Modeling the Production of Co-Verbal Iconic Gestures by Learning Bayesian Decision Networks

Kirsten Bergmann and Stefan Kopp

Sociable Agents Group, CITEC, Bielefeld University

SFB 673 “Alignment in Communication”, Bielefeld University

P.O. Box 100 131, D-33615 Bielefeld, Germany

{kbergman,skopp}@techfak.uni-bielefeld.de

Abstract. Expressing spatial information with iconic gestures is abundant in human communication and requires to transform information about a referent into resembling gestural form. This transformation is barely understood and hard to model for expressive virtual agents as it is influenced by the visuo-spatial features of the referent, the overall discourse context or concomitant speech, and its outcome varies considerably across different speakers. We employ Bayesian Decision Networks (BDN) to achieve such a model. Different machine learning techniques are applied to a data corpus of speech and gesture use in a spatial domain to investigate how to learn such networks. Modeling results from an implemented generation system are presented and evaluated against the original corpus data to find out how BDNs can be applied to human gesture formation and which structure learning algorithm performs best.

1 Introduction

The use of gestures is a ubiquitous characteristic of human-human communication, especially when spatial information is expressed. It is therefore desirable to endow conversational agents with similar gestural expressiveness and flexibility to improve the interaction between humans and machines. *Iconic gestures* are, in contrast to other gesture types (McNeill, 2005), characterized by the fact that they depict some feature of the referent by the gesture’s form. They are produced in combination with speech, a feature distinguishing iconics from pantomime or emblems. In this paper, we focus particularly on iconics used in object descriptions. Modeling the production of such iconic gestures is an ambitious objective, since, as de Ruiter (2007, p. 30) recently put it; “the problem of generating an overt gesture from an abstract [...] representation is one of the great puzzles of human gesture, and has received little attention in the literature". The intricacy of this problem is due to the fact that iconic gestures, in contrast to language or other gesture types such as emblems, have no conventionalized form-meaning mapping. Apparently, iconic gestures communicate through iconicity, i.e., their physical form corresponds with object features such as shape or spatial properties. Empirical studies, however, reveal that similarity with the referent cannot fully account for all occurrences of iconic gesture use (Streeck, 2008). Recent findings actually indicate that a gesture's form is also influenced by specific contextual constraints and the use of more general gestural representation techniques. These techniques, e.g., shaping or drawing, further sub-divide the class of iconic gestures (Müller, 1998; Kendon, 2004; Streeck, 2008).

In addition, human beings with their personal and cultural backgrounds are quite distinct and inter-subjective differences in gesturing are quite obvious (cf. Hostetter & Alibali (2007)). Consider for instance gesture frequency: while some people rarely make use of their hands while speaking, others gesture almost without interruption. Similarly, individual variation becomes apparent in preferences for particular representation techniques or the choices of morphological features such as handshape or handedness (Bergmann & Kopp, 2009b). See Figure 1 for some examples of how people perform different gestures in referring to the same entity, a u-shaped building. The speakers differ, first, in their use of representation techniques. While some speakers perform drawing gestures (the hands trace the outline the referent), others perform shaping gestures (the referent's shape is sculpted in the air). Second, gestures vary in their morphological features even when speakers use the same representation technique: drawing gestures are performed either with both hands (P1) or with one hand (P8), while the shaping gestures are performed with differing handshapes (P7 and P15).

——————————— INSERT FIGURE 1 HERE ———————————

Taken together, iconic gesture generation on the one hand generalizes across individuals to a certain degree, while on the other hand, inter-subjective differences must also be taken into consideration by an account of why people gesture the way they actually do. In previous work we developed an overall production architecture for multimodal utterances (Bergmann & Kopp, 2009b). Integrated into this architecture is *GNetIc* (Bergmann & Kopp, 2009a), a gesture net specialized for iconic gestures. Combining model-based generation techniques as well as data-driven methods GNetIc allows to derive the form of speech-accompanying iconic gestures for virtual agents. Individual as well as general networks are learned from annotated corpora and supplemented with rule-based decision making. The model allows for a speaker-specific gesture production which is not only driven by iconicity, but also by the overall discourse context.

This paper extends and details the GNetIc approach by investigating the task how to best learn network structures from corpus data. We describe the application and evaluation of different state-of-the-art methods to learn and build those networks.

In the following, we survey existing approaches to model gesture generation (Section 2) and describe our empirical basis (Section 3). In Section 4 we present our integrated approach of iconic gesture generation (GNetIc). The overall generation process and modeling results from a prototype implementation are given in Section 5. A comparison of several machine learning techniques applied to our data and an evaluation of the gesturing behavior generated with GNetIc is presented in Section 6. We conclude with a discussion of our results (Section 7).

2 Related Work

Work on the generation of speech-accompanying iconic gestures and its simulation with virtual agents is still relatively sparse. The first systems investigating this challenge were lexicon-based approaches (Cassell et al., 2000). Relying on empirical results, these systems focus on the context-dependent coordination of gestures with concurrent speech, whereby gestures are drawn from a lexicon. Flexibility and generative power of gestures to express new content, therefore, is obviously very limited. A different attempt that is closely related to the generation of speech-accompanying gestures in a spatial domain is Huenerfauth's system which translates English texts into American Sign Language (ASL) and focuses on classifier predicates which are complex and descriptive types of ASL sentences (Huenerfauth, 2008). These classifier predicates have several similarities with iconic gestures accompanying speech. The system also relies on a library of prototypical templates for each type of classifier predicates in which missing parameters are filled in adaptation to the particular context. The NUMACK system (Kopp et al., 2004) tries to overcome the limitations of lexicon-based gesture generation by considering patterns of human gesture composition. Based on empirical results, referent features are linked to morphological gesture features by an intermediate level of image description features. However, as shown empirically, iconic gesture use is not solely driven by similarity with the referent (Kopp et al., 2007). In contrast, in our approach we additionally consider the discourse context and also the use of different gestural representation techniques which are found to have an impact on gesture production in humans.

The research reviewed so far is devoted to build general models of gesture use, i.e., systematic inter-personal patterns of gesture use are incorporated exclusively. What is not considered in these systems is individual variation which is investigated by another line of relatively recent research. Ruttkay (2007) aims at endowing virtual humans with unique style in order to appear prototypical for some social or ethnical group. Different styles are defined in a dictionary of meaning-to-gesture mappings with optional modifying parameters to specify the characteristics of a gesture. The focus of this work is a markup language to define different aspects of style which are handcrafted and model the behavior of stereotypic groups instead of individuals. In a similar account, Hartmann et al. (2006) investigate the modification of gestures to carry a desired expressive content while retaining their original semantics. Bodily expressivity is defined with a small set of dimensions such as spatial/temporal extent, fluidity, or power, which are used to modify gestures. Similar to Ruttkay (2007), the focus of this approach is a framework to individualize gesture style.

Another line of research uses data-driven methods to simulate individual speakers' gesturing behavior. Stone et al. (2004) recombine motion captured pieces with new speech samples to recreate coherent multimodal utterances. Units of communicative performance are re-arranged while retaining temporal synchrony and communicative coordination that characterizes peoples spontaneous delivery. The range of possible utterances is naturally limited to what can be assembled out of the pre-recorded behavior. Neff et al. (2008) aim at generating character-specific gesture style capturing the individual differences of human speakers. Based on statistical gesture profiles learned from annotated multimodal behavior, the system takes arbitrary texts as input and produces synchronized conversational gestures in the style of a particular speaker. The resulting gesture animations succeed in making a virtual character look more lively and natural and have empirically been shown to be consistent with a given performer's style. The approach, however, does not need to account for the fundamental meaning-carrying functions of gestures since it focuses on discourse gestures and beats.

In summary, previous research has either emphasized general patterns in the formation of iconic gestures, or concentrated on individual gesturing patterns. In order to shed light onto the question how a gesture for a particular referent is shaped in a particular discourse situation, we present a modeling approach going beyond previous systems by accounting for both, systematic commonalities across speakers and idiosyncratic patterns of the current individual speaker. In this model we apply machine learning techniques which are increasingly used for modeling behavior of virtual agents (Morency et al., 2010), but have not been applied to the problem of iconic gesture formulation, yet.

3 Empirical Basis

Building a computational model from empirical data requires an adequate and comprehensively annotated corpus. To build such a corpus of multimodal behavior we conducted a study on spontaneous speech and gesture use in direction-giving and landmark descriptions. This data collection, the *Bielefeld Speech and Gesture Alignment* (SaGA) corpus, contains 25 dialogs and ∼5000 iconic/deictic gestures (Lücking et al., 2010). An example to illustrate the kind of communicative behavior we are dealing with is given in Table 1.

———————————— INSERT TABLE 1 HERE ————————————

In the work reported here, we concentrate on descriptions of four landmarks from 5 dyads (473 noun phrases, 288 gestures). That is, all five speakers referred to the same entities in the selected data cases. We transcribed the spoken words and coded further information about the dialog context (for details see Lücking et al., 2010; Bergmann & Kopp, 2010). All coverbal gestures have been segmented and coded for their representation technique. According to our focus on object descriptions we distinguish the following five 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; and (5) *posturing*: the hands form a static configuration to stand as a model for the object itself. In addition, each gesture has been coded for its morphology in terms of handedness, handshape, hand position, palm and finger orientation, and movement features.

Further, each gesture used in the object descriptions has been coded for its referent and some of the spatio-geometrical properties of this referent. These object features are drawn from an imagistic representation we built from the VR stimulus of the study (e.g., houses, trees, streets). This hierarchical representation is called *Imagistic Description Trees* (IDT), and was originally designed to cover all decisive visuo-spatial features of objects one needs in iconic gesture interpretation (Sowa & Wachsmuth, 2005). Each node in an IDT holds a schema representing the shape of an object or object part in terms of its main extents and symmetries. Features extracted from this representation in order to capture the main characteristics of a gesture's referent are (1) whether an object can be decomposed into detailed subparts (whole-part relations), (2) whether it has any symmetrical axes, (3) its main axis (4), its position in the VR stimulus, and (5) its shape properties. The transcription of the interlocutor's words is enriched with further information about the overall discourse context in terms of thematization (theme, rheme), information state (private, shared) and communicative goal in the domain of object descriptions. In Table 2 the complete annotation scheme is summarized.

——————————————— INSERT TABLE 2 HERE ———————————————

As reported in (Bergmann & Kopp, 2009b) individuals differ significantly in their gestural behavior. As concerns the question whether or not a gesture is produced for a particular object (part), gesture rates differ from a minimum of 2.34 to a maximum of 32.83 gestures per minute in our whole corpus (N=25). The mean gesture rate is 15.64 gestures per minute (SD=7.06). For the five dyads which are analyzed in detail here, the gesture rates vary between 12.5 to 25.0 gestures per minute. Another example illustrating how speakers differ inter-subjectively concerns the question which hand(s) to use when referring to an entity. The general distribution of handedness in the five dyads is as follows: With 56.6% the majority of gestures is performed two-handed, while right-handed gestures occur in 28.6% of the cases and left-handed gestures in 14.8%. As shown in Figure 2 this distribution is not shared by all speakers. Similarly, the use of representation techniques and handshapes is rather idiosyncratic, whereas other gesture features are subject to inter-subjective patterns as reported, e.g., in (Kopp et al., 2007). These commonalities concern particularly morphological gesture features such as hand orientation and movement features accounting for the similarity between referent shape and gesture form.

——————————————— INSERT FIGURE 2 HERE ———————————————

4 The GNetIc Gesture Generation Approach

We have proposed GNetIc (Gesture Net for Iconic Gestures), an approach to tackle the challenge of considering both, general and individual patterns in gesture formulation using Bayesian decision networks (BDNs; cf. Bergmann & Kopp, 2009a). Decision networks (also known as influence diagrams) supplement standard Bayesian networks by decision nodes (Howard & Matheson, 2005). The formalism has been proven useful in the simulation of human behavior, e.g., Yu & Terzopoulos (2007) used decision networks to simulate social interactions between pedestrians in urban settings. Decision networks are suited for our purpose since they provide a representation of a finite sequential decision problem, combining probabilistic and rule-based decision-making. Each decision to be made in the process of gesture generation, e.g., whether or not a speech-accompanying gesture will be planned or which representation technique will be used, can be represented in the network either as a *decision node* or as a *chance node* with a specific probability distribution. Chance nodes represent random variables each of which having a conditional probability table specifying the probability of the variable having a particular value given a combination of values of its parent nodes. Decision nodes represent decisions to be made. They can have both chance nodes and other decision nodes as parents indicating that the decision has to be made at a particular point in time, i.e., when all necessary evidence is available. The links between any of the nodes explicitly represent the relations between causes and effects, i.e., (probabilistic) relationships among variables are encoded in the network.

4.1 BDNs for Gesture Generation

Applied to our task of gesture generation, employing a probabilistic approach like BDNs is advantageous for a number of reasons. First, networks can be learned either from the annotated data of single speakers or from a larger corpus containing data from several speakers. This allows for observing inter-subjective differences not only at the level of surface behavior, but also to detect differences in the underlying generation strategies. Second, there are no minimum sample sizes required to perform the analysis. It could be shown that Bayesian networks can have good prediction accuracy even with rather small sample sizes (Kontkanen et al., 1997). Third, a network can be easily extended by introducing further variables, either annotated in the corpus, or inferred from that data. Third, the same network can be used to calculate the likely consequences of causal node states (causal inference), as well as to diagnose the likely causes of a collection of dependent node values (diagnostic inference). In other words, either the gestural behavior of an agent can be generated, e.g., by propagating evidence about an object's properties to gesture characteristics. Or, given a particular gesture, the features of its referent object might be inferred (what is, however, out of the scope of this paper). Further, the use of decision nodes allows to enrich the dependencies directly learned from the data by additional model-based rules. Finally, the approach provides the possibility to detect clusters of speakers who share particular interrelations between causal nodes and observable behavior, i.e., it enables us to determine and to distinguish between different forms of inter- and intrapersonal systematicities in the production of iconic gestures.

With regard to the challenge of considering general and individual patterns in gesture formulation (see Section 3) , we use chance nodes to be learned from corpus data to model choices found to be highly idiosyncratic (gesture, representation technique, handedness, handshape). This learning task will be focussed in Section 6 where we compare several machine learning techniques in application to our data. Choices found to follow general patterns are realized as decision nodes (palm and finger orientation, movement features). Decision nodes specify rules and constraints on the gesture features. Each of them contains a set of 50-100 if-then rules for our current domain of application. The inter-subjective commonalities modeled with these rules realize the form-meaning mapping between referent shape and gesture form, thus, accounting for iconicity in the resulting gestures. The following examples illustrate the character of these rules:

```
if (and (Gesture="true", Handedness="rh", Handshape="ASL-G",
Technique="drawing", ShapeProp="round2d"), "MR>MD>ML>MU").


if (and (Gesture="true", Handedness="lh", Handshape="ASL-C"|"ASL-G"|"ASL-5",
Technique="shaping", Childnodes="0", MainAxis="z", ShapeProp="longish"), "PTR").
```

A schematic of the overall decision network is shown in Figure 3. It contains four chance nodes (drawn as ovals) which are connected to their predecessors by learning the network structure from speaker-specific data. In contrast to the chance nodes, the dependencies of the decision nodes (drawn as rectangles) are defined generally, i.e., they do not vary in the individual networks. Nevertheless, each decision node has chance nodes as predecessors so that these rule-based decisions are also dependent on chance variables whose (individual) values have been found previously. Furthermore, each decision node is informed from the set of referent features accounting for iconicity in the resulting gesture.

——————————————— INSERT FIGURE 3 HERE ———————————————

5 Gesture Formulation Process

Each GNetIc decision network, as described above, can be used directly for gesture formulation. However, a few pre- and post-processing steps are additionally necessary to complete the mapping from representations of imagistic semantics (IDT representation) and discourse context to an adequate speech-accompanying iconic gesture. Figure 4 gives a schematic of the formulation process. The gesture formulator has access to a structured blackboard since it is part of a greater speech and gesture generation architecture, in which all modules operate concurrently and proactively on this blackboard. Details of this architecture are described elsewhere (Bergmann & Kopp, 2009b). Information is accessed from that blackboard and results are written back to it.

The initial situation for gesture formulation is an IDT representation (kind of a 'mental image') of the object to be referred to. In a pre-processing step, this representation is analyzed to extract all features that are required as initial evidence for the network: (1) whether an object can be decomposed into subparts, (2) whether it has any symmetrical axes, (3) its main axis, (4) its position in the VR stimulus, and (5) its shape properties. Further information drawn upon by the decision network concerns the discourse context. It is provided by other modules in the overall generation process and can be accessed directly from the blackboard. All evidence available is then propagated through the network resulting in a posterior probability distribution for the values in each chance node. We make the decision which value is filled in the feature matrix specifying the gesture morphology by selecting the maximum a posteriori value. Alternatively, sampling over the distribution of alternatives (probability matching) could be applied at this stage of decision-making to result in non-deterministic gesture specifications. This would, however, decrease the accuracy of simulation results compared to the human archetype modeled in the network.

To avoid gesture specifications with incompatible feature values, a post-processing of this intermediate result is necessary. Take the example of a posturing gesture referring to a round window: If the speaker whose gesturing behavior is simulated, strongly prefers the handshape ASL-B (flat hand), it is likely that this handshape is also inferred from the network in this case. However, since in posturing gestures the hands themselves form a static configuration to stand as a model for the object itself, the handshape has to be reflective of the object's shape, i.e., the suggested flat handshape is inadequate for its referent. To reveal discrepancies like this one, each gesture feature matrix derived from the decision network is analyzed for its semantics: The morphological gesture features are transformed into an IDT representation according to form-meaning relations analyzed as described by Sowa & Wachsmuth (2005). This *gesture*-IDT is compared to the initial *referent*-IDT by means of formal graph unification. If a discrepancy is detected, the decision network is requested again with the additional constraint to either plan a gesture with a different technique or to return a feature matrix with a different handshape. The implementation of this post-processing is work in progress. It will be particularly important when extending our domain of application.

——————————— INSERT FIGURE 4 HERE ———————————

5.1 Generation Examples

A prototype of the previously described generation model has been realized using the HUGIN toolkit for Bayesian inference (Madsen et al., 2005) and the ACE realization engine (Kopp & Wachsmuth, 2004). In this prototype implementation a virtual agent explains the same virtual reality buildings that we already used in the previously described empirical study. Being equipped with proper knowledge sources, i.e., communicative plans, lexicon, grammar, propositional and imagistic knowledge about the world, the agent randomly picks a landmark and a certain spatial perspective towards it, and then creates his explanations autonomously on the fly. Currently, the system has the ability to simulate five different speakers by switching between the respective decision networks built from the individual speaker data as described above. The system is able to produce utterances for the interaction with a human user in realtime.

The resulting gesturing behavior for a particular referent in a respective discourse context varies in dependence on the decision network which is used for gesture formulation. In Figure 5, examples are given from simulations of five speakers, each of which based on exactly the same initial situation, i.e., all gestures are referring to the same referent (a round window of a church) and are generated in exactly the same discourse context ('landmark construction', 'rheme', 'private'). The resulting nonverbal behavior varies significantly depending on the decision network underlying the simulation: P1 uses a right-handed drawing gesture, whereas for P5 and P7 posturing gestures are produced which, however, differ in their low-level morphology. For P5 the handshape ASL-O is employed using the right hand while in the simulation for P7 ASL-C with both hands. For P8 no gesture is produced at all in the given discourse situation.

——————————— INSERT FIGURE 5 HERE ———————————

6 Evaluation of Different Learning Algorithms

The chance nodes in a BDN build an acyclic, directed graph (DAG) enriched with a set of probability distributions. For each chance node, the probability that the variable will be in one of its possible states given its parents' states can be calculated based on the frequency of their observed co-occurrences in a set of training data. For discrete variables, as in our data, probability distributions are expressed as conditional probability tables. Learning a Bayesian network from a sample of data cases comprises two tasks. First, identifying an adequate structure of the DAG (*structure learning*), and second, estimating a set of corresponding parameters (*parameter estimation*). For our purpose of employing Bayesian networks as a means to gain empirical insights into iconic gesture production, the task of structure learning is of particular importance since the graph structure reveals interrelations between variables. We will, therefore, investigate the question how to best learn a DAG from the SaGA corpus by comparing different structure learning algorithms.

There are two very different approaches to learning the structures of Bayesian networks from given data: *score-based learning* and *constraint-based learning*. The idea in score-based approaches is to define a global measure (score) which evaluates a given Bayesian network model as a function of the data. The problem is solved by searching in the space of possible Bayesian network models trying to find the network with optimal score. As concerns the scoring function, there are two popular choices. The first one, *Bayesian score*, measures the quality of a Bayesian network in terms of its posterior probability given the database (Cooper & Herskovits, 1992). It takes into account a prior and the marginal likelihood. The second one, *Bayesian Information Criterion* (BIC), is a likelihood criterion which takes into account the estimated model parameters and a number of data cases (Schwarz, 1978). The BIC method has the advantage of not requiring a prior.

Even for a relatively small number of variables, however, the estimation of a DAG from given data is difficult and computationally non-trivial due to the fact that the number of possible DAGs is super-exponential in the number of nodes. This is why search heuristics are used with either local or global search algorithms. The *K2* algorithm is a local, greedy search algorithm which is appropriate if the order of nodes is given (Cooper & Herskovits, 1992), as in our case of a sequential decision problem. Initially each node has no parents. The algorithm then adds incrementally that parent whose addition increases the score of the resulting structure most. When the addition of no single parent can increase the score, adding parents to the node is stopped.

A prominent global search method to find adequate network structures is the *Markov Chain Monte Carlo* (MCMC) algorithm called Metropolis-Hastings (Madigan & York, 1995). The basic idea is to construct a Markov Chain whose state space is a set of DAGs. The idea is to sample from this chain for “long enough” (burn-in time) to ensure it has converged to its stationary distribution. Notably, the algorithm is not deterministic; the chain can be run from multiple starting points with different values for the burn-in time whereby convergence is an open problem.

The other main approach to learning network structures is constraint-based learning. It is based on carrying out several independence tests on the database and building a DAG in agreement with the statistical test results. The most popular example of this type is the *PC* algorithm (Spirtes & Glymour, 1991). It starts from a complete, undirected graph and recursively deletes edges on the basis of conditional independence decisions. Statistical tests for conditional independence are then performed for all pairs of variables. An undirected link is added between each pair of variables for which no conditional independencies were found. Colliders are then identified, ensuring that no directed cycles occur. Next, directions are enforced for those links whose direction can be derived from the conditional independencies found and the colliders identified. One important thing to note about the PC algorithm is that, in general, it will not be able to derive the direction of all the links from data, and thus some links will be directed randomly. This means that the learned structure should be inspected, and if any links seem counterintuitive (e.g., gesture features causing object properties, instead of the other way around), one might consider going back and insert a constraint specifying the direction of the link. Correctness of the PC algorithm has been proven under the assumption of infinite data sets (Madsen et al., 2005).

Another constraint-based method is the *NPC* algorithm (Steck & Tresp, 1999) which is particularly useful when structure learning is based on a limited dataset. The NPC algorithm is an extension of the PC algorithm which introduces the notion of a *Necessary Path Condition*: if, at some point, a conditional dependence is established, there must be a connecting path explaining this dependency. The NPC condition is necessary for the existence of a perfect map of the conditional dependence and independence statements derived by statistical tests. Thus, in order for an independence statement to be valid, a number of links are required to be present in the graph.

In general, constraint-based algorithms show one major advantage compared to score-based algorithms: they allow to vary the significance level which is used by the conditional independence tests. This way, it is possible to judge the strength of dependencies among variables.

6.1 Learning Network Structures: Results and Analysis

We have applied all four methods described in the previous section to investigate how to best learn network structures from our data with regard to the challenge of formulating an adequate gesture for a particular referent in a given discourse situation (see Table 2 for the set of variables and values).

Learning network structures is constrained by the following assumptions. First, the order of decisions is given so that the choice to produce a gesture or not for a particular noun phrase is taken first, followed by the decision which representation technique to use. The choices for handedness and handshape are made thereafter. Second, the direction of edges is specified in a way that variables of gesturing behavior are influenced by the other variables, i.e., it is assumed for instance that properties of the gesture referent have an impact on the gesture and not the other way around. And third, no dependencies are assumed to exist among variables characterizing the given situation in terms of referent features, discourse context, and the previously performed gesture.

As score-based algorithms we employed K2 and MCMC as implemented in the BNT toolkit (Murphy, 2001). Both algorithms are parameterized to use the BIC score. MCMC's starting position is a graph without any edges and the burn-in time is set to 5n whereby n is the number of nodes. Since MCMC is not deterministic we have run the algorithm five times for each data set which sometimes resulted in different network structures. For each data set we have chosen the network structure that has been found most often by the algorithm. As constraint-based methods we considered PC and NPC implemented in HUGIN (Madsen et al., 2005) with a significance level of .01. Results of all four algorithms are given in Figure 6 for five speakers individually (rows 1-5) and for a dataset including all five speakers (row 6).

——————————————— INSERT FIGURE 6 HERE ———————————————

A comparison of the networks in each row reveals that learned structures are subject to the different learning algorithms. Network structures learned with K2 typically show the lowest number of edges, whereas skeletons learned with NPC show various links among variables. As concerns the constraint-based algorithms PC and NPC, it is conspicuous that dependencies found by the PC algorithm are always also covered by the network structure found by the NPC algorithm. In most of the cases, NPC detects more edges than PC which is not surprising due to the necessary path condition. This finding is in line with Steck & Tresp (1999) who found that the NPC algorithm learns more edges present in a dataset than the PC algorithm.

Comparing the different network structures in each column shows that the number of links found is depending on the data set, i.e., different networks are resulting from different speakers' data. This result is consistent among the different algorithms (i.e., between columns). For P8, for instance, all four structure learning methods result in a low number of edges. For P7, in contrast, all four algorithms learn a relatively high number of edges.

Further, it turns out that the different solutions have multiple edges in common, however, there are also some edges in which the structures differ (inconsistent edges). In this context it is notable that score-based approaches as K2 and MCMC consider a *global* measure for the entire network. Constraint-based methods, in contrast, remove all those edges from a network for which a conditional independence statement can be derived from the data. They do not take into account the structure of the network as a whole and can therefore be considered as *local*.

To investigate the quality of the different network structures with regard to their prediction accuracy, we also compared each model's choices with the empirically observed gesturing behavior from the data corpus. For each network structure found, its maximum likelihood estimates of parameters are computed employing the standard EM algorithm (Lauritzen, 1995).

In a leave-one-out cross-validation each model has been learned from n-1 selected data cases (n is the number of cases in the data set), leaving out the data case i. Then the model was tested on that data case. For each model the procedure is repeated n times so that each data case is used for testing once. The reported results, as summarized in Table 3, are an average over all n runs.

——————————————— INSERT TABLE 3 HERE ———————————————

It turned out that for each generation choice the prediction accuracy values clearly outperform the chance level baseline. In total, the prediction accuracy achieved with individual networks is, by trend, better than the accuracy achieved with networks learned from non-speaker specific data. This holds for the results of all four learning methods.

A comparison of the different learning techniques shows that networks learned with the score-based K2 algorithm result in the lowest accuracy values, whereas best results are achieved with the constraint-based PC algorithm. There is, however, no significant difference to be obtained between score-based and constraint-based algorithms in application to our data. Therefore, the application of constraint-based methods as a local method provide a more detailed insight into underlying data patterns and allow to gain insights into iconic gesture production in humans. Another advantage of using constraint-based algorithms like PC or NPC is the availability of link strength by varying the significance level which is used by the conditional independence tests. This way, it is possible to judge the strength of the dependencies among variables. As an example consider Figure 7 in which two network structures are displayed. The two networks differ obviously with regard to the number of dependencies learned. Whereas gesture production choices in the left network are predominantly influenced by the discourse context (nodes 'T', 'IS', and 'CG'), gesture features in the right network are additionally influenced by all referent features (nodes 'G', 'CN', 'MA', 'P', and 'SP') and also by the previously performed gesture (nodes 'LG', 'LT', 'LH', and 'LHS'). Moreover, only a part of the learned dependencies is highly significant. These connections are found with a significance level of 0.001, whereas other links are less strong, however, still significant (significance level of 0.01 or 0.05). In the left network, for instance, the decision to use a gesture (node 'G') is strongly influenced by the shape properties of the referent (node 'SP'), whereas link strength for the discourse factors thematization (node 'T') and the number of subparts (node 'CN') are less strong. Depending on the significance level used for learning the network structure, the resulting network structures are different: all three connections would, e.g., be present in the resulting network for a significance level of 0.05, whereas for a significance level of 0.001 only the strongest link would be learned.

——————————————— INSERT FIGURE 7 HERE ———————————————

6.2 Decision Nodes: Results and Evaluation

We have evaluated the performance of the four decision nodes of GNetIc, too, by comparison with actual gesture data. Note that decisions are made in a particular order, which has an impact on the validation results. If one of the earlier choices does not match the observed value in the test data, the following decisions typically cannot match the data either. Assume for instance that the predicted representation technique is ‘indexing’ although the representation technique in the test data is ‘shaping’. The following choices concerning morphological gesture features are accordingly made under false premises. Results for these rule-based decisions therefore are validated locally, i.e., we take the test case data for previous decisions as a basis and evaluate the quality of our decision making rules directly. The detailed results are given in Table 4. Notably, we cannot employ the same measure for all four variables. Palm and finger orientation are compared by calculating the angle between the two orientation vectors. For instance, there is an angle of 90° between ‘left’ and ‘up’, and an angle of 45° between ‘left’ and ‘left/up’. A maximal angle of 180° is present if the two vectors are opposing (e.g. ‘left’ and ‘right’) and can be considered the worst match. Considering this, the mean deviation for palm orientation of 54.6° (SD = 16.1°) and the mean deviation for finger orientation of 37.4° (SD = 8.4°) are quite satisfying results with deviations which lie well within the natural fuzziness of biological gestures.

For the movement direction we distinguish between motions along the following three axes: (1) sagittal axis (forward, backward), (2) transversal axis (left, right), and (3) vertical axis (up, down). Each segment in the generated movement description is tested for co-occurrence with the annotated value, resulting in a accuracy measure between 0 (no agreement) and 1 (total agreement). For multi-segmented movements the mean accuracy is calculated, i.e., if a generated movement consists of two segments from which only one matches the annotation, the similarity is estimated with a value of 0.5. Evaluating GNetIc with this measure gives a mean similarity of .75 (SD = .09). For the movement type (linear or curved) we employed the standard measure of accuracy, i.e., we compared if the generated value exactly matches the annotated value. The mean accuracy for the movement type is 76.4% (SD=13.6).

——————————————— INSERT TABLE 4 HERE ———————————————

7 Conclusion

In this paper we have presented a novel approach to generating iconic gestures for virtual agents with an integrated model, combining probabilistic and rule-based decision making. In particular, we compared the application of different state-of-the-art machine-learning techniques to learn and build Bayesian network structures from corpus data.

The comparison of the different learning techniques shows that networks learned with the constraint-based PC algorithm result in the best accuracy values. In general, there is no significant difference between score-based and constraint-based algorithms in application to our data with respect to their prediction accuracy. We prefer, therefore, the application of constraint-based methods which provide a more detailed insight into the underlying data and allow for elucidating iconic gesture production in humans. The analysis of learned network structures showed that gesture generation choices are actually constrained by referent features, the overall discourse context and also by the previously performed gesture. Moreover, networks built from individual speaker-specific data differ significantly revealing that individual differences are not only present in the overt gestures, but also in the production process they originate from. Whereas gesture production in some individuals is, e.g., predominantly influenced by visuo-spatial referent features, other speakers mostly rely on the discourse context. So there seems to be a set of different gesture generation strategies from which individuals typically apply a particular subset (cf. Dale & Viethen, 2009, for similar findings from the generation of referring expressions). Future research has to be undertaken to further investigate these questions. Overall, the conclusion to be taken is that the GNetIc simulation approach beside being valuable for an adequate simulation of speaker-specific gestures in virtual agents, is an appropriate means to shed light onto the open research questions of how iconic gestures are shaped and which sources individual differences in gesturing may originate from.

To evaluate GNetIc-generated gestures in terms of their impact on the interaction between humans and machines, calls for a study that analyzes if (1) semantic information uptake from gestures, and (2) the perceived interaction quality (expressiveness, naturalness etc.), is influenced by the agent's gesturing behavior. Generated gestures whose features do not fully coincide with our original data may still serve their purpose to communicate adequate semantic features of their referent–even in a speaker-specific way.

As concerns the GNetIc model itself, we are aware that our modeling results also reveal deficiencies of the current model, which mark starting points for further refinements. First and foremost, the amount of corpus data from which the networks are built should be increased. It is to expect that prediction accuracy improves the more data is available

(Steck & Tresp, 1999). In addition, the number of factors we have take into account so far is by no means complete. Since, for instance, gesturing is in particular a phenomenon of face-to-face dialogue (Bavelas et al., 2008), dialogue situation, partner model, gestural mimicry (Kimbara, 2006) etc. also have to be considered. Moreover, we see that gesturing behavior varies over time such that gestures become more simplified in the course of the discourse. This could be accounted for by considering the complete course of the previous discourse (not only the last gesture), which implies a more extensive use of dynamics in the networks. Finally, the application of the gesture formulator in the greater production architecture necessitates to decide between several variants of gesture morphology in combination with alternative results from the speech formulation processes. One particular feature of decision networks lends itself to this purpose, namely, the possibility to incorporate functions judging the utility of single decisions or collections of decisions. These extensibilities substantiate that Bayesian Decision Networks in combination with machine learning techniques can be an adequate formalism to successfully tackle the challenge of iconic gesture generation.

Acknowledgements

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

Bibliography

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

Bergmann, K. & Kopp, S. (2009a). GNetIc–Using bayesian decision networks for iconic gesture generation. In Z. Ruttkay, M. Kipp, A. Nijholt, & H. Vilhjalmsson (Eds.), *Proceedings of the 9th International Conference on Intelligent Virtual Agents* (pp. 76–89). Berlin/Heidelberg: Springer.

Bergmann, K. & Kopp, S. (2009b). Increasing expressiveness for virtual agents–Autonomous generation of speech and gesture in spatial description tasks. In K. Decker, J. Sichman, C. Sierra, & C. Castelfranchi (Eds.), *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems* (pp. 361–368). Budapest, Hungary.

Bergmann, K. & Kopp, S. (2010). Systematicity and idiosyncrasy in iconic gesture use: Empirical analysis and computational modeling. In S. Kopp & I. Wachsmuth (Eds.), *Gesture in Embodied Communication and Human-Computer Interaction*. Berlin/Heidelberg: Springer.

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

Cooper, G. F. & Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning Journal*, 9, 308–347.

Dale, R. & Viethen, J. (2009). Referring expression generation through attribute-based heuristics. In E. Krahmer & M. Theune (Eds.), *Proceedings of the 12th European Workshop on Natural Language Generation* (pp. 58–65). Athens, Greece: Association for Computational Linguistics.

de Ruiter, J. (2007). Postcards from the mind: The relationship between speech, imagistic gesture, and thought. *Gesture*, 7(1), 21–38.

Hartmann, B., Mancini, M., & Pelachaud, C. (2006). Implementing expressive gesture synthesis for embodied conversational agents. In S. Gibet, N. Courty, & J.-F. Kamp (Eds.), *Gesture in Human-Computer Interaction and Simulation* (pp. 45–55). Berlin/Heidelberg: Springer.

Hostetter, A. & Alibali, M. (2007). Raise your hand if you're spatial–Relations between verbal and spatial skills and gesture production. *Gesture*, 7(1), 73–95.

Howard, R. & Matheson, J. (2005). Influence diagrams. *Decision Analysis*, 2(3), 127–143.

Huenerfauth, M. (2008). Spatial, temporal and semantic models for American Sign Language generation: Implications for gesture generation. *International Journal of Semantic Computing*, 2(1), 21–45.

Kendon, A. (2004). *Gesture–Visible Action as Utterance*. Cambridge University Press.

Kimbara, I. (2006). On gestural mimicry. *Gesture*, 6(1), 39–61.

Kontkanen, P., Myllymäki, P., Silander, T., & Tirri, H. (1997). Comparing predictive inference methods for discrete domains. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics* (pp. 311–318). Ft. Lauderdale, USA.

Kopp, S., Tepper, P., & Cassell, J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of the 6th International Conference on Multimodal Interfaces* (pp. 97–104). New York: ACM Press.

Kopp, S., Tepper, P., Ferriman, K., Striegnitz, K., & Cassell, J. (2007). Trading spaces: How humans and humanoids use speech and gesture to give directions. In T. Nishida (Ed.), *Conversational Informatics* (pp. 133–160). New York: John Wiley.

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

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19, 191–201.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2010). The Bielefeld speech and gesture alignment corpus (SaGA). In M. Kipp, J.-C. Martin, P. Paggio, & D. Heylen (Eds.), *Proceedings of the LREC 2010 Workshop on Multimodal Corpora*.

Madigan, D. & York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63, 215–232.

Madsen, A., Jensen, F., Kjærulff, U., & Lang, M. (2005). HUGIN–The tool for bayesian networks and influence diagrams. *International Journal of Artificial Intelligence Tools*, 14(3), 507–543.

McNeill, D. (2005). *Gesture and Thought*. Univ. of Chicago Press: Chicago.

Morency, L.-P., de Kok, I., & Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Journal of Autonomous Agents and Multi-Agent Systems*, 20(1), 70–84.

Müller, C. (1998). *Redebegleitende Gesten: Kulturgeschichte–Theorie–Sprachvergleich.* Berlin: Berlin Verlag.

Murphy, K. (2001). The bayes net toolbox for MATLAB. *Computing Science and Statistics*, 33.

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

Ruttkay, Z. (2007). Presenting in style by virtual humans. In A. Esposito (Ed.), *Verbal and Nonverbal Communication Behaviours* (pp. 23–36). Berlin: Springer Verlag.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.

Sowa, T. & Wachsmuth, I. (2005). A model for the representation and processing of shape in coverbal iconic gestures. In *Proceedings of KogWis05* (pp. 183–188).

Spirtes, P. & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computing Review*, 9(1), 62–72.

Steck, H. & Tresp, V. (1999). Bayesian belief networks for data mining. In *Proceedings of the 2nd Workshop on Data Mining and Data Warehousing*.

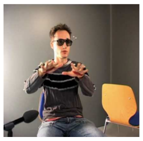
Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., & Bregler, C. (2004). Speaking with hands: Creating animated conversational characters from recordings of human performance. In *Proceedings of SIGGRAPH '04* (pp. 506–513).

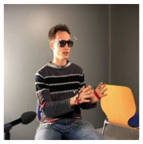
Streeck, J. (2008). Depicting by gesture. *Gesture*, 8(3), 285–301.

Yu, Q. & Terzopoulos, D. (2007). A decision network framework for the behavioral animation of virtual humans. In *Proceedings of SIGGRAPH '07* (pp. 119–128).

Table 1. Example transcript from the corpus (the verbal utterances are translated from German to English).

Router:	There is [a large square]$_{g1}$.
Follower:	Mhm.
Router:	On the left hand there is [a]$_{g2}$ church.
	And on the right [hand]$_{g3}$ there is another church.


Gesture *g1*


Gesture *g2*

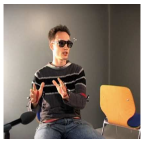

Gesture *g3*

Table 2. Coding scheme for gestures and their discourse context.

	Variable	**Annotation Primitives**
Gesture	Representation Technique (RT)	indexing, placing, shaping, drawing, posturing
	Handedness (H)	rh, lh, 2h
	Handshape (HS)	ASL handshapes, e.g. ASL-B, ASL-C
	Palm Orientation (PO)	up, down, left, right, towards, away
	Finger Orientation (FO)	up, down, left, right, towards, away
	Movement Direction (MD)	up, down, left, right, forward, backward
	Movement Type (MT)	linear, curved
Discourse Context	Thematization (T)	theme, rheme
	Information State (IS)	private, shared
	Communicative Goal (CG)	introduction, description, construct, position
Referent Features	Subparts (CN)	true, false
	Symmetry (S)	mirror, round, none
	MainAxis (MA)	x-axis, y-axis, z-axis, none
	Position (P)	3D vector
	ShapeProp (SP)	round2d, round3d, longish, etc.

Table 3. Prediction accuracy of single features for the networks learned with different structure learning algorithms from individual and combined speaker data, respectively.

		Accuracy (%)							
Generation Choices	**Chance Level Baseline**	K2		MCMC		PC		NPC	
		Indiv.	Comb.	Indiv.	Comb.	Indiv.	Comb.	Indiv.	Comb.
Gesture (y/n)	50.0	82.5	83.3	68.7	60.7	82.5	83.3	84.8	77.6
Technique	20.0	60.6	49.1	50.9	49.1	66.4	62.6	59.5	59.9
Handedness	33.3	60.6	65.4	63.7	59.2	62.3	69.9	61.2	65.1
Handshape	16.7	65.1	49.8	60.6	49.8	71.3	59.2	67.8	53.3
Total Accuracy		69.3	57.4	61.2	55.5	71.3	69.1	67.8	65.8

Table 4. Evaluation results of generation choices made by GNetIc's decision nodes.

Generation Choices	P1	P5	P7	P8	P15	Mean (SD)
Palm Orientation	37.1°	61.9°	76.4°	57.1°	40.5°	54.6° (16.1°)
Finger Orientation	29.0°	41.7°	41.9°	27.9°	46.6°	37.4° (8.4°)
Movement	.69	.84	.84	.56	.89	.75 (.09)
Movement Direction	.82	.76	.82	.61	.76	.76 (.14)

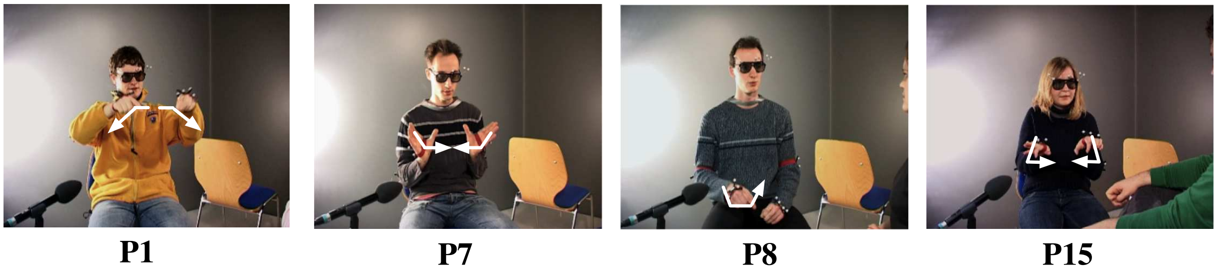


Fig. 1. Example gestures from different speakers, each referring to the same u-shaped building.

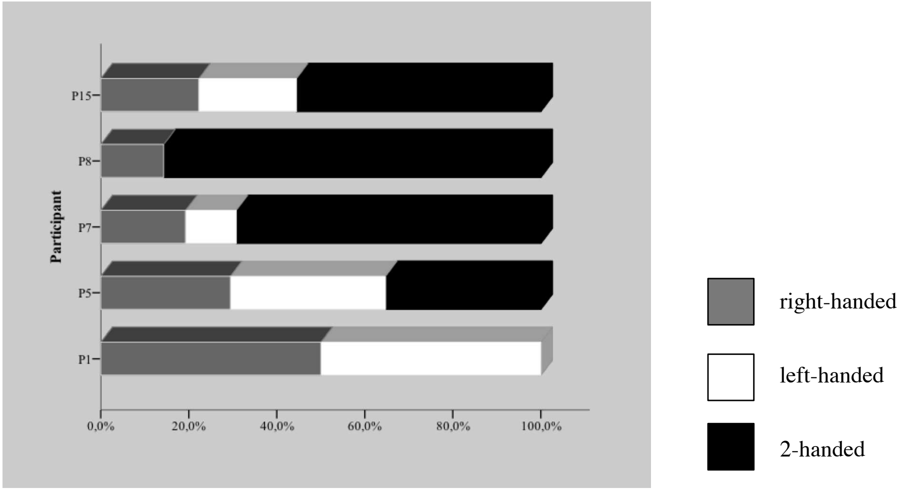


Fig. 2. Example of inter-subjective variability: gesture handedness in placing gestures across five speakers (N=70).

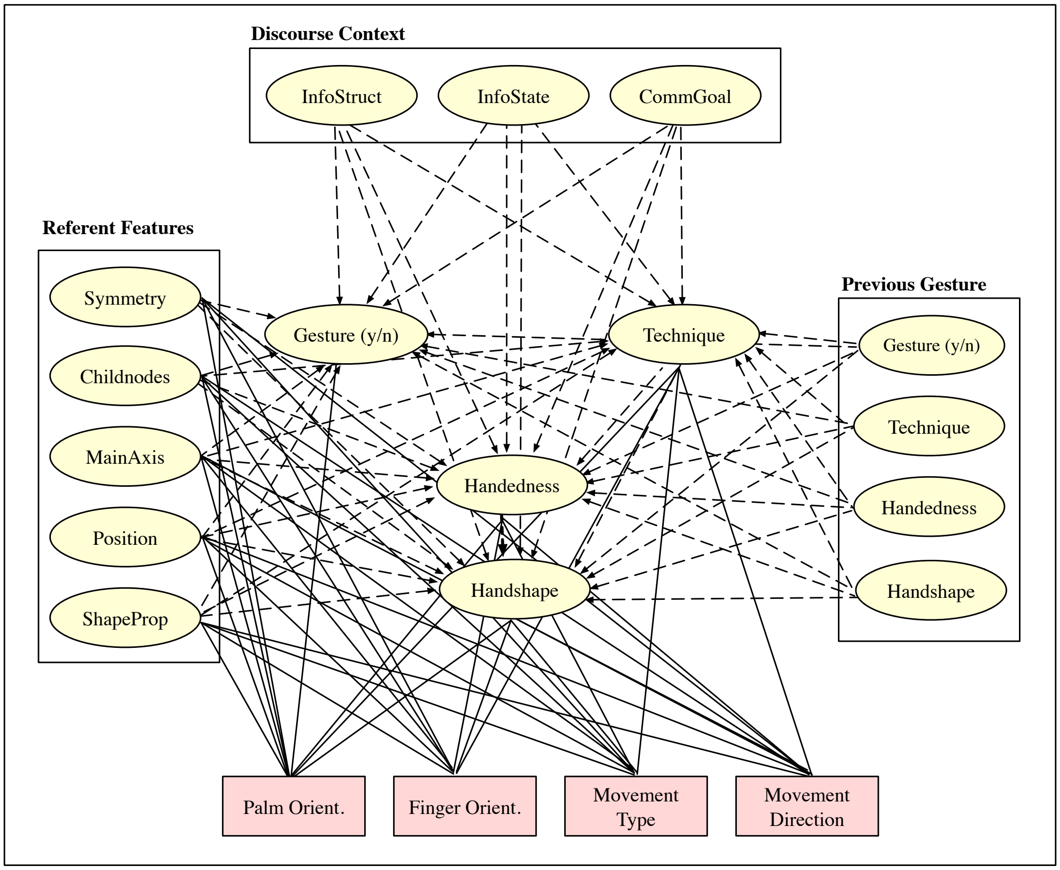


Fig. 3. General structure of a GNetIc decision network. Gesture production choices are considered either probabilistically (chance nodes drawn as ovals) or rule-based (decision nodes drawn as rectangles). Each choice is depending on a number of contextual variables. The links are either learned from corpus data (dotted lines) or defined in a set of if-then rules (solid lines).

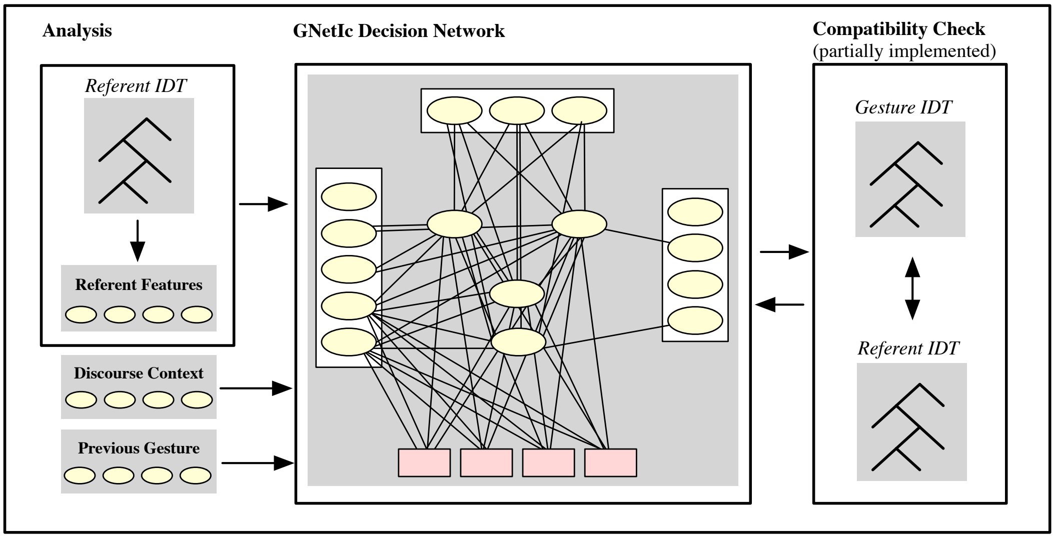


Fig. 4. Schematic of the gesture formulation process (see text for details).

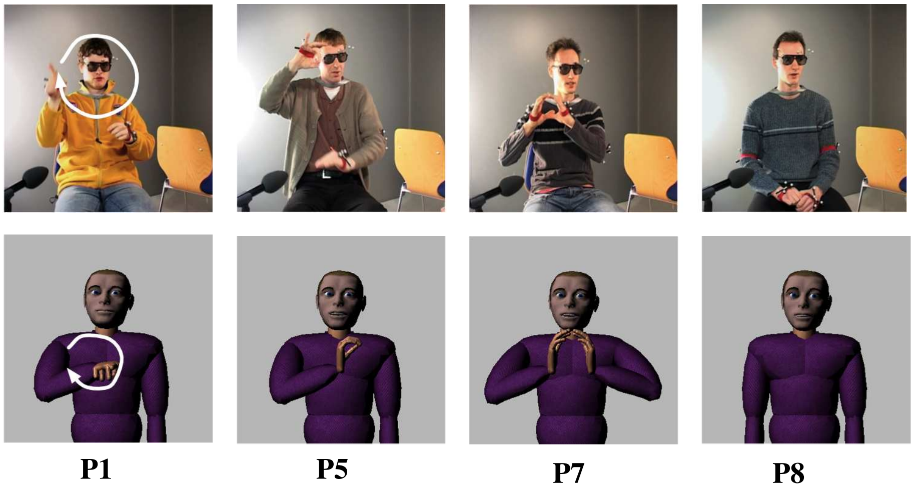


Fig. 5. Examples from different speakers each describing a round window of a church (first row) and GNetIc-generated behavior simulating those speakers in the same initial situation (second row).

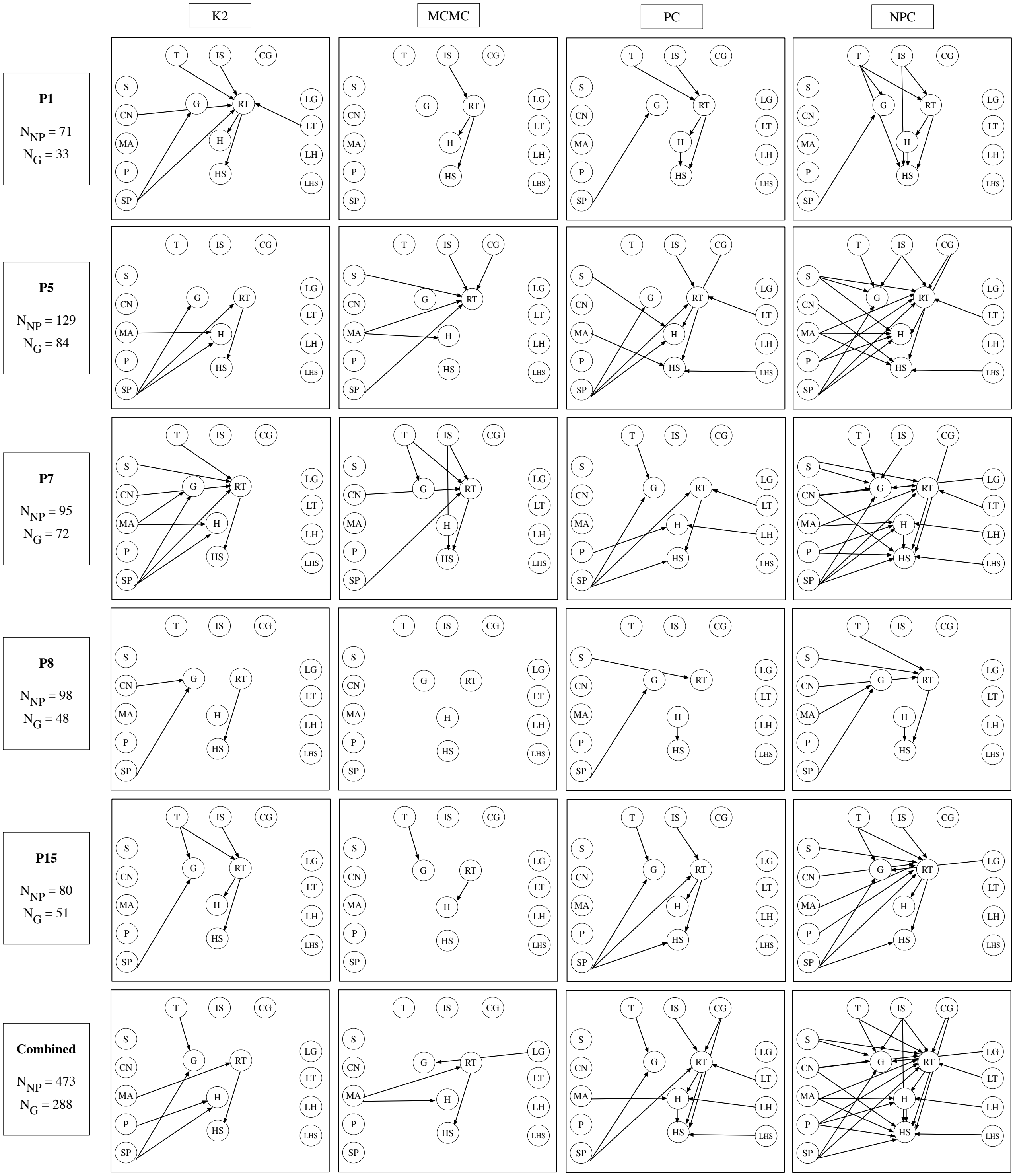


Fig. 6. Network structures obtained with different algorithms (columns) for different data sets (rows). See Table 2 for the reading of the variables and their value sets.

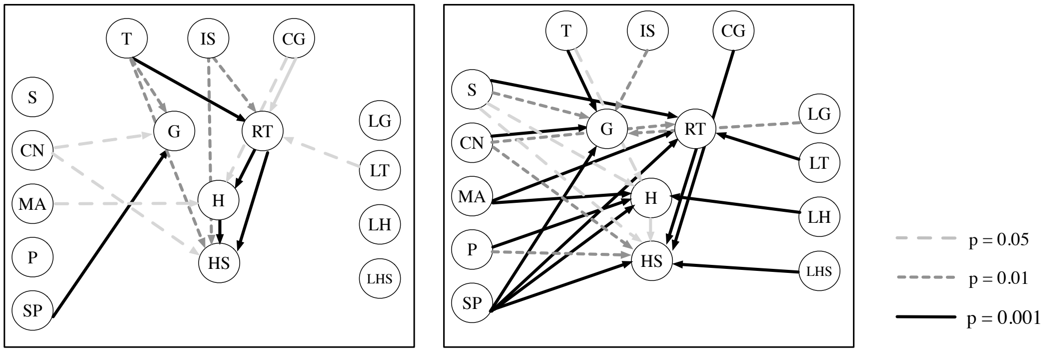


Fig. 7. Network structures obtained with NPC algorithm for two different speakers (left, right): link strength corresponds to significance level.