Individualized Gesturing Outperforms Average Gesturing – Evaluating Gesture Production in Virtual Humans

Kirsten Bergmann[1,2], Stefan Kopp[1,2], and Friederike Eyssel[1]

[1] Center of Excellence in "Cognitive Interaction Technology" (CITEC), Bielefeld University
[2] Collaborative Research Center 673 "Alignment in Communication", Bielefeld University
P.O. Box 100 131, D-33501 Bielefeld, Germany
{kbergmann,skopp}@techfak.uni-bielefeld.de
friederike.eyssel@uni-bielefeld.de

Abstract. How does a virtual agent's gesturing behavior influence the user's perception of communication quality and the agent's personality? This question was investigated in an evaluation study of co-verbal iconic gestures produced with the Bayesian network-based production model GNetIc. A network learned from a corpus of several speakers was compared with networks learned from individual speaker data, as well as two control conditions. Results showed that automatically GNetIc-generated gestures increased the perceived quality of an object description given by a virtual human. Moreover, the virtual agent showing gesturing behavior generated with individual speaker networks was rated more positively in terms of likeability, competence and human-likeness.

Key words: Evaluation, Gesture Generation, Inter-subjective Differences

1 Introduction

A major goal in developing intelligent virtual agents (IVAs) is to advance the interaction between humans and machines towards natural and intuitive conversation. Human-human conversation is characterized by a high degree of multi-modality combining speech and other non-verbal behavior such as gestures, facial expressions, gaze, body posture, and intonation. Thus, IVA researchers are faced with two major problems: first, how to master the technical challenge to generate flexible conversational behavior automatically in IVAs and, second, how to ensure that the produced synthetic behavior improves the human-agent conversation valued by human users. The first issue has sparked the interest of many researchers in the field of IVA. For instance, regarding iconic gestures, different modeling approaches are tested, with the goal of identifying systematic characteristics of co-verbal gestures, shared among speakers, and have tried to cast these commonalities into generative models [6, 16, 19]. Others have emphasized individual differences in communicative behavior, e.g. [27, 8], or have tried to model individual gesture style for IVAs [29, 12, 26]. It is obvious that for the generation of multimodal behavior the consideration of both, commonalities that account for an agreed (or even conventionalized) sign system, and idiosyncrasies that make for a coherent individual style is an important issue. In previous work [1] we have proposed GNetIc (*Gesture Net for Iconic Gestures*) to automatically derive novel gestures from contextual demands, for instance, the given communicative goal, discourse status, or referent features. By combining rule-based and data-based models, GNetIc can simulate both systematic patterns shared among several speakers, as well as idiosyncratic patterns specific to an individual. That is, GNetIc can produce novel gestures simulating a specific speaker.

The second major problem to be addressed concerns the question of how of ensuring positive affect and user acceptance. There is increasing evidence that endowing virtual agents with human-like, non-verbal behavior may lead to enhancements of the likeability of the agent, trust in the agent, satisfaction with the interaction, naturalness of interaction, ease of use, and efficiency of task completion [4, 13]. With regards to effects of co-speech gestures, Krämer et al. [21] found no effect on agent perception when comparing a gesturing agent with a non-gesturing one. The agent displaying gestures was perceived just as likable, competent, and relaxed as the agent that did not produce gestures. In contrast, Cassell and Thórisson reported that non-verbal behavior (including beat gestures) resulted in an increase of perveived language ability and life-likeness of the agent, as well as smoothness of interaction [7]. A study by Rehm and André revealed that the perception of an agent's politeness depended on the graphical quality of the employed gestures [28]. Moreover, Buisine and Martin [5] found effects of different types of speech-gesture cooperation in agent's behavior. They found that redundant gestures increased ratings of explanation quality, expressiveness of the agent, likeability and a more positive perception of the agent's personality. In an evaluation of speaker-specific gesture style simulation, Neff et al. [26] reported that the proportion of subjects who correctly recognized a speaker from generated gestures was significantly above chance.

The goal of this paper is to evaluate the GNetIc production model to explore if and how automatically generated gestures can be beneficial for human-agent interaction. In particular, we were interested in (1) the quality of the produced iconic gestures as rated by human users; (2) whether an agent's gesturing behavior could systematically alter a user's perception of the agent's likeability, competence, and human-likeness; and (3) whether producing gestures like a particular individual or like the average speaker is preferable. To investigate these questions, we exploit the flexibility afforded by GNetIc to generate speech-accompanying gestures in different conditions: individual speaker networks (representing an individualized gesturing style), networks learned from corpus data of several speakers, random gestures, or no gestures at all. The following section briefly describes the GNetIc production model. Section 3 describes the setting and procedure of the evaluation study. Results are presented in Section 4. Finally, we discuss the results and draw conclusions in Section 5.

2 Gesture Generation with GNetIc

Iconic gestures, in contrast to language or other gesture types, such as emblems, have no conventional form-meaning mapping. Apparently, iconic gestures communicate by virtue of iconicity, i.e., their physical form corresponds to object features such as shape or spatial properties. Empirical studies have revealed, however, that similarity with the referent cannot fully account for all occurrences of iconic gesture use [31]. Recent findings actually indicate that a gesture's form can be influenced by a variety of contextual constraints, and that distinctive differences in personal and cultural backgrounds can lead to obvious inter-subjective differences in gesturing (cf. [15]). Consider, for instance, gesture frequency: while some people rarely make use of their hands while speaking, others do so almost without interruption. Similarly, individual variation becomes apparent in preferences for general gestural representation techniques [24, 17, 31] or the choices of morphological features, such as handshape or handedness [2].

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 be taken into consideration by an account of why people use gestures the way they actually do. To tackle the challenge of considering both general and individual patterns in gesture formulation, we have proposed GNetIc [1]. In this approach, we employ Bayesian Decision networks which provide a representation of a finite sequential decision problem, combining probabilistic and rule-based decision-making. Gesture features empirically found to be highly idiosyncratic, *Idiosyncratic Gesture Features* (IGFs) henceforth, are represented as nodes conditioned by probability distributions. These distributions can be learned from corpus data–either from data of several speakers or for an individual speaker's data separately [3]. Resulting networks differ in their global network structure as well as in their local conditional probability distributions, revealing that individual differences are not only present in the overt gestures but can be traced back to the production process they originate from.

Table 1. Gesture features, their types and the values as determined with GNetIc.

Feature Type	Gesture Features	Values
Idiosyncratic Gesture Features (IGFs)	Gesture (G)	*yes, no*
	Representation Technique (RT)	*indexing, placing, shaping, drawing, posturing*
	Handedness (H)	*rh, lh, 2h*
	Handshape (HS)	*ASL handshapes, e.g. ASL-B, ASL-C*
Common Gesture Features (CGFs)	Palm Orientation (PO)	*up, down, left, right, towards, away*
	Finger Orientation (FO)	*up, down, left, right, towards, away*
	Movement Type (MT)	*up, down, left, right, towards, away*
	Movement Direction (MD)	*up, down, left, right, towards, away*

Other gesture features, in contrast, are more universal. These features basically realize the form-meaning mapping between referent shape and gesture form, thus, accounting for most of the iconicity in the resulting gestures. In the following, we will refer to these features as *Common Gesture Features* (CGFs). In GNetIc networks, the use of these features is modeled in a rule-based way, that is, by nodes containing a set of if-then rules. Table 1 summarizes GNetIc gesture features and their properties.

Figure 1 illustrates the overall decision network. It contains four nodes representing the IGFs (see Table 1; drawn as ovals) which are connected to their predecessors by the network structure learned from speaker-specific data. The dependencies (edges) of nodes representing the CGFs (drawn as rectangles) are defined universally and do not vary across individual networks. Nevertheless, since each CGF-node has IGF-nodes as predecessors, the rule-based CGF decisions depend on IDFs whose (individual) values have been determined previously. Furthermore, each CGF-node is determined from the visuo-spatial features of the referent accounting for iconicity in the resulting gesture.

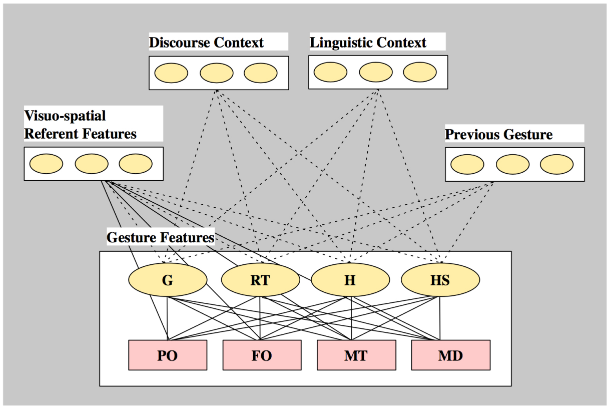


Fig. 1. General structure of a GNetIc decision network. Gesture production choices are taken either probabilistically (IGF-nodes drawn as ovals) or rule-based (CGF-nodes drawn as rectangles), solely depending on the values of connected contextual variables. Links are either learned from corpus data (dotted lines) or defined by a set of if-then rules (solid lines).

Whether the non-verbal behavior produced by GNetIc is a reasonable simulation of real speaker behavior, has been investigated in [1, 3]. To do so, we conducted a corpus-based cross-validation study in which we compared the model's predictions with the actual gestures we had observed empirically. Results for both, IGF- and CGF-nodes, were quite satisfying with deviations lying well within what can be considered the natural fuzziness of human gesturing behavior. However, to find out whether the automatically generated gestures are actually comprehensible as intended and thus helpful in human-agent interaction, we still needed to conduct a study to evaluate GNetIc with real human users. This study is described in the following.

3 Evaluation Study

The present study was designed to investigate three questions. First, can we achieve a reasonable quality in the iconic gestures automatically derived with GNetIc as perceived by users? Second, is the user's perception of an agent in terms of likeability, competence, and human-likeness altered by the agent's gesturing behavior? And third, is it preferable to produce gestures like a particular individual or like the average speaker?

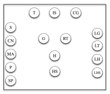
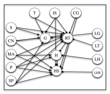
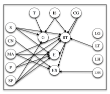
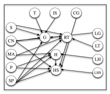
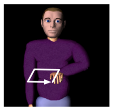
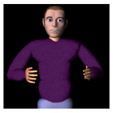
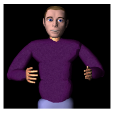
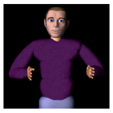
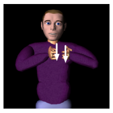
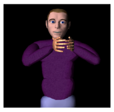
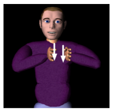
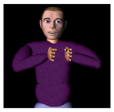
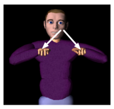
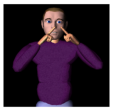
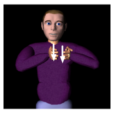
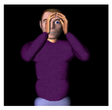
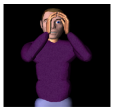
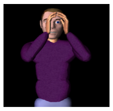
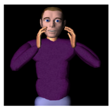
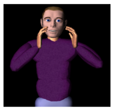
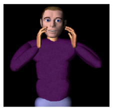
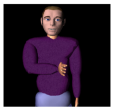
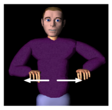
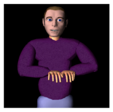
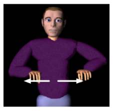
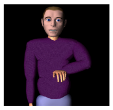
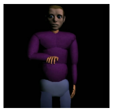
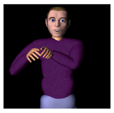
3.1 Independent Variables

In a between-subject design, participants were presented with a description of a church building given by the virtual human Max [20]. All descriptions were produced fully autonomously at runtime using a speech and gesture production architecture into which GNetIc was integrated [1]. We manipulated the gesturing behavior of the agent resulting in five different conditions in which Max, notably, received identical communicative goals and produced identical verbal utterances throughout. Furthermore, all gestures were generated from the same knowledge base, a visuo-spatial representation of referent features (IDT, [30]). In two individual conditions, *ind-1* and *ind-2*, the GNetIc networks were learned from data of individual speakers from our SaGA corpus [22] (subject P5 in *ind-1*, subject P7 in *ind-2*). We have chosen these two speakers because both speakers gestured quite frequently and approximately at the same rate. In a *combined* condition, the GNetIc network was generated from data of five different speakers (including P5 and P7). These speakers' gesture styles are thus amalgamated in one network. As a consequence, Max' gesturing behavior was not as consistent as with individual networks with regard to the IDF choices. Finally, we added two control conditions: in the first one, *no gestures* were produced at all, whereas in the second, values in the four IGF-nodes were determined by chance (*random*). The latter condition can result, for instance, in gestures occurring at atypical (e.g., thematic) positions in a sentence since the network was applied for every noun phrase in the verbal description.

Overall, the virtual agent's verbal utterances were held constant and all gestures were created fully autonomously by the system. There was no within-condition variation, because choices in the Bayesian networks were not made via sampling, but by choosing the values with maximum a-posteriori probability. Furthermore, the values for the CGFs were determined in the same rule-based way in all conditions, to ensure that no "non-sense" gestures were produced throughout.

Table 2 shows the stimuli that resulted from the five different conditions. There was no wide difference across conditions in gesture frequency (either five, six or seven gestures in six sentences). However, the two individual GNetIc conditions are characterized by less variation in the production choices. In condition *ind-1* gestures are predominantly static, whereas there are more dynamic shaping gestures in condition *ind-2*. Moreover, the gestures in condition *ind-1* are mostly performed with c-shaped hands, whereas in *ind-2* some gestures are performed with a flat handshape. In the *combined* GNetIc condition, a combination of different techniques is obvious. A similar mixture of techniques is observable in the *random* condition which is further characterized by inconsistency in handedness and handshapes. Moreover, gestures in this condition can occur at atypical positions in a sentence.

Table 2. Stimuli presented in the five different conditions: verbal description given in each condition (left column; translated to English; gesture positions labelled with squared brackets); GNetIc networks from which the gesturing behavior were produced (top row); gestures produced (right columns).

	no gesture	*random*	*combined*	*ind-1*	*ind-2*
					
[The church is squared]...					
...and in the middle there is [a small spire.]					
[The spire]...					
...has [a tapered roof].					
And [the spire]...					
has [a clock].					
There is [a door] in front.					
And in front of the church there is [a low, green hedge].					
There is [a large deciduous tree] to the right of the church.					

3.2 Dependent Variables

Immediately upon receiving the descriptions of the church by Max, participants filled out a questionnaire in which two types of dependent variables had to be assessed on seven-point Likert-scales. First, participants were asked to evaluate the presentation quality with respect to Max's language capability (eloquence) as well as gesture quantity and quality. With regard to gesture quality we used the attributes spatial extent, temporal extent, fluidity, and power as proposed in [12]. Further, we measured the degree as to how comprehensible Max's explanations were, as well as how vivid the agent's mental concept (or mental image) of the environment was. Second, participants were asked to report their perception of the virtual agent's personality. To this end, we chose 18 items [9, 14], e.g., 'pleasant', 'friendly', 'helpful' (translated from German) to access the degree to which they applied to Max using a 7-point Likert scale.

3.3 Participants

At total of 110 participants (22 in each condition), aged from 16 to 60 years (M = 23.85, SD = 6.62), took part in the study. 44 participants were female and 66 were male. All of them were recruited at Bielefeld University and received 3 Euros for participating.

3.4 Procedure

Participants were instructed to carefully watch the presentation given by the virtual agent Max in order to be able to answer questions regarding content and subjective evaluation of the presentation afterwards. Figure 2 shows the setup used for stimulus presentation: Max was displayed on a 80 x 143 cm screen and thus appeared in life-size of 1.25 m. Life-sized projections have been shown to yield visual attention and fixation behavior towards gestures that is similar to behavior in face-to-face interactions [11]. Participants were seated 170 cm away from the screen and their heads were approximately leveled with Max's head.

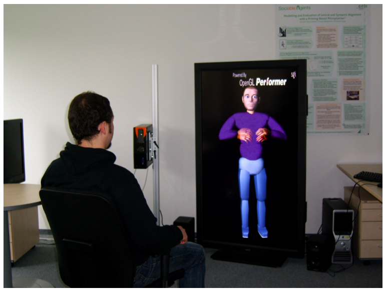


Fig. 2. Set-up of the stimulus presentation phase.

Participants were randomly assigned to one of the five conditions. The object description given by Max was preceded by a short introduction: Max introduced himself and repeated the instruction already given by the experimenter to get participants used to the speech synthesis. The following object description was always six sentences long and took 45 seconds. Each sentence was followed by a pause of three seconds. Participants have been left alone for the stimulus presentation, and after receiving the questionnaire to complete it (neither experimenter nor Max present).

Table 3. Mean values for the dependent variables of presentation quality in the five conditions (standard deviations in parentheses).

	ind-1	***ind-2***	***combined***	***no gestures***	***random***
Gesture Quantity	3.91 (1.15)	3.95 (0.95)	3.59 (0.91)	2.48 (1.21)	3.55 (1.22)
Spatial Extent	3.77 (0.87)	4.14 (0.83)	3.59 (1.05)	–	3.55 (1.05)
Temporal Extent	3.68 (0.83)	3.64 (0.66)	3.50 (1.01)	–	3.30 (0.87)
Fluidity	4.09 (1.48)	4.00 (1.57)	3.05 (1.32)	–	3.65 (1.53)
Power	3.59 (1.10)	4.09 (1.27)	3.91 (1.38)	–	3.90 (1.48)
Eloquence	3.50 (1.74)	4.91 (1.14)	3.05 (1.46)	3.69 (1.11)	3.25 (1.61)
Comprehension	5.18 (1.33)	5.27 (1.16)	4.68 (1.49)	4.95 (1.32)	4.18 (1.37)
Gestures helpful	5.68 (1.56)	5.82 (0.85)	4.70 (1.62)	1.82 (1.14)	4.10 (2.05)
Vividness	5.32 (1.62)	5.45 (1.13)	4.18 (1.81)	4.08 (1.32)	3.81 (1.80)

4 Results

In the following we report results regarding the effect of experimental conditions on perceptions of presentation quality and agent perception. The third methodological issue will be discussed based on these results in Section 5.

4.1 Quality of Presentation.

We investigated the perceived quality of presentation with regard to gestures, speech, and content. Participants were asked to evaluate each variable on a seven-point Likert-scale. To test the effect of experimental conditions on the dependent variables, we conducted analyses of univariate variance (ANOVA) and paired-sample *t*-tests with 95% condence intervals (CI) for these pairwise comparisons between condition means. Mean values and standard deviations are summarized in Table 3 and visualized in Figure 3 for dependent variables with significant main effects.

Gesture Quantity. With regard to gesture quantity, the overall mean value for the four gesture conditions was M=3.75 (SD=1.06) on a seven-point Likert-scale (too few–too many). There was no significant main effect for experimental conditions. That is, participants were quite satisfied with the gesture rate. For the *no gesture* condition participants rated gesture quantity as rather too low (M=2.48, SD=1.21).

Gesture Quality. No main effect for experimental conditions was obtained for the four attributes characterizing gesture quality: spatial extent (too small–too large, M=3.77, SD=0.97), temporal extent (too slow–too fast, M=3.53, SD=0.85), fluidity (not fluid–very fluid, M=3.70, SD=1.51), and power (weak–powerful, M=3.87, SD=1.30). In all four gesture conditions the four quality attributes were rated with mean values between 3.0 and 4.0 on a seven-point Likert-scale.

Eloquence. With regard to perceived eloquence of the virtual agent (Max is not eloquent–Max is eloquent), there was a significant main effect (*F*(4,79)=3.12, *p*=.02). This is due to the fact that the mean of condition *ind-2* differed from all other conditions (*ind-2/no gesture*: *t*(21)=2.64, *p*=.02, CI=[0.26;2.17]; *ind-2/random*: *t*(25)=2.94, *p*=.01, CI=[0.50;2.82]; *ind-2/combined*: *t*(25)=4.02, *p*=.001, CI=[0.91;2.82]; *ind-2/ind-1*: *t*(31)=2.43, *p*=.02, CI=[0.23;2.59]). That is, gestures produced with a suitable individual gesture network have the potential to increase the perceived eloquence (recall that the verbal explanations were identical in all conditions).

Overall Comprehension. Another variable we were interested in was the comprehensibility of the overall description (not comprehensible–easily comprehensible). Although the ANOVA marginally failed to reach significance (*F*(4,105)=2.37, *p*=.057), we analyzed simple effects for experimental conditions. The means for both individual GNetIc conditions significantly outperformed the mean of the *random* gesture condition (*ind-1/random*: *t*(42)=2.46, *p*=.018, CI=[0.18; 1.82]; *ind-2/random*: *t*(41)=2.85, *p*=.007, CI= [0.32;1.86]). In tendency, the *no gesture* mean differed from the *random* mean. That is, participants reported greater comprehension of the presentation when the agent produced no, rather than random gestures.

Gesture's Helpfulness for Comprehension. With regard to perceived helpfulness of gesturing (not helpful–very helpful) we obtained a significant main effect (*F*(4,104)= 25.86, p<.001). Not surprisingly, participants in the *no gesture* condition rated gesturing as less helpful than participants in the other conditions (*t*-test, p<.001 in each case). In addition, gestures in both individual conditions (*ind-1, ind-2*) were rated more helpful than in the *random* condition (*ind-1*: *t*(41)=2.87, *p*=.006, CI=[0.47;2.70]; *ind-2*: *t*(41)=3.63, *p*=.001, CI=[0.77;2.68]).

Vividness. Furthermore, we asked participants to rate the vividness of the agent's conception of the presented content (not vivid–vivid). Random gesturing tended to hamper this impression even more than no gesturing and combined gesturing. Furthermore, the ANOVA revealed a significant main effect (*F*(4,79)=3.50, *p*=.01). Results of *t*-tests showed significant mean differences between both individual GNetIc conditions and the other three conditions (*ind-1/no gesture*: *t*(29)=2.47, *p*=.02, CI=[0.16;2.32]; *ind-1/random gestures*: *t*(30)=2.66; *p*=.01, CI=[0.38;2.63]; *ind-1/combined*: *t*(41)=2.19, *p*=.03, CI=[0.09;2.18]; *ind-2/no gesture*: *t*(22)=2.76, *p*=.01, CI=[0.33;2.43]; *ind-2/random gestures*: *t*(25)=2.91, *p*=.01, CI=[0.38;2.90]; *ind-2/combined*: *t*(31)=2.12, *p*=.04, CI=[0.05; 2.50]). That is, producing gestures with an individualized network helps a virtual agent to create the impression of having a better idea of what is being described in human recipients.

4.2 Agent Perception

We assessed how Max is perceived using several items, e.g. 'pleasant', 'friendly', 'helpful', on seven-point Likert scales (not apprpriate–very appropriate). To measure the re-

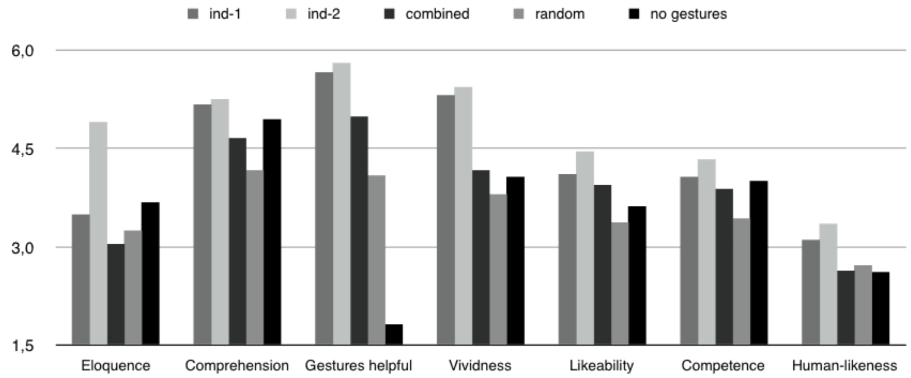


Fig. 3. Mean values of the dependent variables in the five conditions (see Tables 3 and 5 for SDs).

liability of these items we grouped them into three scales ‘likeability’, ‘competence”, and ‘human-likeness’ (see Table 4) and calculated Cronbach’s alpha for the indeces. Alpha values for all three scales were above 0.7, which justifies combining these items into one mean value as a single index for this scale. We analyzed the main effect for experimental conditions by applying ANOVAs and further investigated the pattern of means by computing paired-samples *t*-tests with 95% confidence intervals (CI) for pairwise comparisons between condition means. Mean values and standard deviations are summarized in Table 5 and visualized in Figure 3.

Table 4. Reliability analysis for the three scales ‘likeability’, ‘competence’, and ‘human-likeness’.

Scale	Items	Cronbach’s Alpha
Likeability	pleasant, sensitive, friendly, likeable, affable, approachable, sociable	.86
Competence	dedicated, trustworthy, thorough, helpful, intelligent, organized, expert	.84
Human-likeness	active, humanlike, fun-loving, lively	.79

Likeability. Regarding likeability, we found a significant main effect for experimental conditions (*F*(4,104)=3.88, *p*=.01). Mean ratings for the two individual GNetIc conditions were higher than in the other conditions. In particular, this relationship was significant when comparing the *ind-2* condition with *no gesture* (*t*(36)=2.68, *p*=.01, CI=[0.21;1.48]) and *random* conditions (*t*(38)=3.58, *p*=.001, CI=[0.47;1.67]). The mean difference between *ind-2* and the *combined* condition marginally failed to reach significance (*t*(40)=1.99, *p*=.054; CI=[-0.01;1.02]). For individual condition *ind-1*, the difference of mean evaluation of likeability in comparison with *random* gestures is significant (*t*(42)=2.08, *p*=.05, CI=[0.02;1.43]). In addition, means for the *combined* GNetIc condition were higher than in both control conditions. In other words, all three GNetIc conditions outperformed the control conditions, whereby best evaluations for likeability were obtained by participants in the individual GNetIc conditions.

Competence. With regard to the evaluation of the agent's competence we also found a significant main effect (*F*(4,101)=2.65, *p*=.04). The GNetIc condition *ind-2* received higher mean evaluations than the *random* condition (*t*(42)=3.51, *p*=.001, CI=[0.38;1.42]). The *combined* GNetIc condition also received a higher mean evaluation than the *random* condition which is, however, not significant. Notably, there were no significant differences between the GNetIc conditions and the *no gesture* condition.

Human-likeness. Finally, the analysis of ratings for human-likeness revealed a main effect (*F*(4,104)=2.08, *p*=.09). Both individual GNetIc conditions outperformed the other conditions. Again, this relationship is stronger for the condition *ind-2* (*ind-2*/*no gesture*: *t*(42)=2.40, *p*=.02, CI=[0.12;1.38]; *ind-2*/*random gestures*: *t*(42)= 2.09, *p*=.04, CI=[0.02;1.27]; *ind-2*/combined: *t*(41)=2.30, *p*=.03, CI=[0.09;1.38]). For the other individual GNetIc condition *ind-1*, the mean rating of human-likeness is also higher than in the *combined* GNetIc condition and the two control conditions, but these differences are not significant. No difference was found between the *combined* GNetIc condition and the two control conditions (*random* and *no gesturing*).

Table 5. Mean values for the agent perception scales in the five different conditions (standard deviations in parentheses).

	ind-1	***ind-2***	***combined***	***no gestures***	***random***
Likeability	4.12 (1.18)	4.47 (0.81)	3.95 (0.87)	3.62 (1.24)	3.39 (1.14)
Competence	4.07 (1.11)	4.34 (0.55)	3.89 (0.84)	4.01 (1.09)	3.44 (1.07)
Humanlikeness	3.11 (1.29)	3.38 (1.07)	2.64 (1.01)	2.62 (1.00)	2.73 (0.98)

5 Discussion and Conclusion

The goal of this paper was to evaluate the GNetIc production model and to explore the impact of automatically generated gestures on human-agent interaction. A network learned from a corpus of several speakers was compared with networks learned from individual speaker data, as well as two control conditions (no and random gestures). Results can be summarized in four points: First, Max's gesturing behavior was rated positively regarding gesture quantity and quality, and we found no difference across gesture conditions concerning these issues. Second, both individual GNetIc conditions outperformed the other conditions in that gestures were perceived as more helpful, overall comprehension of the presentation was rated higher, and the agent's mental image was judged as being more vivid. Similarly, the two individual GNetIc conditions outperformed the control conditions regarding agent perception in terms of likeability, competence, and human-likeness. Third, the *combined* GNetIc condition was rated worse than the individual GNetIc conditions throughout. Fourth, the *no gesture* condition was rated more positively than the *random* condition, in particular for the subjective measures of overall comprehension, the gesture's role for comprehension, and vividness of the agent's mental image. That is, with regard to these aspects it seems even better to make no gestures than to randomly generate gestural behavior even though it is still considerably iconic (cf. gestures in Table 2). It is remarkable that the significant effects reported in this paper already occur after the presentation of 45 seconds lasting stimuli each of them containing up to seven gestures. Future research should, however, also investigate how users perceive longer presentations or interactions between agent and user who was just a passive recipient in the present study.

The results reported here bear important and exciting consequences for IVA research. First, from the methodological point of view of building IVAs, we now have evidence that building generative models of co-verbal iconic gesture use, going beyond gesture lexicons is possible and can yield good results with actual users. Notably, we did not reproduce individual speaker's behavior "literally". Rather, we trained the model from their data so as to extract their preferences and strategies in composing gestures. In result, we can say that we obtained models that create novel gestures as if being the respective speaker and users rated the produced gestures positively.

Second, from the point of view of human communication research our results show that computational modeling with IVAs is a highly valuable tool to discover mechanisms and principles of communicative behavior. Here, we explicated process models of how speakers form gestures and we showed that these models actually produce reasonable behavior. Furthermore, different models result in perceivably different behavior, with consistently differing perception and evaluation by human recipients.

Third, and probably most surprising, we found that the common approach to inform behavior models from empirical data by averaging over a population of subjects is not necessarily the best choice. Our findings suggest that modeling individual speakers with proper abilities for the target behavior (in our case good iconic gesturing) results in even better behavior judged from the perspective of human interaction partners. This may be due to the fact that individual networks ensure a greater coherence of the produced behavior. As a consequence, the agent may appear more coherent and self-consistent which, in turn, may make its behavior more predictable and easier to interpret for the user. This is in line with Nass et al. who found that people like ECAs more when they show consistent personality characteristics across modalities [25]. On the contrary, however, Foster & Oberlander recently argued for more variation in the generation of non-verbal behavior based on evidence from the evaluation of automatically produced head and eyebrow motion [10]. In any case, as a consequence for future IVA research, it seems reasonable according to our results to detect particularly appropriate speakers and to individualize agents in their way (e.g., in our data *ind-2* outperforms *ind-1*). This may also help to point up a solution to the task of producing iconic gestures, which is daunting because of the seemingly under-constrained problem of having to pick from a myriad of possible options, which appear to be more or less equivalent and whose contingencies are hardly known. Adhering to an individual style of gesturing can provide additional constraints to resolve this problem of behavior formulation, and it can actually help to produce good behavior and increase the acceptance of the agent.

Such an individualization, however, bears the danger of narrowing acceptance down to a certain population of users, since gesture perception, as well as production, may be subject to inter-individual differences. For instance, Martin et al. found the rating of gestural expressivity parameters to be influenced by human addressee's personality traits [23]. Although our data do not suggest such a risk immediately, since in these conditions standard deviations were not notably higher, an elaborated study should also take the addressee into account. Further, a quest for individualization should, in our view, be accompanied by efforts to also make agents able to deviate from this individualized behavior in reciprocal interaction, in order to achieve inter-personal coordination and to induce social resonance [18].

Acknowledgements

This research is partially supported by the DFG in the CRC 673 "Alignment in Communication" and the Center of Excellence in "Cognitive Interaction Technology".

References

1. K. Bergmann and S. Kopp. GNetIc–Using Bayesian decision networks for iconic gesture generation. In Z. Ruttkay, M. Kipp, A. Nijholt, and H. Vilhjalmsson, editors, *Proceedings of the 9th Intern. Conf. on Intelligent Virtual Agents*, pages 76–89. Springer, Berlin, 2009.
2. K. Bergmann and S. Kopp. Increasing expressiveness for virtual agents–Autonomous generation of speech and gesture in spatial description tasks. In K. Decker, J. Sichman, C. Sierra, and C. Castelfranchi, editors, *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 361–368, Budapest, Hungary, 2009.
3. K. Bergmann and S. Kopp. Modeling the production of co-verbal iconic gestures by learning bayesian decision networks. *Applied Artificial Intelligence*, to appear.
4. T. Bickmore and J. Cassell. Social dialogue with embodied conversational agents. In J. van Kuppevelt, L. Dybkjaer, and N. Bernsen, editors, *Advances in Natural, Multimodal Dialogue Systems*, New York, 2005. Kluwer Academic Publishers.
5. S. Buisine and J.-C. Martin. The effects of speech-gesture cooperation in animated agents' behavior in multimedia presentations. *Interacting with Computers*, 19:484–493, 2007.
6. J. Cassell, M. Stone, and H. Yan. Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of the First Intern. Conf. on NLG*, 2000.
7. J. Cassell and K. Thórisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13:519–538, 1999.
8. R. Dale and J. Viethen. Referring expression generation through attribute-based heuristics. In E. Krahmer and M. Theune, editors, *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 58–65, Athens, Greece, 2009.
9. S. T. Fiske, A. J. Cuddy, and P. Glick. Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Science*, 11(2):77–83, 2006.
10. M. Foster and J. Oberlander. Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, 41:305–323, 2007.
11. M. Gullberg and K. Holmqvist. What speakers do and what listeners look at. Visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition*, 14:53–82, 2006.

12. B. Hartmann, M. Mancini, and C. Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In S. Gibet, N. Courty, and J.-F. Kamp, editors, *Gesture in Human-Computer Interaction and Simulation*, pages 45–55. Springer-Verlag, 2006.
13. D. Heylen, I. van Es, A. Nijholt, and B. van Dijk. Experimenting with the gaze of a conversational agent. In J. van Kuppevelt, L. Dybkjær, and N. Bernsen, editors, *Proceedings International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, pages 93–100, 2002.
14. A. Hoffmann, N. Krämer, A. Lam-Chi, and S. Kopp. Media equation revisited. Do users show polite reactions towards an embodied agent? In Z. Ruttkay, M. Kipp, A. Nijholt, and H. Vilhjálmsson, editors, *Proceedings of the 9th Intern. Conf. on Intelligent Virtual Agents*, pages 159–165, Berlin, 2009. Springer.
15. A. Hostetter and M. Alibali. Raise your hand if you're spatial–Relations between verbal and spatial skills and gesture production. *Gesture*, 7(1):73–95, 2007.
16. M. Huenerfauth. Spatial, temporal and semantic models for American Sign Language generation: Implications for gesture generation. *Semantic Computing*, 2(1):21–45, 2008.
17. A. Kendon. *Gesture–Visible Action as Utterance*. Cambridge University Press, 2004.
18. S. Kopp. Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Communication*, 52:587–597, 2010.
19. 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*, pages 133–160. John Wiley, New York, 2007.
20. S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52, 2004.
21. N. Krämer, B. Tietz, and G. Bente. Effects of embodied interface agents and their gestural activity. In T. Rist, R. Aylett, D. Ballin, and J. Rickel, editors, *Proceedings of the 4th Intern. Workshop on Intelligent Virtual Agents*, pages 292–300, Berlin, 2003. Springer.
22. A. Lücking, K. Bergmann, F. Hahn, S. Kopp, and H. Rieser. The Bielefeld speech and gesture alignment corpus (SaGA). In M. Kipp, J.-C. Martin, P. Paggio, and D. Heylen, editors, *Proceedings of the LREC 2010 Workshop on Multimodal Corpora*, 2010.
23. J.-C. Martin, S. Abrilian, and L. Devillers. Individual differences in the perception of spontaneous gesture expressivity. In *Integrating Gestures*, page 71, 2007.
24. C. Müller. *Redebegleitende Gesten: Kulturgeschichte–Theorie–Sprachvergleich*. Berlin Verlag, Berlin, 1998.
25. C. Nass, K. Isbister, and E.-J. Lee. Truth is beauty: Researching embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 374–402, Cambridge, 2000. MIT Press.
26. 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.
27. J. Oberlander and A. Gill. Individual differences and implicit language: personality, parts-of-speech and pervasiveness. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 1035–1040, Chicago, IL, 2004.
28. M. Rehm and E. André. Informing the design of agents by corpus analysis. In T. Nishida and Y. Nakano, editors, *Conversational Informatics*. John Wiley & Sons, Chichester, 2007.
29. Z. Ruttkay. Presenting in style by virtual humans. In A. Esposito, editor, *Verbal and Nonverbal Communication Behaviours*, pages 23–36. Springer Verlag, Berlin, 2007.
30. T. Sowa and I. Wachsmuth. A computational model for the representation an processing of shape in coverbal iconic gestures. In K. Coventry, T. Tenbrink, and J. Bateman, editors, *Spatial Language and Dialogue*, pages 132–146. Oxford University Press, 2009.
31. J. Streeck. Depicting by gesture. *Gesture*, 8(3):285–301, 2008.