MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents

Alfred Kranstedt
Artificial Intelligence Group
Faculty of Technology
University of Bielefeld
D-33594 Bielefeld, Germany
akranste@techfak.uni-bielefeld.de

Stefan Kopp
Artificial Intelligence Group
Faculty of Technology
University of Bielefeld
D-33594 Bielefeld, Germany
skopp@techfak.uni-bielefeld.de

Ipke Wachsmuth
Artificial Intelligence Group
Faculty of Technology
University of Bielefeld
D-33594 Bielefeld, Germany
ipke@techfak.uni-bielefeld.de

ABSTRACT

This paper presents work on an artificial anthropomorphic agent with multimodal interaction abilities. It focuses on the development of a markup language, MURML, that bridges between the planning and the animation tasks in the production of multimodal utterances. This hierarchically structured notation provides flexible means of describing gestures in a form-based way and of explicitly expressing their relations to accompanying speech.

Keywords

Multimodal Communication, Form-based Gesture Description, Markup Languages, Conversational Agents

1. INTRODUCTION

Interaction metaphors of human-computer-interaction have changed in the past from typed input and output to WIMP-style interaction (Windows, Icons, Menu, Pointing). And it changes again with the outcome of new fields of applications in virtual reality. These environments raise new demands on how to interact with and in virtual worlds. In response, there are numerous attempts to utilize natural multimodal communication including the field of embodied conversational agents in which several techniques from artificial intelligence, computer graphics and human-computer-interaction are recently converging [2]. Our research focuses on the interrelations between the various modalities in face-to-face conversation by equipping the agents with similar interactional abilities as humans employ. This includes the reception and generation of simultaneous and synchronized verbal and nonverbal utterances. The automatic generation of natural speech with simultaneous gesture requires a time-critical production process with high flexibility. This problem can be solved by generating gesture and speech in real-time from flexible representations that specify the visible features of the mandatory parts of a gesture in time correlation with the spoken output.

In this paper we present our approach we realized in *Max* (see Fig. 1), a virtual anthropomorphic agent that acts as a mediator in an immersive 3D virtual environment for simulated assembly and design tasks [14]. Max is capable of producing smooth coverbal gestures in synchrony with synthetic speech solely from application-independent descriptions of their outer form. The gesture animation process builds on a hierarchical model of planning and controlling the upper-limb movements of an articulated figure and is described in [8, 9]. We propose an XML-based representation language for multimodal utterances, MURML, that can be processed by our generation model. After discussing related work in the next section, Section 3 explains how to describe target utterances that comprise coordinated verbal and gestural behaviors in MURML and, in particular, how to flexibly specify the desired gesture using a feature-based description. In Section 4, we explain how the XML description is utilized as representation system during several processing steps that are taken in preparation of utterance planning and execution.

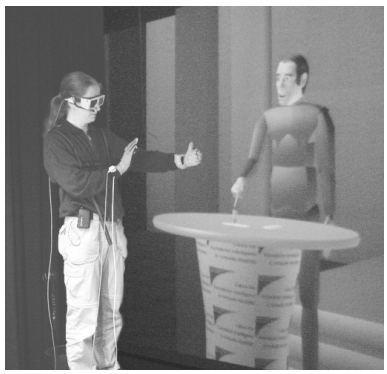


Figure 1: Multimodal interaction with Max

2. RELATED RESEARCH

In recent times a growing number of researchers focus on the development of representation systems for multimodal utterances. They approach this topic from different points. Some researchers [12, 3] replenish typed input with synchro-

nized nonverbal behaviors, manually inserted [12] or automatically determined during a linguistic and contextual analysis of the text [3]. These approaches identify speech as the dominant modality responsible for time structure of the utterance. Synchrony is achieved by starting the animation of nonverbal behaviors simultaneously with the correlated verbal phrases. This approach is in a strict way behavior-based. The nonverbal behaviors are referred to by unique identifiers and are drawn from a behavior database. The gesture ontology in these sytems does not allow to create gestures from atomic elements and to adapt their structure in the synchronization process.

With notation systems for sign languages like HamNoSys [13], on the other hand, researchers have developed form-based gesture descriptions that allow to specify a wide range of manual gestures by symbolically composing form and movement primitives in a structured way. These approaches strictly distinguish between the syntactic specification of gesture features and the semantics of this behavior. Building on these notations, automatic translation systems from speech to sign language with an animated visual interface agent are being developed [5, 7]. Time relations in the form-based gesture descriptions are utilized for synchronization purposes. The aim of this work is the animation of sign language and not coverbal gesture. It incorporats nonmanual gestures like facial expressions and body language relying on manual gesture as the leading modality. There is no necessity to adapt the temporal structure of the manual gesture in a flexible way. Similar to Cassell et al [3], the ViSiCAST Project [7] uses an XML-based language for the sake of flexibility and adaptability.

Our work can be seen as a synthesis of both approaches described. We use the potency and high granularity of form-based gesture descriptions and augment them with suitable means of describing cross-modal time relations.

3. UTTERANCE SPECIFICATION

Our approach to synthesizing multimodal utterances starts from straightforward descriptions of their overt form in an XML-based specification language (see Fig. 3). Such a description contains the verbal utterance augmented by nonverbal behaviors including gestures. The correspondence between gesture and speech at this surface level is commonly assumed to exist between certain units on different levels of the hierarchical structure of both modalities [6, 11]. Kendon [6] defined *units* of gestural movement to consist of *gesture phrases* which comprise one or more subsequently performed movement *phases*, notably *preparation*, *holds*, *stroke*, and *retraction*. Speech is produced in several locutions, whose intonational contour in English and other languages (including German) is organized over *intonational phrases* (cf. [10]). Such phrases are separated by significant pauses and display a meaningful pitch contour with exactly one pitch accent being most prominent[1] (the *nucleus*). We adopt the empirical assumption [11] that continuous speech and gesture are co-produced in successive units each expressing a single idea unit. We define *chunks* of speech-gesture production to consist of a intonational phrase in overt speech and a co-expressive gesture phrase (see Fig. 2), i.e., complex utterances with multiple gestures are conceived divided in several chunks. Within each chunk, the gesture stroke corresponds to the focussed constituent (a single word or a subphrase of a few words) in the intonational phrase (the *affiliate*), which has the nuclear accent.

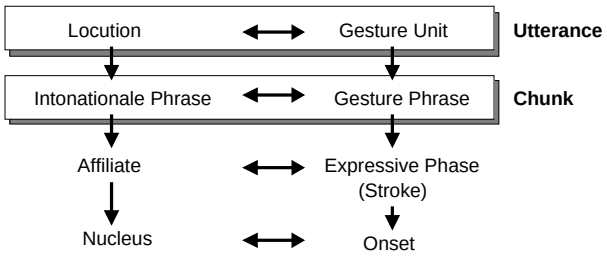


Figure 2: Units of speech-gesture correspondence.

Our XML utterance specifications start from the textual output, during which certain points in time can be defined by markups. The verbal part of a complex multimodal utterance that comprises multiple gestures must be divided into chunks by annotating the corresponding time tags. Then, in the subsequent definition of the nonverbal behavior, correspondence between speech and gesture is expressed by specifying the affiliate's onset and end. Gestures can be stated by specifying a required communicative function sufficient for the agent to choose an appropriate behavior from a gesture lexicon. Alternatively, the desired gesture can be described explicitly in terms of its spatiotemporal features as explained in the following subsection. Fig. 3 shows a description of an utterance that comprises a deictic and an iconic gesture. The first one is selected from the agent's repository (a lexicon of XML definitions) based on a provided communicative function ('refer_to_loc') which in this case is parametrized by a certain object's position. The second iconic gesture, illustrated in the right part of Fig. 3, is stated in terms of the gesture's spatiotemporal features as described below. An animation created from this specification by our generation model can be found at our web page [1].

In addition to gestures, further behaviors can be incorporated such as arbitrary body movements, defined as timed parametric keyframe animations using ease in/out, and facial animations given as sequences of face muscle values.

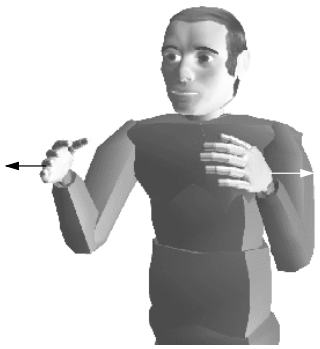
3.1 Gesture specification

With respect to its communicative intent, a gesture is sufficiently described by specifying the spatio-temporal features of its stroke as the meaningful phase of the gesture. To this end, we define a hand/arm configuration in terms of three features whose values can be defined either numerically or symbolically (using augmented *HamNoSys* [13] descriptions; see Fig. 4): (1) the *location* of the wrist, symbolically specified by unique identifiers for the position in the frontal, transversal, and sagittal plane (Figure 6); (2) the *shape* of the hand, compositionally described by the overall hand shape and additional symbols denoting the kind and degree of flexion within each finger; (3) the *orientation* of the wrist, compositionally described by symbols representing a vector, also w.r.t. the three planes located in the agent, originating at the wrist, running along the length of

[1] If an elements is prosodically focussed, the primary pitch accent expresses its prominence.

```
<definition><utterance>
    <specification>
     And now take <time id="t1"/> this bar <time id="t2" chunkborder="true"/>
     and make it <time id="t3" /> this big. <time id="t4"/>
    </specification>
    <behaviorspec id="gesture_1"><gesture>
      <affiliate onset="t1" end="t2"/>
      <function name="refer_to_loc">
       <param name="refloc" value="$Loc-Bar_1"/>
      <function/>
     </gesture></behaviorspec>
```

```
    <behaviorspec id="gesture_2"><gesture>
      <affiliate onset="t3" end="t4"/>
      <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"/>
         <dynamic slot="HandLocation">
          <dynamicElement type="linear">
           <value type="start" name="LocShoulder LocCenterRight LocNorm"/>
           <value type="direction" name="DirR"/>
           <value type="distance" name="125.0"/>
          </dynamicElement>
         </dynamic>
        </parallel>
       </symmetrical>
      </constraints>
     </gesture></behaviorspec>
```



```
  </utterance></definition>
```

Figure 3: Sample XML utterance specification including an iconic gesture as illustrated right.

the back of the hand (*extended finger orientation; EFO*), and the normal vector of the palm (*palm orientation; PO*). The latter may be given either absolutely w.r.t. the agent's overall frame of reference or relatively w.r.t. to the extended finger orientation (interpreted as a rotation about this axis).

Flexibility of gesture generation means that, on the one hand, all spatiotemporal features of a gesture can be specified in accordance to the individual context of accompanying speech. On the other hand, one may want to define templates for frequently used gestures like pointing. To this end, a gesture description can accommodate parametrizeable feature values as in the following example. Global parameter tags can be defined at the beginning of the overall utterance specification that set up the individual context of this utterance in terms of slot-value pairs.

```
 ...
<parameter slot="object_loc_1" value="1500 10 100"/>
...
<static slot="ExtFingerOrientation" value="$object_loc_1"
mode="pointTo"/>
...
```

The underlying idea of a MURML gesture representation is that the stroke phase can be considered as an arbitrarily complex combination of submovements within the three features described. As illustrated with the example in Fig. 3, two different types of *movement constraints* are provided for specifying a feature over a certain period of time: (1) a *static* constraint defines a postural feature, which is to be held for a certain period of time before retraction; (2) a *dynamic* constraint specifies a significant submovement within

HandLocation	LocFrontal LocTransversal LocSagittal
LocFrontal	"LocAboveHead", "LocHead", "LocForehead", "LocEyes", "LocNose", "LocMouth", "LocChin", "LocLowerChin", "LocNeck", "LocShoulder", "LocUpperChest", "LocLowerChest", "LocStomach", "LocBelowStomach", "LocHip", "LocBelowHip"
LocTransversal	"LocCCenter", "LocBackof", "LocCCenterRight", "LocCCenterLeft", "LocCenterRight", "LocCenterLeft", "LocPeripheryRight", "LocPeripheryLeft", "LocExtremePeripheryRight", "LocExtremePeripheryLeft", "LocLeft", "LocRight"
LocSagittal	"LocContact", "LocNear", "LocNorm", "LocFar", "LocStreched"
Handshape	Basicsymbol [(Thumbsymbol)] {(Fingersymbol)}
Basicsymbol	"BSfist", "BSflat", "BSffinger"
Thumbsymbol	"ThExt", "ThAcr", "ThCpart" [Degree]
Fingersymbol	"FBstr" Finger, "FBangle" Finger [Degree], "FBround" Finger [Degree], "FBroll" Finger, "FBbent" Finger, "FBstiff" Finger
Finger	"all", "p", "m", "r", "l"
Degree	"c", "no", "o", "wo"
Handorientation	ExtFingerOrientation PalmOrientation
ExtFingerOrientation	"Dir" Directions
PalmOrientation	absPalmOrientation | relPalmOrientation
absPalmOrientation	"Dir" Directions
Directions	Direction[Directions]
Direction	"A", "T", "R", "L", "U", "D"
relPalmOrientation	"PalmL", "PalmLU", "PalmLD", "PalmR", "PalmRU", "PalmRD", "PalmU", "PalmD", "PalmA", "PalmAL", "PalmAR", "PalmT", "PalmTL", "PalmTR", "PalmAU", "PalmAD", "PalmTU", "PalmTD"

Figure 4: The three features, and possible values (in EBNF), of the MURML spatiotemporal description of the gesture stroke.

Tag	Content elements	Attributes
parallel	symmetrical, repeat, repeat_alt, sequence, static, dynamic	start, end
sequence	symmetrical, parallel, static, dynamic, repeat, repeat_alt	start, end
symmetrical	parallel, sequence, static, dynamic	dominant, symmetry, center
repeat	symmetrical, parallel, sequence, dynamic	number
repeat_alt	symmetrical, parallel, sequence, dynamic	number

Figure 5: Notation elements for the describtion of submovement relations.

a feature that is fluently connected with adjacent movement phases. While start and end time for static constraints can be directly assigned, dynamic constraints are made up of segments whose timing (start/end time, moment of peak velocity) has to be defined explicitly. In order to allow for complex trajectories, each segment of a dynamic wrist location constraint may be defined either as "linear" (default), "curve", or "circle" with respective location specifications.

The overall structure of a gesture is given by the relationships between the feature constraints, e.g., moving the hand up while keeping a fist. To this end, *simultaneity*, *posteriority*, *repetition*, and *symmetry* of submovements can be denoted by specific MURML elements constituting a constraint tree for the gesture. Fig. 5 itemizes the elements and their possible content elements and attributes.

For symmetric two-handed gestures HamNoSys discriminates between the movement of the dominant and the following hand. Analogue to this approach, we define eight different symmetries (given as attribute value to the "symmetry" tag) made up of combinations of mirror symmetries w.r.t. the frontal, transversal, and sagittal plane of the agent's body. Fig. 6 shows the three main body planes and the permutations (l(eft)-r(ight), u(p)-d(own), f(orward)-b(ackward)) each symmetry causes in hand location, extended finger orientation, and palm orientation to the configuration of the following hand w.r.t. the dominant one.

The example gesture specified in Fig. 3 ("gesture_2") is defined by the movement of the dominant right hand in addition to static a wrist orientation and hand shape. Left hand movement results from mirror symmetry w.r.t. the sagittal plane ("SymMS"). In addition, the gesture is defined to correspond to the verbal phrase "this big" by the "affiliate" tag.

4. UTTERANCE PROCESSING

During the parsing of the XML specification, a hierarchically structured representation of the utterance is constructed (each non-empty tag corresponding to a node; see Fig. 7). This tree acts as the central data structure during the first processing steps in which the utterance is divided into chunks (*chunking*) as well as cross-modal correspondence within each chunk is established in preparation of behavior planning and execution (as explained in [8, 9]). To this end, a set of *visitors* is applied that manipulate and transform the tree while traversing it.

The first step in utterance processing is to apply a visitor that textually replaces all variables in the utterance definition by their respective values, e.g., a target location in the definition of a pointing gesture (see above). Such parameter values constitute the utterance's context and may be stated as slot-value pairs in global parameter tags at the beginning of the XML file.

In the next step, the verbal utterance tree (below the specification tag; Fig. 7) is decomposed into several subtrees according to the utterance's chunk structure as specified by the annotated time tags. For every leaf node containing data that is to be verbalized in speech, an augmented SABLE[2] tag for emphasis (EMPH) is inserted, in preparation of assigning pitch accent as well as of retrieving time information from our TTS system (see [16]) for the verbal parts.

Finally, cross-modal correspondence is established by appending all coverbal behaviors to their respective chunk trees (see Fig. 8). A gesture is considered coverbal when its timing is defined in terms of the onset and the end of a verbal affiliate, i.e., when the behavior is specified as being speech-related in a strict sense. Since we assume that a gesture supports the conveyance of the most prominent concept, which in speech carries the nucleus, the affiliate of each coverbal gesture receives a pitch accent by setting the appropriate attributes for level, duration, and intonational contour of the EMPH tag (the X-BI- prefix denotes supplementary extensions to the original SABLE tag). All remaining speech-independent behaviors are generated separately and executed in an absolutely timed fashion w.r.t. the onset of the overall utterance, i.e., the first chunk.

[2]SABLE is an international standard for marking up text input to speech synthesizers.

Ident	Symmetrie	HandLoc.	EFO and PO
Sym	equal	-	-
SymMS	symm. sag.	r-l	r-l
SymMT	symm. trans.	u-d	r-l, u-d
SymMF	symm. front.	f-b	r-l, f-b
SymMST	sag., trans.	r-l, u-d	r-l, u-d
SymMSF	sag., front.	r-l, f-b	r-l, f-b
SymMTF	trans., front.	u-d, f-b	r-l, u-d, f-b
SymMSTF	sag., trans., front.	r-l, u-d, f-b	r-l, u-d, f-b

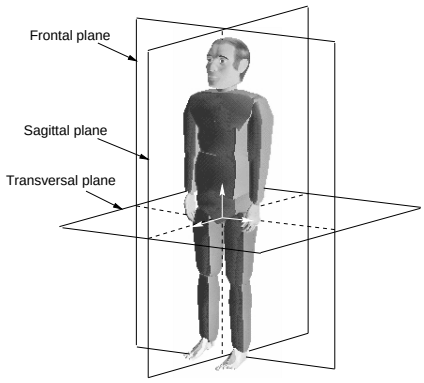


Figure 6: Effect of different symmetry constraints w.r.t. the frontal, transversal, and sagittal plane.

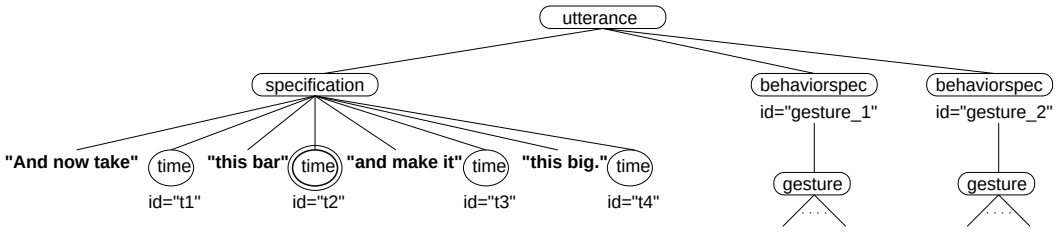


Figure 7: Tree structure of the utterance in Fig. 3.

5. CONCLUSION AND FURTHER WORK

In this paper, the MURML notation system was presented. It combines the represention of cross-modal relationships within complex multimodal utterances with flexible means for describing particulary gestural behaviors. In addition, MURML descriptions can be augmented by facial animations as well as arbitrary body movements. However, such behaviors have to be defined as timed keyframe animations in joint angles or face muscle values. A next step in our work could be to define descriptions of facial expressions on a higher level of abstraction, probably using action units as suggested by Ekman and Friesen [4] in their FACS-System. In specifying gestures in MURML, one has the options to either describe spatiotemporal features of the gesture or state a required communicative function. Ultimately, we want to create a description system that allows to specify utterances for conversational agents on different levels of abstraction in a coherent way, enabling an automatic breakdown of higher level into lower levels by modifying an XML tree.

In our lab, research on gesture *recognition* pursues the same approach. Starting from empirical studies about the use of gesture and speech in our application scenario [15], form features of the input data are extracted and transformed into a similar HamNoSys-based notation system (currently restricted to static hand constraints). In ongoing work, gesture perception and generation are integrated such that symmetric representations on both sides of the cognitive processing in communication are used. One particular goal is the realization of imitation games between user and agent based on feature-based representations of only the mandatory gesture features in order to evaluate the usefulness and efficiency of our approach. In this work, the representation of multimodal utterances on different levels of abstraction can bridge between the interpretation of input signals and the generation of believable output.

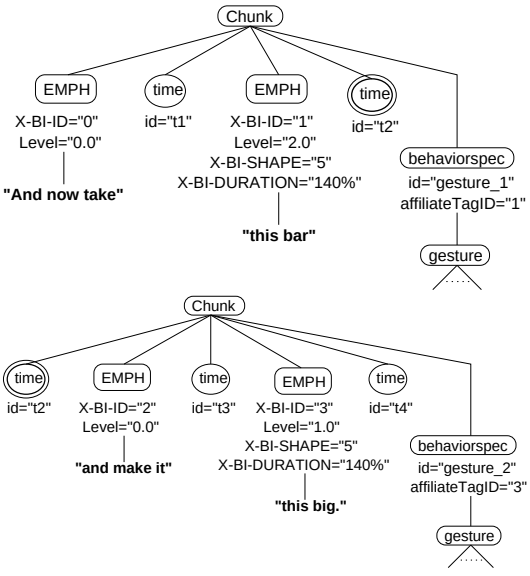


Figure 8: Trees after chunking.

6. ACKNOWLEDGMENT

This research is partially supported by the Deutsche Forschungsgesellschaft (DFG) in the Collaborative Research Center "Situated Artificial Communicators"(SFB 360).

7. REFERENCES

[1] Project page of the Articulated Communicator/Max. *http://www.techfak.uni-bielefeld.de/~skopp/max.html*.

[2] J. Cassell. More Than Just Another Pretty Face: Embodied Conversational Interface Agents. *Communications of the ACM*, 43(4):70–78, 2000.

[3] J. Cassell, H. Vilhjalmsson, and T. Bickmore. BEAT: the Behavior Expression Animation Toolkit. In *SIGGRAPH '01*, 2001.

[4] P. Ekman and W. Friesen. *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press, Palo Alto, Calif., 1978.

[5] S. Gibet, T. Lebourque, and P.-F. Marteau. High-level Specification and Animation of Communicative Gestures. *Journal of Visual Languages and Computing*, 12:657–687, 2001.

[6] A. Kendon. Gesticulation and speech: Two aspects of the process of utterance. In M. Key, editor, *The Relationship of Verbal and Nonverbal Communication*, pages 207–227. The Hague, Mouton, 1980.

[7] R. Kennaway. Synthetic Animation of Deaf Signing Gestures. In *Proceedings of the International Gesture Workshop, GW2001, London, UK, April 2001*, volume 2298 of *LNAI*, pages 146–157. Springer-Verlag, 2002.

[8] S. Kopp and I. Wachsmuth. A Knowledge-based Approach for Lifelike Gesture Animation. In W. Horn, editor, *ECAI 2000 - Proceedings of the 14th European Conference on Artificial Intelligence*, pages 663–667, Amsterdam, 2000. IOS Press.

[9] S. Kopp and I. Wachsmuth. Model-based Animation of Coverbal Gesture. In *Proceedings of Computer Animations 2002*. IEEE Press, 2002. to appear.

[10] W. Levelt. *Speaking*. MIT press, Cambridge, Massachusetts, 1989.

[11] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.

[12] T. Noma and N. Badler. A Virtual Human Presenter. In *Proceedings of the IJCAI Workshop on Animated Interface Agents: Making Them Intelligent.*, pages 45–51, 1997.

[13] S. Prillwitz. *HamNoSys. Version 2. Hamburger Notationssystem für Gebärdensprachen. Eine Einführung*. SIGNUM-Verlag, 1989.

[14] T. Sowa, S. Kopp, and M. Latoschik. A Communicative Mediator in a Virtual Environment: Processing of Multimodal Input and Output. In *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue, Verona, Italy, 2001*, pages 71–74, 2001.

[15] T. Sowa and I. Wachsmuth. Interpretation of Shape-Related Iconic Gestures in Virtual Environments. In I. Wachsmuth and T. Sowa, editors, *Proceedings of the International Gesture Workshop, GW2001, London, UK, April 2001*, volume 2298 of *Lecture Notes in Artificial Intelligence*, pages 21–33. Springer-Verlag, 2002.

[16] I. Wachsmuth and S. Kopp. Lifelike Gesture Synthesis and Timing for Conversational Agents. In I. Wachsmuth and T. Sowa, editors, *Proceedings of the International Gesture Workshop, GW2001, London, UK, April 2001*, volume 2298 of *Lecture Notes in Artificial Intelligence*, pages 120–133. Springer-Verlag, 2002.