML. *Control of Human Movement*. Human Kinetics: Champaign, IL, 1993.
23. Morasso P, Mussa Ivaldi FA. Trajectory formation and handwriting: a computational model. *Biological Cybernetics* 1982; **45**: 131–142.
24. Perlin K. Real-time responsive animation with personality. *IEEE Transactions on Visualization and Computer Graphics* 1995; **1**(1): 5–15.

Authors' biographies:

Stefan Kopp is a faculty research assistant in the Artificial Intelligence Group at the University of Bielefeld, Germany. In 1998, he obtained a masters degree in computer science from the University of Bielefeld, where he also received his Ph.D. in 2003 for research on the synthesis and coordination of gesture and speech for virtual multimodal agents. His areas of interest include intelligent agents, in particular embodied conversational ones, human computer animation, multimodal systems, and virtual reality. Stefan Kopp is member of the Collaborative Research Center SFB 360 'Situated Artificial Communicators' at Bielefeld.



Ipke Wachsmuth is director of the Center for Interdisciplinary Research (ZiF) and chair of Artificial Intelligence at the University of Bielefeld, Germany. He holds a Mathematics Master's degree and a Ph.D. both from the University of Hannover, Germany, and an Informatics Habilitation degree from the University of Osnabruck, Germany. A major part of his current research is related to the Collaborative Research Center 'Situated Artificial Communicators' where he conducts joint projects with linguists and psycholinguists on situated language and coverbal gesture. Practical aspects of his research aim at advanced human-machine interfaces which realize embodied communication in fully immersive virtual reality.