Increasing the Expressiveness of Virtual Agents– Autonomous Generation of Speech and Gesture for Spatial Description Tasks

Kirsten Bergmann
Sociable Agents Group, CITEC
Bielefeld University
P.O. 100 131, D-33615 Bielefeld
kbergman@techfak.uni-bielefeld.de

Stefan Kopp
Sociable Agents Group, CITEC
Bielefeld University
P.O. 100 131, D-33615 Bielefeld
skopp@techfak.uni-bielefeld.de

ABSTRACT

Embodied conversational agents are required to be able to express themselves convincingly and autonomously. Based on an empirial study on spatial descriptions of landmarks in direction-giving, we present a model that allows virtual agents to automatically generate, i.e., select the content and derive the form of coordinated language and iconic gestures. Our model simulates the interplay between these two modes of expressiveness on two levels. First, two kinds of knowledge representation (propositional and imagistic) are utilized to capture the modality-specific contents and processes of content planning. Second, specific planners are integrated to carry out the formulation of concrete verbal and gestural behavior. A probabilistic approach to gesture formulation is presented that incorporates multiple contextual factors as well as idiosyncratic patterns in the mapping of visuo-spatial referent properties onto gesture morphology. Results from a prototype implementation are described.

Categories and Subject Descriptors

I.2.0 [**Artificial Intelligence**]: General—*Cognitive Simulation*; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*Natural Language Interfaces*; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Intelligent Agents*; D.2.2 [**Software Engineering**]: Design Tools and Techniques—*User Interfaces*

General Terms

Design, Experimentation, Theory

Keywords

Gesture, language, expressiveness, multimodal output, embodied conversational agents

1. INTRODUCTION

One key issue in the endowment of virtual agents with human-like expressiveness, i.e., richness and versatility, is the autonomous generation of language and accompanying gestures. Current literature on gesture research states that the question "why different gestures take the particular physical form they do is one of the most important yet largely unaddressed questions in gesture research" [1, p. 499]. This holds especially for iconic gestures, for which information has to be mapped from some mental image into (at least partly) resembling gestural form. This transformation is neither direct nor straightforward but involves a number of issues like the composability of a suitable linguistic context, the choice of gestural representation technique (e.g., placing, drawing etc.), or the low-level choices of morphological features such as handshape or movement trajectory.

In this paper, we present a novel approach to generating coordinated speech and iconic gestures in virtual agents. It comprises an architecture that simulates the interplay between these two modes of expressiveness on two levels. First, two kinds of knowledge representation–propositional and imagistic–are utilized to capture the modality-specific contents and processes of content planning (i.e., what to convey). Second, specific planners are integrated to carry out the formulation of concrete verbal and gestural behavior (i.e., how to convey it best). The overall interplay of these modules is modeled as a multi-agent cooperation process in order to meet the low latency and realtime requirements that hold for behavior generation in interactive agents.

In the following, we will put special focus on the above-described puzzle of gesture formulation. After discussing related work in the following section, we report in Section 3 on an empirical study on spontaneous speech and gesture use in VR direction-giving. Its results indicate that a model for autonomous generation of expressive gestures must take into account both, *inter*-personal commonalities in terms of contextual factors constraining the involved decisions [11], and *intra*-personal systematics as apparent in idiosyncratic gesture patterns. We thus present in Section 4, after introducing the overall architecture of our framework, a simulation account going beyond recent systems that either derive iconic gestures from systematic meaning-form mappings (e.g. [12]), or model the individual gesturing patterns of specific speakers (e.g., [16]). Based on data from our annotated corpus, we employ machine learning and adaptation algorithms to build Bayesian networks that allow to model both kinds of constraining factors. Finally, Section 5 presents results from an application of an implementation of our model in a spatial-description task domain.

Cite as: Increasing the Expressiveness of Virtual Agents– Autonomous Generation of Speech and Gesture for Spatial Description Tasks, Kirsten Bergmann, Stefan Kopp, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009),* Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. 361–368

2. RELATED WORK

The computational modeling of multimodal communicative behavior and its simulations with embodied (virtual or robotic) agents is still a relatively unexplored field. The Behavior Expression Animation Toolkit BEAT [3] employed a set of behavior generators to select and schedule conversational behaviors like hand gesture, head nod, or pitch accents for a given text. Relying on empirical results, specialized generators supplemented the verbal description of actions as well as of rhematic object features with gestures drawn from a lexicon. REA [2] extended this to generation of context-appropriate natural language and gesture by employing a natural language generator (SPUD [21]). Specific constituents were needed to coordinate linguistic expressions and iconic gestures with each other in meaning, and with the discourse context within which they occur. Gestures were hence lexicalized like words, selected using a lexical choice algorithm and incorporated directly into sentence planning.

Huenerfauth [8] proposed an approach to translate English texts from a spatial domain into American Sign Language (ASL). The system turns the text into a 3D model of how the described objects are arranged and move. This model is "overlaid" onto the space in front of the ASL signer. A planning-based NLG approach selects sign templates from a library, fills in missing parameters, and builds animation events to produce them.

The NUMACK system [12] has tried to overcome the limitations of lexicon-based gesture generation by formulating the meaning-form mapping on a level of single gesture features. A gesture planner (GP) composed morphological features (hand shape or movement trajectory) according to how they would depict features of visuo-spatial meaning (so-called IDFs; shapes, extent, relative locations, etc.), considered part of the current communicative intention. By exploring all possible combinations, the GP produces a set of possible gestures, each of which annotated with the IDFs it encodes. A sentence planner selects and combines them with words in order to derive multimodal utterances. NUMACK proved that gestures can be automatically assembled based on an representation of visuo-spatial meaning. However, an empirical study revealed that similarity with the referent cannot account for all occurrences of iconic gesture [12].

Another line of research adopts a stronger data-driven approach. Stone et al. [20] pick units of communicative performance from a database, while retaining the synchrony and communicative coordination that characterizes peoples' spontaneous expressiveness. Motion samples are recombined with new speech samples and blended in order to re-create coherent multimodal phrases. The range of possible utterances is naturally limited to what can be assembled out of the pre-recorded behavior. Further, combinatorical search approaches suffer from high computational costs.

Hartmann et al. [6] enabled the virtual agent Greta to process an utterance marked up with communicative functions, and to match these functions to a library of prototype gestures. Greta's gesture planner is unique in that it allows for parametric control of six dimensions of expressivity of a gesture, such as spatial extent or the amount of energy. This approach is limited to the set of prototype gestures, although it can create a range of variants on them.

Neff et al. [16] recently focused on data-driven simulation of small discourse gestures, conduit gestures and beats. Based on probabilistically modeled gesture profiles, the system takes arbitrary texts as input and produces synchronized conversational gestures in the style of a particular speaker. This approach succeeds in making a virtual character look more lively and natural. Yet, it does not account for the meaning-carrying functions of gestures.

In summary, automatizing the generation of multimodal behavior is a daunting task. Numerous systems were built and they have focused either on more or less stipulated regularities in multimodal utterances to come up with generalized production models, or on data-driven modeling and editing of a limited number of gesture patterns.

3. EMPIRICAL BASIS

There is only few empirical work that has tried to analyze the meaning-form mapping in expressive gesture. Kopp et al. [12] describe a gesture study indicating that people tend to employ their hands in particular ways (e.g., flat handshape, vertically oriented palm) in order to create gestural images of objects with particular shape properties (e.g., flat upright walls). However, the shape properties in this analysis were idealized and failed to fully account for many of the movement forms seen. A more general conclusion, though, which concurs with gesture literature, is that iconicity with particular spatial features may not be the sole driving force behind an iconic gesture.

Figure 1: Experiment setting: Stimulus presentation (left) and dialog phase with the speaker uttering "the left church has two towers" (right).

In order to investigate this question we have started a study on spatial communication dyads involving direction-giving and sight description. In this study, we employed Virtual Reality technology to exactly control the stimuli that the subjects perceive prior to their descriptions (see Fig. 1).

The goal of our study is to find out which contextual factors govern to which extent the use of words and gestures for objects descriptions. One aspect we focused on here for the first time as a candidate for inter-speaker systematics is the use of *representation techniques* [9] for transforming perceived object information into a gesture. By and large, they can be distinguished according to the following categories: (1) indexing: pointing to a position within the gesture space; (2) placing: an object is placed or set down within gesture space; (3) shaping: an object's shape is contoured or sculpted in the air; (4) drawing: the hands trace the outline of an object's shape; (5) posturing: the hands form a static configuration to stand as a model for the object itself. We suspected that this differentiation may be an important step in the formation of a gesture beyond merely establishing iconicity. Note that several representation techniques may be found in one gesture. Our empirical analyses thus aimed to test whether any factors correlate with the choice of tech-

nique. We investigated if different speakers gesture similarly depending on the dialog context as well as the visuo-spatial properties of possibly different objects. Further, we wanted to find out the individual differences between speakers who gesture about the *same* object.

3.1 Method

We collected a dialog corpus of 25 dyads with one participant giving directions to another participant. The direction giver had a virtual ride through a VR town out of simplified 3D objects (see Fig. 1). After finishing the ride, the direction giver had to explain the route to the follower (Fig. 1), who was to be enabled to find the same way through the virtual town and identify five important landmarks afterwards.

Audio- and videotapes were taken of each dialog. For the videotape, three synchronized camera views were recorded. All coverbal gestures have been coded according to five categories of representation technique. From the transcripts, we further coded for the task-related communicative intention of the speaker. Denis [5] developed several categories of communicative goals that can be distinguished in route directions. As we were mainly interested in object descriptions, we revised and refined it for this case into three categories: *introducing* an object without further description, *describing* an introduced object, and *positioning* an object without providing further description of it.

3.2 Results

3.2.1 Inter-subjective correlations

In our analysis we concentrated on the descriptions of the most significant part of the route, the church square (see Fig. 5), from 10 dyads with 174 speech-accompanying gestures in total. Our first analysis aimed to correlate the use of representation techniques with the spatio-geometrical properties of the objects described. It turned out that there is a significant difference between object shapes which can be decomposed into detailed subparts (whole-part relations) and objects without any subparts ($\chi^2 = 61.776, df = 4, p < 0.001$). For objects without subparts, the number of shaping and drawing gestures is increased, whereas the rate of indexing and placing gestures is decreased. For for objects which have at least one subpart, indexing and placing gestures occur more often than expected, while shaping and drawing gestures occur less often (Tab. 1). Thus, if an object is minimally complex in the sense that it has no subparts, and therefore seems more amenable to gestural reconstruction, depicting gestures are preferred. For more complex objects, with at least one sub-part, people prefer positioning gestures.

Another way to assess shape complexity of an object is in terms of its inherent symmetry: The more symmetric axes an object has, the less complex it is perceived. We analyzed the correlation between representation technique and existence of symmetry in the reference object. Again, we found a significant relationship ($\chi^2 = 79.028, df = 4, p < 0.001$): If an object has no symmetrical axis, i.e., is more complex, indexing and placing gestures are used relatively often, while drawing, shaping and modeling gestures are used less often than expected (see Tab. 1). In contrast, if an object has at least one symmetrical axis, the number of drawing, shaping and modeling gestures is increased, whereas the number of indexing and placing gestures is decreased. This is in line with the above finding, that complex objects are likely to be positioned gesturally, while less complex objects are more likely to be depicted by gesture.

A further analysis investigated the correlation of representation technique and dialog context, in particular the communicative goal. We found a significant relationship between the two variables ($\chi^2 = 81.206, df = 8, p < 0.001$). Descriptions come along with significantly more depicting gestures (shaping, drawing, posturing), while the spatial arrangement of entities is accompanied by indexing and placing gestures in the majority of cases (see Tab. 1).

3.2.2 Individual differences

Analyzing the individual differences in the use of representation techniques revealed that individuals differ significantly in the way they gesture about the same thing, and these differences concern multiple decision levels involved in the process of gesture formation. To illustrate this, we will compare the gesturing behavior of two particular individuals, speakers P5 and P16. First of all, the two individuals differ in their gesture rate: While speaker P5 performs on average 23.8 gestures per minute, speaker P16 performs 29.5 gestures per minute. In the whole corpus (N=25), gestures rates differ from a minimum of 2.34 to a maximum 32.83 gestures per minute. Second, the two speakers differ significantly in their use of representation techniques (see Tab. 2): Speaker P5 uses much more shaping and posturing gestures than P16, whereas speaker P16 uses much more placing gestures than P5. So, speaker P5 seems to use gesture to depict the entities she is talking about, while speaker P16 uses her hands rather to position entities in space. This difference is highly significant ($\chi^2 = 398.565, df = 50, N = 1919, p < 0.001$).

Table 2: Distribution of representation technique for two speakers P5 and P16.

	Speaker P5			**Speaker P16**		
Repr. Techn.	#	%	Δ%	#	%	Δ%
Indexing	52	10.9	–8,8	25	14.1	–5.6
Placing	50	10.5	–2.2	57	32.2	+19.5
Shaping	172	36.1	+5.4	49	27.7	–3.0
Drawing	11	2.3	+5.7	12	6.8	–1.2
Posturing	53	11.1	+5.6	2	1.1	–4.4
Other	138	29.0	+5.6	32	18.1	–5.3
Handshape						
ASL-B	97	20.4	–23.9	119	67.2	+22.9
ASL-C	122	25.6	+10.7	12	6.8	–8.1
ASL-G	91	19.1	+1.8	14	7.9	–9.4
ASL-O	11	2.3	+0.2	0	0.0	–2.1
ASL-5	57	12.0	+6.0	12	6.8	+0.8
Other	98	20.6	+5.2	20	11.3	–4.1

Finally, individual differences of the speakers become apparent in the morphological features of their gestures, such as the handshapes chosen (Tab. 2): Speaker P16 uses the handshape ASL-B in the majority of her gestures (67.2 %) while other handshapes are chosen in relatively equal minor parts. In contrast, speaker P5 diversifies her handshapes: ASL-B, ASL-C, and ASL-G are used more or less equally often. Again, this difference is highly significant ($\chi^2 = 614.779, df = 50, N = 1919, p < 0.001$).

Table 1: Interrelation of representation technique and visuo-spatial features of the referent (number of subparts and number of symmetrical axes). Parenthetical values are expected occurrences.

Representation	Number of subparts		Number of symmetrical axes		Communicative goal		
Technique	**0**	**1 or more**	**0**	**1-3**	**Intro.**	**Descr.**	**Pos.**
Indexing	6 (13.4) *	18 (10.6) *	18 (7.7) ***	6 (16.3) *	0 (2.5)	6 (14.8) *	18 (6.8) ***
Placing	4 (20.1) ***	32 (15.9) ***	27 (11.6) ***	9 (24.4) **	6 (3.7)	7 (22.1) **	23 (10.1***)
Shaping	45 (32.3) *	13 (25.7) *	9 (18.7) *	49 (39.3)	31 (33.0)	6 (6.0) *	3 (16.3) ***
Drawing	21 (13.4) *	3 (10.6) *	0 (7.7) **	24 (16.3) *	15 (13.7)	2 (2.5)	2 (6.8) *
Posturing	21 (17.8)	11 (14.2)	2 (10.3) **	30 (21.7)	4 (3.3)	25 (19.7)	3 (9.0) *

In conclusion we see systematic factors having an impact on which representation technique is chosen across speakers. This inter-personal systematics concerns not only the correlation between a gesture and its referent, but also the communicative intention underlying an utterance. At the same time, there is a form of intra-personal systematics, both of which must be taken into consideration by an account of why people gesture the way they actually do.

4. COMPUTATIONAL MODELING

Our goal is a model that, given a particular discourse context and a given communicative intention such as, in lay terms, "describe how object X looks like", allows virtual agents to automatically select the content and derive the form of coordinated language and iconic gestures. Looking at current psycholinguistic production models, Kita & Özyürek's [10] *Interface Hypothesis Model* (IH model, henceforth) provides inspiration in several ways. First, information packaging for iconic gestures parallels linguistic packaging [10]. A computational model, therefore, has to consider in particular the ways in which the syntactical and grammatical structures of a language are reflected in its conceptual structure, and the influence of this structure on the content of coverbal iconic gestures. Second, performing representational gestures impacts content planning and conceptualization [7] as well as the lexical choices [15] in speech. A computational model thus needs to account for how "gesture thinking" interacts with the propositional thinking that underlies speech. None of the existing systems (see Sect. 2) complies with this insight.

Our architecture (shown in Fig. 2) is inspired by, but extends and substantiates Kita & Özyürek's IH model in several ways. We distinguish four processing modules to be involved in content planning and micro-planning of speech and gesture: *Image Generator, Preverbal Message Generator, Speech Formulator*, and *Gesture Formulator*. That is, in contrast to the IH model, we adopt the idea advocated in other production models [4, 12] of two functionally separable modules, one for activating and selecting features of visuo-spatial imagery (the Image Generator) and one for turning these features into gestural form (Gesture Formulator). In addition, two dedicated modules (Motor Control and Phonation) are concerned with the realization of synchronized speech and gesture movements with the virtual human Max [13]. Further components are a discourse model and distinct long-term memories for imagery and propositional knowledge.

The overall production process is treated as a multi-agent problem solving task. All modules are modeled as software agents that operate concurrently and proactively on a central working memory, realized as a globally accessible, structured blackboard. As opposed to message-passing architectures the overall production process thus evolves by each module observing entries in the working memory, taking local action if possible, and modifying or posting entries in result. The production process is finished when all entries associated with a multimodal communicative delivery stabilize and specifications of verbal and gestural acts have been formed in working memory. In this way, interaction among the modules realizes content planning and micro-planning in an interleaved and interactive manner, and it enables bottom-up processes in both modalities.

4.1 Representation of content

Humans adopt different perspectives towards a spatial scene, describe objects at different levels of detail, and at different scales. In doing so, speakers transform, scale, or arrange their mental visuo-spatial representation with great flexibility and adaptivity, and a propositional formalism can only hardly, if at all, account for this. We thus face the necessity to account for the dichotomic nature of multimodal expressivity already at the level of mental representations. Thus, going beyond previous approaches that were based on one common propositional representation [12], we adopt a dual coding perspective for our content representation. In his dual coding theory, Paivio [17] distinguishes two functionally independent systems, verbal memory and image memory, with associative links within each memory and possible referential links across the two systems. Likewise, we propose a representational layer underlying the production of multimodal behavior that consists of three parts, (1) an imagistic description, (2) propositional knowledge, and (3) an interface in the form multimodal concepts that associate imagistic and propositional knowledge.

4.1.1 Imagistic Descriptions

To realize computational imagery for the agent, we employ an hierarchical model called *Imagistic Description Trees* (IDT) [18]. IDT has been developed based on an empirical study in order to capture the imagistic content of shape-related gestures in a gesture interpretation system. Thus it is designed to cover all decisive visuo-spatial features one finds in iconic gesture.

Each node in an IDT contains an Imagistic Description (IMD) which holds an schema representing the shape of an object or object part. Object schemas contain up to three axes representing spatial extents in terms of a numerical measure and an assignment value like "max" or "sub", classifying this axis' extent relative to the other axes. Each axis

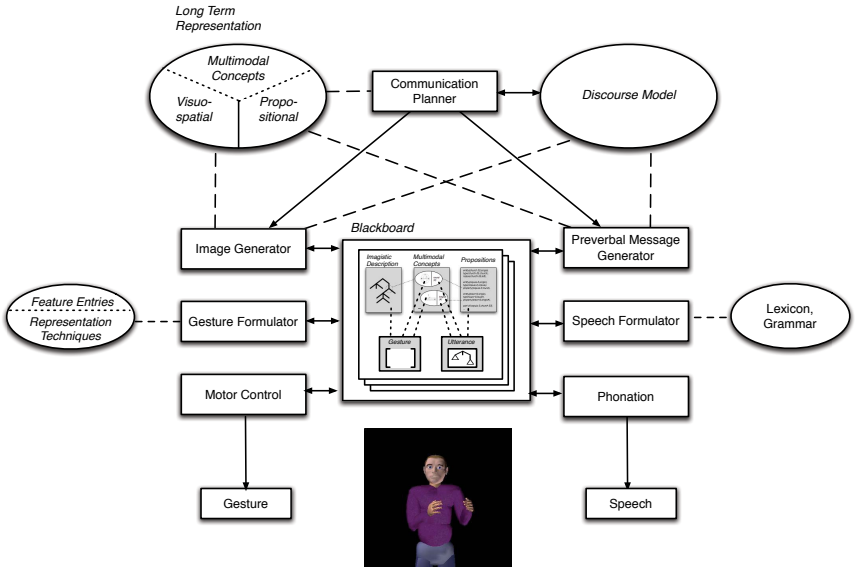

Figure 2: Overview of the architecture of the speech and gesture production model.

can cover up to three dimensions to account for rotation symmetries (becoming a so-called "integrated axis"). The boundary of an object can be defined by a profile vector that states symmetry, size, and edge properties for each object axis or pair of axes. The links in the tree structure represent spatial relations between parts and wholes, and are defined by transformation matrices. Note that the IDT model explicates those spatial features of complexity of object shape that we have found to significantly influence the choice of the gesture representation technique. Fig. 3 illustrates how imagery can be operationalized with the IDT model. We have modeled extensive parts of the virtual world from our study analogously.

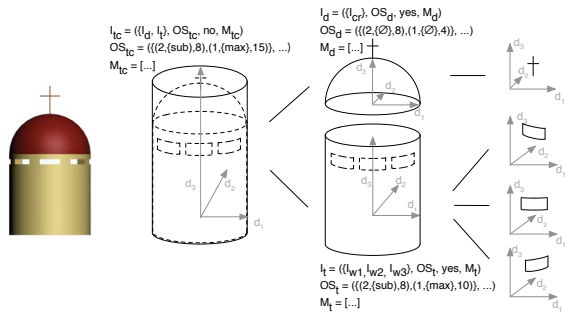

Figure 3: IDT representation of a church tower.

4.1.2 Propositional representation

Speech formulation needs to draw upon a proper representation of spatial knowledge, as well as conceptual background knowledge, about the considered entities. As common in computational approaches to language semantics, we employ a propositional format for this, i.e., logical formulae that state propositions over symbols that represent objects and relations according to a given ontology. Since we focus on object descriptions, the spatial knowledge pertains to objects (houses, streets, etc.), their properties (proper name, color, quality, size, shape etc.), and the taxonomic (is-a), partonomic (part-of) or spatial relations (on-top-of, left-of) between them. An according knowledge base has been defined for parts of the virtual world.

4.1.3 Multimodal concepts

The third part of the long-term memory are a number of abstract *multimodal concepts*, which are bindings of IDTs with corresponding propositional formulations. The imagistic part of multimodal concept is characterized by underspecification, i.e., it allows more than one interpretation. This underspecification draws the distinction between the imagistic part of a multimodal concept and the imagistic description of a concrete spatial object. For example, the property of being "longish" is represented as an underspecified IDT in which one axis dominates the other axes, as well as in terms of a logical formula"longish(?X)". As such an underspecified IDT can be matched by means of formal graph unification [18] with any other IDT, e.g. of a specific tree "tree-2", the corresponding formula is concretized by binding the variable to yield "longish(tree-2)". This mechanism anchors the underspecified semantics of an iconic gesture into linguistic context. Currently, multimodal concepts for dimensional adjectives (longish, round, tall, etc.), stereotyped object shapes (box, circle, etc.), and basic spatial relations (right-of, side-by-side, above, etc.) are defined.

4.2 Production process

4.2.1 Image and Preverbal Message Generation

The production of a chunk of multimodal object description starts upon the arrival of a message from the Communication Planner, containing a new communicative intent.

Such a communicative intent comprises a specification of the intended communicative act (e.g., introduce, describe, describe-position, describe-construction) along with the entities to be referred to or the properties to be predicated.

The Image Generator accesses the IDTs stored in long-term memory and activates the imagistic descriptions of all objects involved in the communicative goal. This activation generally leads to import into the respective working memory structure. Likewise, the Preverbal Message Generator starts by selecting knowledge from propositional long-term memory and asserts the selected facts into working memory.

For IMDs with a significantly high activation, the Image Generator performs spatial perspective taking. That is, it determines which spatial perspective to adopt towards the objects. Direction-givers usually adopt either a route (1st person) or survey (3rd person) perspective [14], with frequent switches between the two. For simplicity, we currently assume that our agent adopts the more prominent route perspective in describing an object. Still, the Image Generator has to figure out how the objects look like from the particular point of view adopted and along that particular view direction. This operation is directly implemented as a transformation on object schemas.

The Image Generator tries to map the perspective IMD onto the imagistic parts of multimodal concepts in long-term memory (see Fig. 2). If this succeeds, the corresponding multimodal concept is added to the working memory unit. Likewise, multimodal concepts are asserted when they unify with propositions selected by the Preverbal Message Generator.

4.2.2 Speech Formulation

The Speech Formulator monitors the unit on the blackboard and carries out sentence planning for each set of propositions posted by the Preverbal Message Generator. As related systems we employ SPUD [21], a grammar-based micro-planner using a Lexicalized Tree Adjoining Grammar (LTAG) to generate natural language sentences. The grammar and the lexicon have been defined from the object descriptions in our empirical data.

In contrast to the NUMACK system, we avoid extending the linguistic formalism to account for gesture integration on the level of speech formulation. Instead, we let this come about via multimodal content planning and only leave it to the Speech Formulator to lay down the necessary temporal constraints for speech-gesture synchrony at surface structure level. To this end, SPUD delivers back information on the semantics of each linguistic constituent. The multimodal concepts are utilized to find the words that are maximally co-expressive with each gesture the Gesture Formulator has proposed. The derived temporal constraints then mark the onset and end of the lexical affiliate on word level and are asserted to the blackboard unit, too.

4.2.3 Gesture Formulation

The Gesture Formulator has to compose and specify the morphology of a gesture in terms of a typed attribute-value matrix. For this purpose, it can take as input all IMDs comprised by the working memory unit but also has access to all other information on the blackboard.

The results of our corpus analyses suggest both intra- and inter-personal systematics in gesture generation. We employ a Bayesian network to model those choices in the production

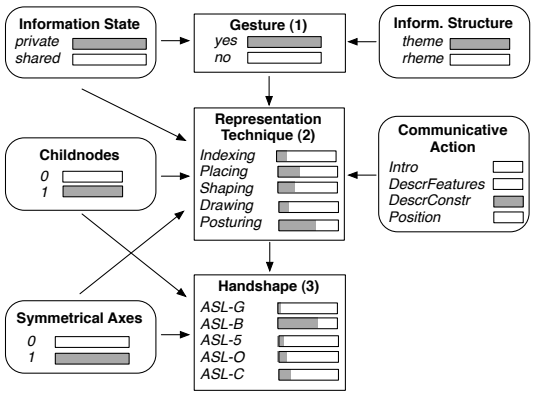

Figure 4: Bayesian network built from data of one speaker, allowing to decide (1) if the production of a gesture is adequate, (2) which representation technique to perform, and (3) which handshape to use.

process which we found to be highly idiosyncratic. These include, first, the question if a gesture will be produced to refer to a particular entity, second, which representation technique to perform, and finally, the handshape to be used if it correlates with the technique. In order to find the most likely causal influences between these features, we learned the structure of the Bayesian network from our corpus data using the Necessary Path Condition (NPC) algorithm [19]. The NPC algorithm is a constraint-based structure learning algorithm that identifies the structure of the underlying graph by performing a set of statistical tests for pairwise independence between each pair of variables. In other words, the independence of any pair of variables given any subset of other variables is tested. Once the structure of the network has been found, its maximum likelihood estimates of parameters are computed. The resulting Bayesian network is shown in Fig. 4. It is built from annotated data of one particular speaker (P5) who gestured a lot, i.e., we have learned an individual speaker model explicating the idiosyncratic patterns that need to be followed in behavior modeling.

Using the Bayesian network, the Gesture Formulator decides if the production of a gesture is adequate, which representation technique to perform, and which handshape to use given the evidence about the communicative context at hand (see Fig. 4). One result is the most probable gesture technique, for which a template is selected from an exhaustive set. This template provides a partially filled gesture form feature matrix, along with specific procedures to determine and fill the remaining open slots in order to maximize iconicity between the visuo-spatial shape features provided by the IDT model and gesture morphology. This includes mapping the IMD into the gesture space coordinate frame to define the gesture position, choosing an adequate palm orientation, or (if not implied by the Bayesian net) selecting the most resembling handshape. This "iconic mapping" employs, again, unification of the referent IMD with IMDs predefined for prominent handshapes.

The process of gesture formulation results in a matrix of form features as proposed by [12]: handshape, wrist location, palm direction, extended finger direction, movement

trajectory and direction. These features lay down the gestural movements and can take on either symbolic or numerical values. For realtime surface realization of such a gesture specification, as well as of a verbal utterance, we employ the *ACE* realization engine [13] which turns them into synthetic speech and synchronized gesture animations.

5. RESULTS FROM A FIRST APPLICATION

A first prototype of the previously described computational model has been realized using our own multi-agent system toolkit, a Prolog implementation of SPUD, the Hugin toolkit for Bayesian inference, and the ACE realization engine. This prototype was employed in a first application scenario during an open house day in our lab. In this application, about 150 visitors of the lab could engage in a game with the virtual human Max, where Max explains multimodally a building of the virtual environment. Being equipped with proper knowledge sources, i.e., communicative plans, lexicon, grammar, propositional and imagistic knowledge about the world, Max randomly picks a landmark and a certain perspective towards it, and then creates his explanations autonomously. The system achieves near real-time performance. In the following we illustrate the production process of a particular multimodal utterance step by step and give generation results.

Let us assume the communicative goal being "describe-construction churchtower-1 roof-3". The IDT nodes in Max's imagery labeled "churchtower-1" and "roof-3" get activated. This activation propagates down the tree so that each child node gets half of its parent's activation (Fig. 5). The activated part of the IDT is imported into working memory as well as a set of significantly activated propositions, which contain "churchtower-1" or "roof-3" as argument (high activation) or are related to one of the referents, e.g., by a part-of relation (less highly activated).

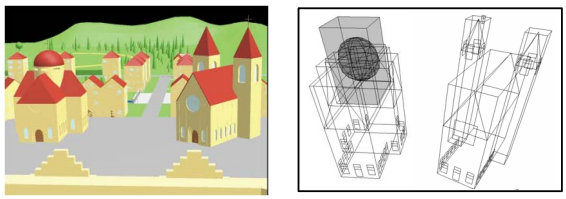

Figure 5: A churchsquare in the virtual world (left) and schematic of the corresponding IDT content (right); activated parts are marked.

Let us further assume that Max has chosen a point of view on the church square between the two churches, facing the left church at a distance of 10m. Each IMD is transformed into the adopted perspective, using a standard view transformation as in computer graphics, and all IDT nodes which are occluded are culled. The resultant IDMs of churchtower-1 and roof-3 are written back to the blackboard, upon which the Image Generator tries to match them with imagistic parts of any multimodal concept defined in long term memory. In our example, the IMD of roof-3 unifies with the concept "round", which is therefore added to the working memory. The variable in the propositional part of the multimodal concept gets bound to "roof-3".

Now the formulator modules take action and the production process hence enters the microplanning stage. The Speech Formulator lets the SPUD system search for an adequate verbalization of the propositions on the blackboard. Figure 6 shows the resulting LTAG tree generated in our case; for illustration, each surface element is annotated with information about the propositional meaning it conveys.

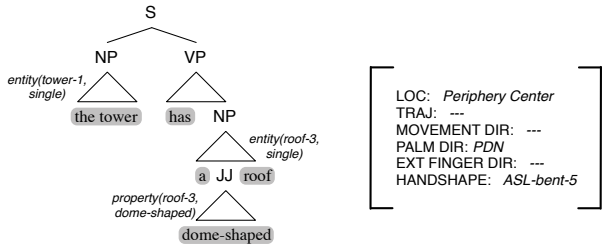

Figure 6: LTAG tree generated for the utterance "the tower has a dome-shaped roof" (left) and gesture form feature matrix generated by the Gesture Formulator (right).

The Gesture Formulator runs once for churchtower-1 and once for roof-3. All facts available are entered into and propagated through the Bayesian network learned beforehand from our empirical data. We assume churchtower-1 to be already introduced into discourse, thus having the information state "shared". The corresponding nominal phrase "the tower" hence fulfills the role of the sentence's theme. These facts are entered into the network and result in a likelihood of .17 for producing a gesture. Consequently, the gesture Formulator stops the generation of the gesture and goes on with the second entity, roof-3. Its information state is "private", since it has not been introduced yet, and the referring expression "a dome-shaped roof" is the sentence's rheme. Propagated through the network, the resulting likelihood to produce a gesture is 1.0. The Gesture Generator enters further facts: the communicative action (describe-construction), the number of childnodes (1) and the entity's symmetry property (round with one symmetrical axis). What results is the highest probability of .45 for a posturing gesture. The Formulator now chooses the template for posturing to be filled. Since posturing is a technique that depicts mainly by its handshape, this slot is filled via an iconic mapping. Here, the IMD for the ASL-bent-5 handshape is found most similar to the referent IMD. Likewise, the gesture location is inquired from the IMD as well as the palm orientation towards the object's main plane. The resultant feature matrix for the gesture is given in Figure 6. Finally, the verbal utterance and the gesture's feature matrix are translated into a MURML description, which is then sent for realtime realization to the virtual agent Max. Figure 7, bottom-left, shows a screenshot of the resulting gesture.

So far, we have trained the Bayesian network for Gesture Formulation only from the empirically observed behavior of subject P5 for the descriptions of four different landmarks in the virtual world. As can be seen from the comparison in Figure 7, our system succeeds in reproducing similar gestural expressions in corresponding contexts.

6. CONCLUSION

In this paper we have presented a new approach to generating coordinated speech and iconic gestures in virtual agents. The necessary interplay between these two modes

Figure 7: Examples of speaker P5 from our corpus and similar movements simulated with our system.

of expressiveness is modeled, first, at the level of two kinds of knowledge representations and processes and, second, between specific planners formulating modality-specific behavior. We have put special emphasis on the question how to bridge the gap between the imagistic representation of an object and its gestural counterpart, whereby we account for empirical findings by considering both, systematic *inter*personal factors as well as idiosyncratic gesturing patterns.

We are confident that our approach is a step forward towards increased expressiveness of virtual agents. We will extend the degree of interaction in our implementation both, horizontally, of two modes of thinking and, vertically, of the two levels of planning. Further, we are continuing to generalize across further speakers, and to employ adaptation algorithms to detect clusters of speakers who share particular interrelations between causal variables and observable behavior. This will enable us to distinguish between and incorporate in our model different degrees of inter- and intrapersonal systematics in gesture production. Finally, a more thorough evaluation of our system will show how far our simulation approximates human multimodal behavior.

7. ACKNOWLEDGMENTS

This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 “Alignment in Communication” as well as the Center of Excellence in “Cognitive Interaction Technology” (CITEC).

8. REFERENCES

[1] J. Bavelas, J. Gerwing, C. Sutton, and D. Prevost. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58:495–520, 2008.

[2] J. Cassell, M. Stone, and H. Yan. Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of the First International Conference on Natural Language Generation*, 2000.

[3] J. Cassell, H. Vilhjalmsson, and T. Bickmore. Beat: The behavior expression animation toolkit. In *Proceedings of SIGGRAPH*, 2001.

[4] J. de Ruiter. The production of gesture and speech. In D. McNeill, editor, *Language and gesture*. Cambridge University Press, 2000.

[5] M. Denis. The description of routes: A cognitive approach to the production of spatial discourse. *Current Psychology of Cognition*, 16:409–458, 1997.

[6] B. Hartmann, M. Mancini, and C. Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In *Gesture in Human-Computer Interaction and Simulation*, 2005.

[7] A. Hostetter, M. Alibali, and S. Kita. Does sitting on your hands make you bite your tongue? The effects of gesture inhibition on speech during motor descriptions. In D. S. McNamara and J. G. Trafton, editors, *Proc. 29th meeting of the Cognitive Science Society*, pages 1097–1102. Erlbaum, 2007.

[8] M. Huenerfauth, L. Zhou, E. Gu, and J. Allbeck. Design and evaluation of an american sign language generator. In *Proceedings of the Workshop on Embodied Language Processing*, 2007.

[9] A. Kendon. *Gesture–Visible Action as Utterance*. Cambridge University Press, 2004.

[10] S. Kita and A. Özyürek. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48:16–32, 2003.

[11] S. Kopp, K. Bergmann, and I. Wachsmuth. Multimodal communication from multimodal thinking - towards an integrated model of speech and gesture production. *International Journal of Semantic Computing*, 2(1):115–136, 2008.

[12] S. Kopp, P. Tepper, K. Ferriman, K. Striegnitz, and J. Cassell. Trading spaces: How humans and humanoids use speech and gesture to give directions. In T. Nishida, editor, *Conversational Informatics*, chapter 8, pages 133–160. John Wiley, 2007.

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

[14] S. Levinson. Frames of reference and molyneux’s question: Cross-linguistic evidence. In *Space and Language*, pages 109–169. MIT Press, 1996.

[15] E. Morsella and R. Krauss. The role of gestures in spatial working memory and speech. *The American Journal of Psychology*, 117:411–424, 2004.

[16] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 27/1:1–24, 2008.

[17] A. Paivio. *Mental Representations*. Oxford Univ. Press, 1986.

[18] T. Sowa and I. Wachsmuth. A model for the representation and processing of shape in coverbal iconic gestures. In *Proc. KogWis05*, pages 183–188, Basel, 2005. Schwabe.

[19] H. Steck and V. Tresp. Bayesian belief networks for data mining. In *Proceedings of the 2nd Workshop “Data Mining und Data Warehousing als Grundlage moderner enscheidungsunterstuetzender System”*, 1999.

[20] M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler. Speaking with hands: Creating animated conversational characters from recordings of human performance. In *Proceedings of SIGGRAPH*, pages 506–513, 2004.

[21] M. Stone, C. Doran, B. Webber, T. Bleam, and M. Palmer. Microplanning with Communicative Intentions: The Spud System. *Comput. Intelligence*, 19(4):311–381, 2003.