From observation to simulation: generating culture-specific behavior for interactive systems

Matthias Rehm · Yukiko Nakano · Elisabeth André · Toyoaki Nishida · Nikolaus Bee · Birgit Endrass · Michael Wissner · Afia Akhter Lipi · Hung-Hsuan Huang

Abstract In this article we present a parameterized model for generating multimodal behavior based on cultural heuristics. To this end, a multimodal corpus analysis of human interactions in two cultures serves as the empirical basis for the modeling endeavor. Integrating the results from this empirical study with a well-established theory of cultural dimensions, it becomes feasible to generate culture-specific multimodal behavior in embodied agents by giving evidence for the cultural background of the agent. Two sample applications are presented that make use of the model and are designed to be applied in the area of coaching intercultural communication.

M. Rehm (✉) · E. André · N. Bee · B. Endrass · M. Wissner
Faculty of Applied Informatics, Augsburg University,
Universitätsstr. 6a, 86159 Augsburg, Germany
e-mail: rehm@informatik.uni-augsburg.de

E. André
e-mail: andre@informatik.uni-augsburg.de

N. Bee
e-mail: bee@informatik.uni-augsburg.de

B. Endrass
e-mail: endrass@informatik.uni-augsburg.de

M. Wissner
e-mail: wissner@informatik.uni-augsburg.de

Y. Nakano · A. A. Lipi
Faculty of Science and Technology, Seikei University,
3-3-1, Kichijoji-Kitamachi, Musashino,
Tokyo 180-8633, Japan
e-mail: y.nakano@st.seikei.ac.jp

A. A. Lipi
e-mail: 50007646211@st.tuat.ac.jp

T. Nishida · H.-H. Huang
Department of Intelligence Science and Technology,
Kyoto University, Yoshida-Honmachi, Sakyo-ku,
Kyoto 606-8501, Japan
e-mail: nishida@i.kyoto-u.ac.jp

H.-H. Huang
e-mail: huang@ii.ist.i.kyoto-u.ac.jp

1 Introduction

At first sight, culture seems a far-fetched concept to consider for the development of interactive computer systems. But a user's cultural upbringing establishes heuristics for behaving and interpreting behavior in others that are deemed "natural" in a given cultural group and thus strongly influence the user's interactions. This influence is not only apparent in face-to-face encounters but has other direct consequences, for instance how information is evaluated that is presented on website. Marcus (2000) gives some interesting examples on different styles of information presentation on websites that are influenced by cultural parameters. To give another example, Hofstede (2001) reports on a study by Schmidt and Yeh (1992) about influence tactics in different cultures and shows that people from cultures accepting distinct hierarchies (see Sect. 3) tend to argue by invoking a higher authority, whereas people from cultures with flatter hierarchies tend to argue more friendly and by reasoning. Hofstede has termed these heuristics for behaving and interpreting behavior *mental programs*.

If we take the evidence from the literature seriously that users from different cultures interact based on such culture dependent heuristics, then it is necessary to acknowledge these differences in the design of interfaces. In this article we focus on embodied conversational agents (Cassell et al. 2000), which allow the user to interact with virtual characters relying on everyday communicative abilities. Thus,

such agents serve as anthropomorphic communication devices and thus create severe expectations regarding their behavior (verbal as well as non-verbal, see, Reeves and Nass 1996). On the other hand, due to this challenge, embodied conversational agents as an interface metaphor have a great potential to realize culture specific interaction behavior in several fields of human computer interaction:

- Information presentation: By adapting their communication style to the culturally dominant persuasion strategy, agents become more efficient in delivering information or selling a point or a product (persuasive technology).
- Intelligent Tutoring Systems (ITS): (i) For educational purposes, experience-based role-plays become possible, e.g., for increasing cultural awareness of users or for augmenting the standard language textbook with behavioral learning scenarios. (ii) Additionally, on a higher level, cultural adaptation is necessary for the underlying teaching concepts, e.g., realizing a more discussion-based or a more fact-based learning concept (Hofstede 1986).
- Entertainment: Endowing characters in games with their own cultural background has two advantages. It makes the game more entertaining by providing coherent behavior modifications based on the cultural background and it lets characters react in a believable and consistent way to (for them) weird behavior of other agents and the user.

In this article we present work from the international German-Japanese project CUBE-G.[1] Based on a theory of cultural dimensions (Hofstede 2001), we investigate whether and how the non-verbal behavior of agents can be generated from a parameterized computational model. Specifying a culture's position on the basic dimensions allows the system to generate appropriate non-verbal behaviors for the agents. The project combines a top down model-based approach with a bottom-up corpus-based approach which allows empirically grounding the model in the specific behavior of two cultures (Japanese and German), and challenges the following objectives:

1. To investigate how to extract culture-specific behaviors from corpora.
2. To develop an approach to multimodal behavior generation that is able to reflect culture specific aspects.
3. To demonstrate the model in suitable application scenarios.

In the remainder of this article, we review other work in the area of culture-specific interactions (Sect. 2) and present our corpus study including the dimensional theory of culture that we employ in our approach and which is used in most of the related work (Sect. 3). With the empirical data from the corpus study at hand, we realize a Bayesian network model of cultural adaptation (Sect. 4), which then is employed in two different sample applications that illustrate the great potential of culture adaptive systems in the ITS-domain (Sect. 5).

2 Related work

In order to model culture-specific interactive behavior for embodied conversational agents, information of such heuristics has to be available. Unfortunately, the information found in the literature is often too unspecific on a technical level to serve as an empirical basis for modeling the behavior of the agents and make it necessary to collect and analyze multimodal data. The use of such annotated corpora has started to spread over from the social sciences to computer science over the last years due to a number of different reasons. Often data on human interaction is lacking information necessary for developing a model to control the behavior of a conversational agent (e.g., about the synchronization of different modalities). To keep the intuition of the researcher at bay, it is indispensable to collect and annotate this data. Once created, such a database can serve to extract rules or statistical information for behavior generation and analysis or it can serve as a benchmark against which the resulting system can be tested. Especially the last point is interesting for enculturating interfaces and developing conversational agents with a cultural background. A number of large corpora of multimodal behavior already exist but all of them focus on interactions in same-culture groups. Examples of such corpora include the AMI (Augmented Multiparty Interaction) corpus[2] that comprises around 100 h of meeting recordings featuring verbal and non-verbal interactions between multiple interlocutors (Jovanovic et al. 2006). The Smartkom corpus[3] focuses on human computer interaction and was recorded in a Wizard of Oz setting to access users' interaction habits with a virtual character (Wahlster 2006). The SAL corpus (Sensitive Artificial Listener) is mainly concerned with investigating facial expressions of emotion (Douglas-Cowie et al. 2008). Because Ekman (1992) has shown the existence of display rules for emotions that vary from culture to culture, it seems inevitable that the SAL corpus has to be augmented with recordings from different cultures.

[1] CUlture-adaptive BEhavior Generation.

[2] http://corpus.amiproject.org/ (last visited: 02 April 2009).

[3] http://www.bas.uni-muenchen.de/Bas/BasMultiModaleng.html (last visited: 02 April 2009).

Caridakis et al. (2007) give an account on how the data from such a corpus can be used to directly mirror the behavior of a human speaker with an agent. This approach goes under the name of copy synthesis and is limited insofar as the agent can only directly reproduce aspects of the corpus data. A similar approach is described by Kipp et al. (2007). Whereas Caridakis et al. aim at real-time mirroring of human behavior, Kipp et al. try to extract specific behavioral data from the corpora that describe the "style" of the human speaker, which is then mimicked by the agent. A different type of approach tries to extract general behavioral information in the form of statistical data or behavioral rules that can then be employed to control an agent's behavior. Lee and Marsella (2006) extract statistical rules from a corpus of natural dialogues that allow them to generate appropriate head and hand gestures for their agent that accompany the agent's utterances. An example rule would be something like "if the utterance contains a negation, shake the head". Thus, their approach exploits the relation between words and gestures. Nakano et al. (2003) concentrate on grounding phenomena in interactions with virtual characters and also extract rule-like regularities for gaze behavior from a corpus of human interactions. The same corpus is later used to judge the results of the human-agent dialogues. Instead of rules, Rehm and André (2007) have shown how statistical information can be extracted from a multimodal corpus and used as control parameters for a virtual character. To this end they analyzed what kind of relation exists between certain types of gestures and verbal strategies of politeness.

All of the above work focuses on multimodal aspects of interaction and does not regard culture as a crucial parameter although embodied conversational agents are ideal candidates for integrating cultural aspects of interaction. The need to do so has been acknowledged (Payr and Trappl 2004) but there are few systems that actually try to tackle this challenge in a principled manner. De Rosis et al. (2004) illustrate this problem by their survey of the Microsoft Agents web site which shows that the appearance, as well as the animations of the characters are primarily based on Western cultural norms. To make such systems adaptable to cultural differences in interaction behavior, a set of parameters or rules is needed that allow influencing the system processes. Most approaches in this area concentrate on learning environments or interactive role-plays with virtual characters. Khaled et al. (2006) focus on cultural differences in persuasion strategies and present an approach of incorporating these insights into a persuasive game for a collectivist society. Johnson et al. (2004) describe a language tutoring system that also takes cultural differences in gesture usage into account. The users are confronted with some prototypical settings and apart from speech input, have to select gestures for their avatars. Moreover, they have to interpret the gestures by the tutor agents to solve their tasks. Warren et al. (2005) as well as Rehm et al. (2007) aim at cross-cultural training scenarios and describe ideas on how these can be realized with virtual characters. Jan et al. (2007) describe an approach to modify the behavior of characters by cultural variables relying on Hofstede's dimensions. The variables are set manually in their system to simulate the behavior of a group of characters. Whereas all of the above systems focus on existing cultures, Aylett et al. (2009) present a quite different approach introducing an invented culture to teach cultural awareness in an experienced based role-play. It remains to be shown that children really transfer what they learn in this approach to their interactions with real cultures.

Even though there are a number of approaches to simulate culture-specific agents, a principled approach to the generation of cross-cultural behaviors is still missing. Furthermore, there is no empirically validated approach that maps cultural dimensions onto expressive dimensions. In order to realize cultural agents, we need to move away from generic behavior models and instead simulate individualized agents that portray idiosyncratic behaviors, taking into account the agent's cultural background. To this end, we propose a combination of an empirical data-driven and a theoretical model-driven approach, which is presented in the remainder of this article.

3 Comparative corpus analysis

The rationale for creating the CUBE-G corpus was the lack of principled studies analyzing and comparing observational data from different cultures in a standardized way. Our starting point was Hofstede's dimensional model of culture that allows for unambiguously distinguishing given cultures on five dimensions. For each of these dimensions, Hofstede et al. (2002) present what they call synthetic cultures for the endpoints of the dimensions and give details on how non-verbal behaviors differ according to the position on the specific dimension. This is exemplified in the following by the volume of speech and proxemics, i.e., spatial behavior, based on examples taken from Hofstede et al. (2002). As can be seen in the examples, a different position on a given dimension does not necessarily imply a difference in behavior.

1. Hierarchy: This dimension deals among other things, with superiors' decision-making styles and with the decision-making style that subordinates prefer in their boss. Hofstede concludes that more coercive and referent power is used in high-H societies and more

reward, legitimate, and expert power in low-H societies. Whereas individuals from high-H societies tend to speak with a soft voice and stand further apart in face-to-face encounters, those from low-H societies speak rather loud and stand closer together.

2. Identity: The degree to which individuals are integrated into a group is defined with this dimension. On the individualist side, we find societies in which the ties between individuals are loose: everyone is expected to look after him/herself. On the collectivist side, we find societies in which people are integrated into strong, cohesive in-groups. Members of individualistic groups speak louder and stand further apart compared to those from collectivistic groups.
3. Gender: The gender dimension describes the distribution of roles between the genders. In feminine cultures the roles differ less than in masculine cultures, where competition is rather accepted and status symbols are of importance. In more masculine societies it is accepted to speak loud and stand close in face-to-face encounters, whereas in more feminine societies, people tend to speak in a softer voice but also stand close together.
4. Uncertainty: The tolerance for uncertainty and ambiguity is defined in this dimension. It indicates to what extent a culture programs its members to feel either uncomfortable or comfortable in unstructured situations. Unstructured situations are novel, unknown, surprising, or different from usual. Individuals from uncertainty avoiding cultures tend to speak louder and stand further apart than those from uncertainty accepting cultures.
5. Orientation: Values associated with long-term orientation are thrift and perseverance, whereas values associated with short-term orientation are respect for tradition, fulfilling social obligations, and protecting one's face. Long-term orientation might lead to speaking with a soft voice and standing further apart, whereas short-term orientation may lead to talking in a soft voice but standing close together.

To gather information about cultural heuristics in face-to-face interactions, which can serve as an empirical basis for modeling the behavior of an embodied conversational agent, we devised a standardized observational study starting with two cultures that are located on different areas of the Hofstede dimensions, namely Germany and Japan (see Fig. 2). Three prototypical interaction scenarios were defined that are found in every culture to allow for comparing the verbal and non-verbal behavior (see Fig. 1 for an impression).

1. Meeting someone for the first time: This is the standard first chapter of every language learning textbook and one of the most fundamental interactions in everyday communication.
2. Negotiating: Coming to an agreement with others can also be considered as a fundamental interaction especially in intercultural communication. This scenario allows us to compare different negotiation styles and the accompanying verbal and non-verbal behavior.
3. Interacting with higher status individual: Cultures differ in how they interpret the unequal distribution of power and status among the members of the culture, resulting in different behaviors towards interaction partners with a higher status.

These scenarios have been chosen due to two reasons. First of all, we claim that they represent situations every expatriate and even every tourist might easily encounter. Moreover, we expect different verbal and/or non-verbal behavior patterns in the German and the Japanese culture due to their different locations on Hofstede's dimensions. This hypothesis is supported by a number of findings for each of the scenarios. According to Ting-Toomey (1999), the actual greetings at the beginning of the first meeting scenarios can be supposed to take longer in Japan, which is a representative of a collectivistic culture. For individualistic countries, more frequent use of gestures can be expected. For the negotiation task, Teng et al. (1999) give some insights in the organization of the interaction. For short-term oriented (Western) cultures a stronger focus on the task itself can be expected, whereas for long-term oriented (Eastern) cultures a slower and more exhaustive way of problem solving can be expected, where every opinion is taken into account and harmony is at stake resulting in an increased frequency of contributions that are related to communication management. For the third scenario, Leffler et al. (1982) suggest differences in spatial behavior and according to Johnson (1994) differences in the use of verbal facilitators like "yeah" or "mhmm" should occur.

3.1 Design of the study

Dyadic interactions between human subjects were recorded in the three scenarios mentioned above. Table 1 gives an overview of the design. One of the interaction partners in each scenario was an actor following a script for the specific situation. The rationale for using actors was that we would be able to elicit sufficient interactions from the subjects and to control the conditions for each participant more tightly. To control for gender effects, a male and a female actor were employed in each scenario interacting with the same number of male and female subjects.

The actual number of participants differed between Germany and Japan due to some over-recruiting.

Fig. 1 German and Japanese participants interacting in the three prototypical situations

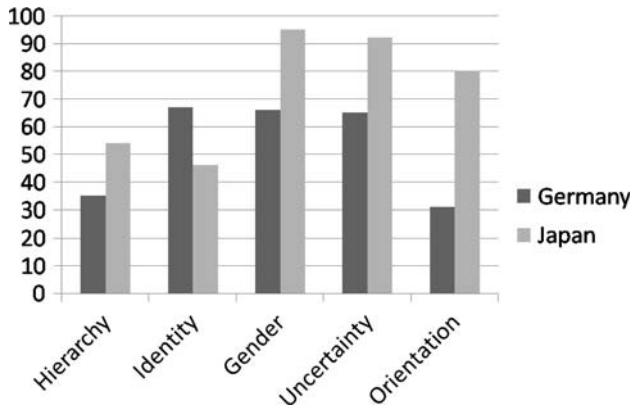

Fig. 2 Germany and Japan on Hofstede's dimensions

Twenty-one subjects (11 male, 10 female) participated in the German data collection, 26 subjects (13 male, 13 female) in the Japanese collection. For each subject, around 25 min of video material was collected, 5 min for the first meeting, 10–15 min for the negotiation, and 5 min for the status difference. Participants were told that they take part in a study by a well-known consulting company for the automobile industry, which would take place at the same time in different countries. To attract their interest in the study, a monetary reward was granted depending on the outcome of the negotiation task. To control for personality traits like extroversion, participants had to fill out a NEO-FFI personality questionnaire (McCrae and John 1992). More information on the specifics of this corpus study can be found in Rehm et al. (2009).

Table 1 Design of the corpus study

First time meeting		Negotiation		Social status	
Actor	Subjects	Actor	Subjects	Actor	Subjects
M_{A1}	M_{S1}–M_{S5}	M_{A1}	M_{S1}–M_{S5}	M_{A2}	M_{S1}–M_{S5}
	F_{S1}–F_{S5}		F_{S1}–F_{S5}		F_{S1}–F_{S5}
F_{A1}	M_{S6}–M_{S10}	F_{A1}	M_{S6}–M_{S10}	F_{A2}	M_{S6}–M_{S10}
	F_{S6}–F_{S10}		F_{S6}–F_{S10}		F_{S6}–F_{S10}

The study was conducted to shed light on pertinent non-verbal behavior patterns found in the two cultures for the

Table 2 Posture types frequently observed in German and Japanese data

	German	Japanese
Head	THdAP (turn head away from person)	THdAP
Leg	LSF (lean sideways on foot)	LSF
Arm	PHIPt (put hand(s) into pocket)	PHFe (put hand to face)
	PHEw (put hand to elbow)	PHWr (put hand to wrist)
	FAs (fold arms)	JHs (join hands)

three scenarios. To this end, the analysis concentrated on posture and gestural activity as two prominent non-verbal behaviors.

3.2 Posture analysis

We employ Bull's (1987) posture coding scheme to categorize posture shifts observed in our corpus. For the current analysis, we annotated head, arm, and leg postures for 8 German and 9 Japanese first time meeting conversations. Table 2 describes the frequently observed categories in German and Japanese data.

3.3 Leg posture analysis

The average number of leg posture shifts in the German data was 9.5 and that in the Japanese data was 16.56 per conversation. A weak trend was found in a *t*-test ($t(15) = 1.764$, $p < 0.1$). The average duration of each posture in the German data was 19.93 sec and that in the Japanese data was 24.64 sec, but the difference was not statistically significant ($t(15) = 0.409$, ns). LSFs (lean sideways on foot) were observed most frequently in both countries.

3.4 Arm posture analysis

The average number of arm posture shifts in the German data was 40.38 and that in the Japanese data was 22.8 per conversation. A weak trend was found in a *t*-test ($t(16) = 1.931$, $p < 0.1$). On the other hand, the average duration of each posture in the German data was 7.79 sec and that in the Japanese data was 14.08 sec ($t(16) = 2.061$, $p < 0.1$).

More interestingly, posture shapes were also very different depending on the country. Hand-to-head postures more frequently occurred in the Japanese data than the German data, and PHFe (put hand to face) was the most frequent in the Japanese data. One-handed postures were very different depending on the culture. The most frequent category in the German data was PHEw (put hand to elbow), and that in the Japanese data was PHWr (put hand to wrist). Intriguingly, German people rarely did PHWr, and Japanese people rarely did PHEw. For two-handed postures, German people mainly used their arms, such as folding their arms (FAs) and putting their hands on the elbows (PHEw). On the contrary, Japanese people mainly used their hands, such as joining the hands (JHs), putting their hands on the wrists (PHWr). Hand-to-cloth postures were rarely observed in the Japanese data, but, especially for PHIPt (put hand into pocket), they were very frequent in the German data.

3.5 Discussion of posture analysis

Generally, head postures and leg postures were not very different depending on the culture. The most frequent head posture in both countries is THdAP (turn head away from person), which is a typical turn taking signal observed at the beginning of a new turn (Duncan 1974). Such communication signals are similar in both countries. Cultural difference was clearer in arm postures. German people more frequently changed arm postures than Japanese people, and Japanese people kept the same posture longer than German. Arm posture shapes were also very different. German people mainly used their arms. On the contrary, Japanese people mainly used their hands, and their postures looked smaller and less powerful than German postures. Moreover, Japanese people frequently touched their heads by their hands, and German people put their hands in the pockets. Although Japanese people did not move their upper bodies as frequently as German people, they used more leg postures.

In addition to these results above, we also found that the total number of posture shifts per conversation was not different depending on the culture: 71.88 in the German data and 58.56 in the Japanese data, ($t(15) = 1.154$, ns). All these results suggest that the frequency of posture shifts is not different depending on the culture, but the posture shape is one of the important factors for characterizing the culture.

3.6 Difference in gestural expressivity

The coding scheme and the analysis of gestural expressivity follow Pelachaud (2005). So far, the first meeting scenarios have been annotated for both cultures. Gestural expressivity was coded for the five parameters repetition, fluidity, power, speed, and spatial extent. Each parameter was coded using a seven-point scale, where 1 denotes small values and 7 large values for the parameter (except for repetition where it denotes the number of repetitions of a given gesture). The distinction between power and speed is taken over from Bevacqua et al. (2006). In order to gain insights in the supposed differences in the use of gestures, we compared expressivity parameters of the German and the Japanese samples. Moreover, we looked into gender

Table 3 General results for gestural activity

	#Gesture/min	#Adaptor/min	GA ratio
G	1.80	0.92	3.38
JP	1.65	1.60	1.84
F	0.120	4.770*	2.272

* $p < 0.05$

specific differences. Some general statistics about gestural activity are given in Table 3 (ANOVA). Due to the slightly different length of the video recordings, number of gestures and adaptors is normalized and given in number of gestures/adaptors per minute. The differentiation between gestures and adaptors follows McNeill's (1992) categorization and was suggested by the material because we observed more frequent use of self-touching hand movements for the Japanese participants. Number of gestures per minute is comparable in both cultures but the number of adaptors is significantly higher for the Japanese samples. This effect does not carry over to the gesture-adaptor ratio. Regarding gestural expressivity, the analysis revealed highly significant differences for all parameters. Table 4 gives the results for this analysis (ANOVA). Compared to the Japanese participants, Germans repeat gestures less, have more fluid motions, gesture more powerful and faster and use more space in gesturing. The gender-specific analysis revealed some additional effects. For the German samples, the duration of gestures is significantly longer for female participants. Regarding the Japanese samples, a weak trend has been found for spatial extent with male participants using more space (see Table 5 for both results (ANOVA)). At last we looked into the influence of the interaction partner's gender on behavior (Table 6 (ANOVA)). For male Germans the only significant effect could be seen for the duration of gestures. Interacting with females, participants' gestures took longer. We found the most effects with the Japanese male participants that used significantly shorter gestures with females. But at the same time gestures were significantly more fluid, powerful, and faster. Additionally, a weak trend could be seen for more expansive gestures. The only effect for female Japanese participants was significantly more powerful gestures with other females.

Table 4 Results of expressivity analysis

	Repetition	Fluidity	Power	Speed	Sp. Ext.	Duration
G	1.43	3.96	3.50	4.32	3.23	1.64
JP	1.90	3.48	2.75	3.33	2.67	3.55
F	18.264**	68.434**	57.998**	99.144**	22.688**	63.853**

** $p < 0.01$

3.7 Discussion of gestural expressivity

The results show how gestures are expressed in the two cultures and reveal strong differences for the examined parameters. Reasons for two of the differences are apparent from the video recordings. Higher duration of gestures for the Japanese participants is attributable to long holds of the stroke. Figure 3 gives an impression. In the depicted example, the position is held for 11 s after the gesture stroke has been performed. Such a prolonged hold happened only once in the German data but frequently occurred with the Japanese participants. Less spatial extent is attributable to the fact that Japanese participants in general perform gestures only with the lower arms, whereas this is rarely seen in the German samples. Figure 4 gives an impression of this difference. The gender-specific analyses gave no conclusive picture except for the male Japanese participants that obviously adapted their gestural activity to the gender of their interaction partners.

With the results of our corpus study at hand, the next section describes in detail how this statistical information can now be employed to model culture-specific non-verbal behavior for embodied conversational agents.

4 A Bayesian network model of culture-specific non-verbal behavior

To adapt a system's interactive behavior to the (assumed) cultural background of the user, two challenges have to be tackled. First, the user's cultural background has to be inferred preferably from his observable behavior. Second, the system has to generate culturally adequate behavior

Table 5 Results of gender-specific expressivity analysis for both cultures

Culture	Gender	Repetition	Fluidity	Power	Speed	Sp. Ext.	Duration
German	Male	1.48	3.95	3.54	4.38	3.23	1.51
	Female	1.34	3.98	3.41	4.21	3.23	1.89
	F	0.816	0.250	0.603	0.816	0.000	7.219**
Japanese	Male	2.03	3.50	2.66	3.21	2.82	3.69
	Female	1.82	3.47	2.80	3.40	2.58	3.47
	F	1.650	0.098	1.355	2.615	2.802^{+}	0.314

$^{+}$ $p < 0.1$, ** $p < 0.01$

Table 6 Results of gender-specific analysis taking the gender of the interaction partner into account

Culture	Gender	Condition	Repetition	Fluidity	Power	Speed	Sp. Ext.	Duration
German	Male	Same	1.33	3.98	3.67	4.49	3.47	1.27
		Mixed	1.56	3.94	3.47	4.31	3.10	1.65
		F	0.516	0.179	1.137	0.655	1.848	7.207**
	Female	Same	1.46	3.96	3.29	4.18	3.36	1.96
		Mixed	1.24	4.00	3.52	4.24	3.12	1.82
		F	1.332	1.182	0.560	0.048	0.518	0.255
Japanese	Male	Same	2.16	3.36	2.50	3.02	2.60	4.56
		Mixed	1.87	3.66	2.85	3.45	3.09	2.61
		F	1.109	4.387*	5.113*	5.833*	3.889^{+}	8.444**
	Female	Same	1.86	3.43	2.97	3.42	2.63	3.68
		Mixed	1.75	3.55	2.46	3.38	2.48	3.04
		F	0.308	1.294	8.157**	0.068	0.812	1.825

$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

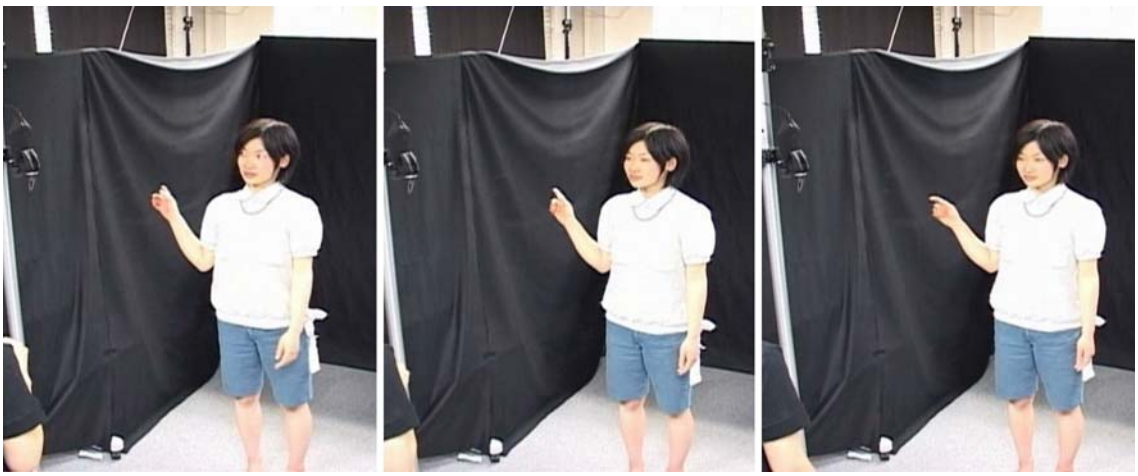

Fig. 3 Prolonged hold of Japanese participant. Images taken at 1:48 min, 1:51 min, and 1:57 min

based on this information. We propose to use Bayesian networks as they address both challenges in a single model. Bayesian networks as described in Jensen (2001) are a formalism to represent probabilistic causal interactions. By modeling such causal relations between concepts they allow for two different types of inferences, causal inferences that follow the causal interactions from cause to effect, and diagnostic inferences that allow for introducing evidence for effects and infer the most likely causes of these effects.

Based on Hofstede's theoretical approach of cultural dimensions, we exploit the relation between a culture's position on Hofstede's dimensions and observable behavior in these cultures like gestural expressivity or postural preferences. Causes in this model are then the positions of a culture on the single dimensions and corresponding effects are observable behaviors like speed or spatial extent of gestures. Following our corpus analysis, we created two different Bayesian networks, one concentrating primarily on gestural expressivity, and the second one concentrating on the effects on posture. This division is not essential and a next step will be the integration into one large network. But for reasons of clarity, the division is kept for the rest of the article.

4.1 Expressivity model

A culture's position on the five dimensions is reduced to two values, high and low, which allow reducing the complexity of the modeling endeavor. Observable behavior is given in three different gradations high, medium, and low. Because the gender-specific analysis was not conclusive, only the results for culture-specific differences have been integrated. The model that was created for expressive behavior does not only take the gestural behavior into account but is extended with information from the

Fig. 4 Difference in using upper and lower arms for Japanese and German participants

literature concerning synthetic cultures to capture additional non-verbal behavior, i.e., proxemics (spatial behavior) and volume (loudness of speech). Whereas the data concerning gestural expressivity derives from our corpus analysis, the other data comes from Hofstede et al. (2002). The model allows us to tackle the above mentioned two challenges:

1. Inferring the user's cultural background: The user's gestural activity is analyzed (e.g., Rehm et al. 2008) and set as evidence for the output nodes of the Bayesian network. A diagnostic inference then yields the most likely causes, i.e., the most likely positions on Hofstede's dimensions, which again can be used to infer the user's cultural group. Additionally, making use of the cultural dimensions allows abstracting from the specific culture of the user to a distribution on the five dimensions. Thus, deviating behavior of the user, i.e., behavior that is not in accordance to known patterns of behavior for the user's culture, results in a different interpretation of the single user's position on the cultural dimensions. Thus, we capture the effect that cultural patterns of behavior are group phenomena and that individuals can deviate from these heuristics. It remains to be shown if the user is then irritated by the system's behavior which is not in accordance with his "real" cultural background.
2. Setting the agent's non-verbal behavior: In this case, the Bayesian network delivers information about dominant patterns of behavior in a culture that is found at the corresponding locations of the cultural dimensions, for instance low on hierarchy, low on identity, high on gender, medium on uncertainty, high on orientation. This results in a probability distribution for each behavior, e.g., for volume the probabilities are 70% high, 29% medium, and 1% low. In Sect. 5.1, this information is used directly to set the behavior of a group of agents, who will then speak with high volume.

4.2 Posture model

We also established a parallel model for arm posture prediction by employing Hofstede's five dimensions as a middle layer in a Bayesian network. Since arm postures strongly characterize the cultures, we focused on modeling arm posture prediction. The behavioral layer was designed similarly to that in the Bayesian network for gesture expressivity prediction. A few nodes which are not suitable in characterizing posture shifts, such as "Speed", were deleted. "Mirroring" was added as a specific aspect in posture shifts. To specify the values for "Spatial extent" and "Power", we conducted an experiment using 10 German and 10 Japanese subjects. The subjects looked at posture shift video clips, and rated Spatial extent and Power of each posture in the video using a seven point Likert scale. "Mirroring", "Frequency", and "Duration" were assigned by calculating the average numbers observed in our corpus data. Our posture model is shown in Fig. 5. Although we admit that a formal model evaluation is necessary, our model outputs reasonable predictions for both German and Japanese postures. As shown in the figure, when Japanese is assigned as evidence at the top node, the model predicts small spatial extent, less power, more frequent mirroring, less posture shift occurrence, and longer duration. These results are very similar to what we found in the empirical study in Sect. 3.2.

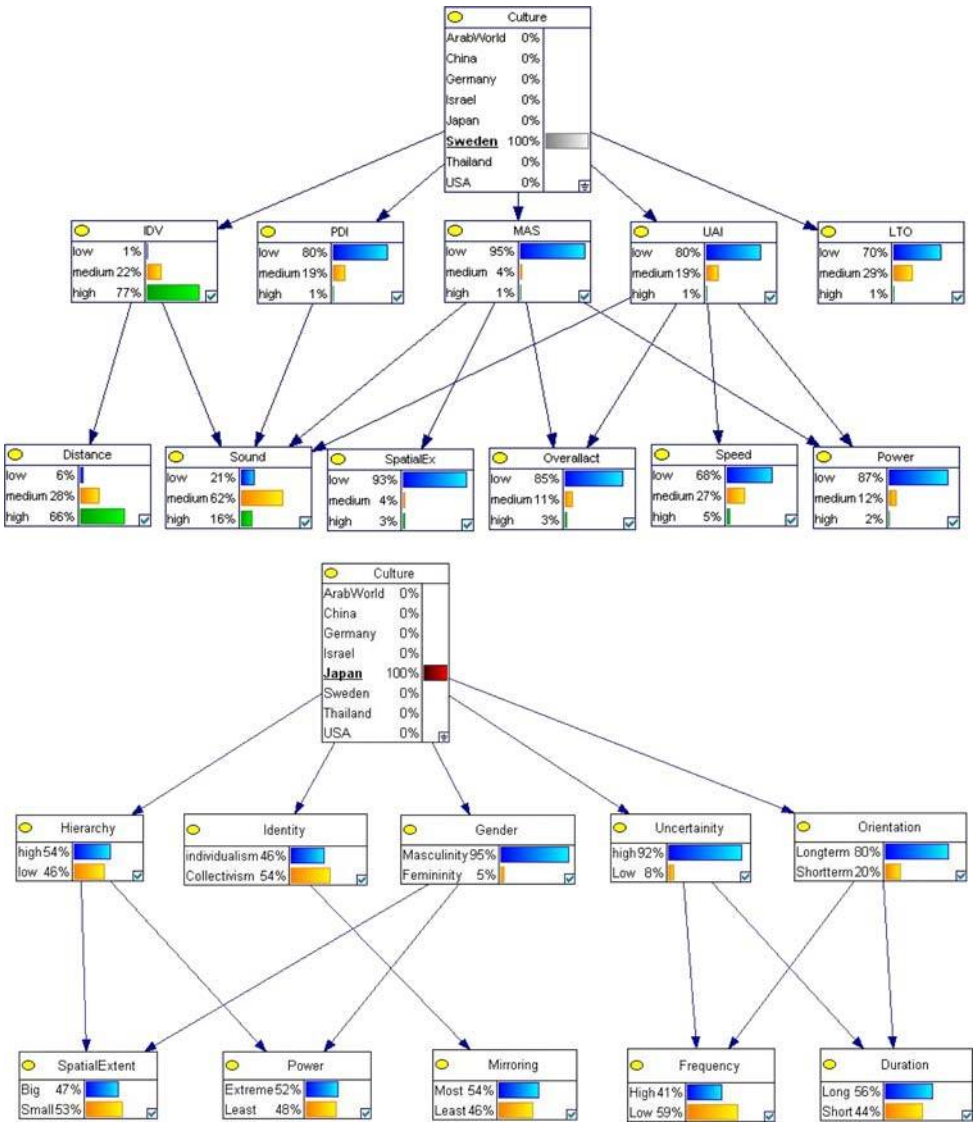

Fig. 5 Bayesian networks for expressivity (*above*) and posture prediction (*below*)

5 Simulating culture-specific non-verbal behavior with embodied conversational agents

Two prototypes have been developed to test the applicability of the Bayesian network modeling of culture-adaptive behavior. The first one aims at increasing the user's awareness of cultural differences in expressive non-verbal behavior and is called the cultural mirror. The second one is a system that supports distance learning of culturally adequate behavior by the use of animated agents.

5.1 The cultural mirror

The most severe misunderstandings in intercultural communication stem from differences in non-verbal behavior (e.g., Ting-Toomey 1999). The reason is a missing awareness of these differences. According to Hofstede, culture gives us heuristics for behavior that are deemed "natural" by members of a given culture. Thus, such heuristics become apparent mainly when confronted with behavior that deviates from this implicit norm. But assuming that one's own behavior is the "natural" one, such deviating behavior is often interpreted as "wrong". Thus, training programs for intercultural communication in general start with increasing the awareness of cultural differences and that behaviors are just different not "right" or "wrong" (e.g., Hofstede 1991; Bennett 1986).

To further such an approach, we developed the cultural mirror that lets a user explore differences in non-verbal behavior based on his own gestural expressivity. Analyzing the user's gestural expressivity, the classification result is set as evidence for the output nodes of our network model. To infer the user's cultural background we make use of acceleration based gesture recognition with

the Wiimote. In Rehm et al. (2008), we have shown that the Wiimote is suitable for such an approach and allows to reliably classify the gestural expressivity of the user. The user's expressivity is analyzed by using the classification result for the expressive dimensions (power, speed, spatial extent) as evidence for the output nodes of the Bayesian network. By a diagnostic inference, the user's cultural background is estimated and this information is then set as evidence to the input nodes of a second network. A causal inference results in a probability distribution for the different observable behaviors of the agents, i.e., spatial behavior, volume of speech as well as gestural expressivity. A group of agents is animated making use of this information and resulting in behavior that is congruent to the user's input.

Figure 6 gives examples of different user input and the resulting behavior of the agents. In the first case, gestural activity in terms of spatial extent, speed, power, and activation is generally low. The cultural background of the user is inferred as Swedish with a high position on the identity dimension and low positions on the other dimensions. In the second case, the user exhibits high spatial extent and low speed. The other parameters are not set exemplifying the advantage of such a model that is able to cope with incomplete information. Based on this evidence the user's cultural background is inferred as probably Chinese or US American with a slightly higher probability for Chinese. Thus, for the behavior generation, Chinese is set as evidence. It remains to be seen which kind of decision procedure should be implemented at this step. The agents directly react to the user's gestural activity and adapt their behavior accordingly. Based on this sample application, we envision a system to increase the user's awareness of cultural difference in behavior patterns by letting the agents react to the user's input based on their own cultural background. For instance, if low gestural activity with low spatial extent is preferred in a given cultural setting and the user exhibits powerful and expansive gestures, the agent could react irritated by this display, allowing the user to examine different reactions of the agents to different patterns of behavior in an embarrassment-free way.

5.2 Distance learning of non-verbal behavior

As another direction which focuses on automatic generation of cultural specific non-verbal behaviors, this section presents conversational agents that play as partners in distance language learning. This technology not only allows the users to present their cultural background without showing their real pictures, but also gives the participants the opportunity to learn non-verbal behaviors of their partners when they learn the language.

Figure 7 shows an overview of the system usage. A student first chooses which language she wants to learn. When she chooses Japanese, a human Japanese teacher types in Japanese texts. The text is sent to a TTS and appropriate postures are determined by a Posture selection mechanism.

The architecture for selecting appropriate postures is given in Fig. 8. Basically it is divided into three main modules. The input to the mechanism is a country name and a text that the agent speaks, which is produced by a TTS.

The Probabilistic Inference Module takes country name as input and outputs the non-verbal parameters for that country. In computing the parameters, this module refers to our Bayesian network model given in Fig. 5. We used the JAVA version of Netica as an inference engine. The outputs of this module are values of non-verbal expressive parameters of each culture: spatial extent, power, duration, and frequency.

The Decision Module is the most important in determining appropriate postures. This module has two sub-modules. Posture computing sub-module takes the estimation results from the Bayesian network (BN) as inputs, and uses them as weights for each empirical data. Then, it calculates the sum of all the weighted values for each posture using the equation given below, and finally outputs a list of all the postures as the posture candidates.

$$\begin{aligned}\text{Posture Score} = {} & \text{bse} * \text{SE} + \text{bpw} * \text{PW} \\ & + \text{bfr} * \text{FR} + \text{bdu} * \text{DU}\end{aligned}$$

Note that the parameters indicated by capital letters are scores from empirical data, and those in small letters are probabilities obtained from the BN. SE: Spatial Extent score in the empirical study, bse: Spatial Extent probability in Bayesian network model, PW: Power, bpw: probability for Power in BN model, FR: Frequency, bfr: probability for Frequency in BN model, DU: Duration, bdu: probability for Duration in BN model.

An example of how a posture score for PHFe (put hand to face) is calculated is shown below.

$$\begin{aligned}\text{PHFe} = {} & \{(0.5183 * 4.19) + (0.507 * 4.4) + (0.58 * 2.725) \\ & + (0.56 * 1.01)\} * 10 = 65.49\end{aligned}$$

where 0.5183, 0.507, 0.58, and 0.56 are weights for spatial extent, power, frequency, and duration, respectively, which are given by the Bayesian network. On the other hand, 4.19, 4.4, 2.725, and 1.01 are values obtained from our empirical studies in Sect. 3.[4] Then, the Decision Module selects appropriate postures by checking the thresholds for a given country. In the previous example, the score for

[4] Since various kinds of measures were used in the empirical data, they are normalized into 1 to 7.

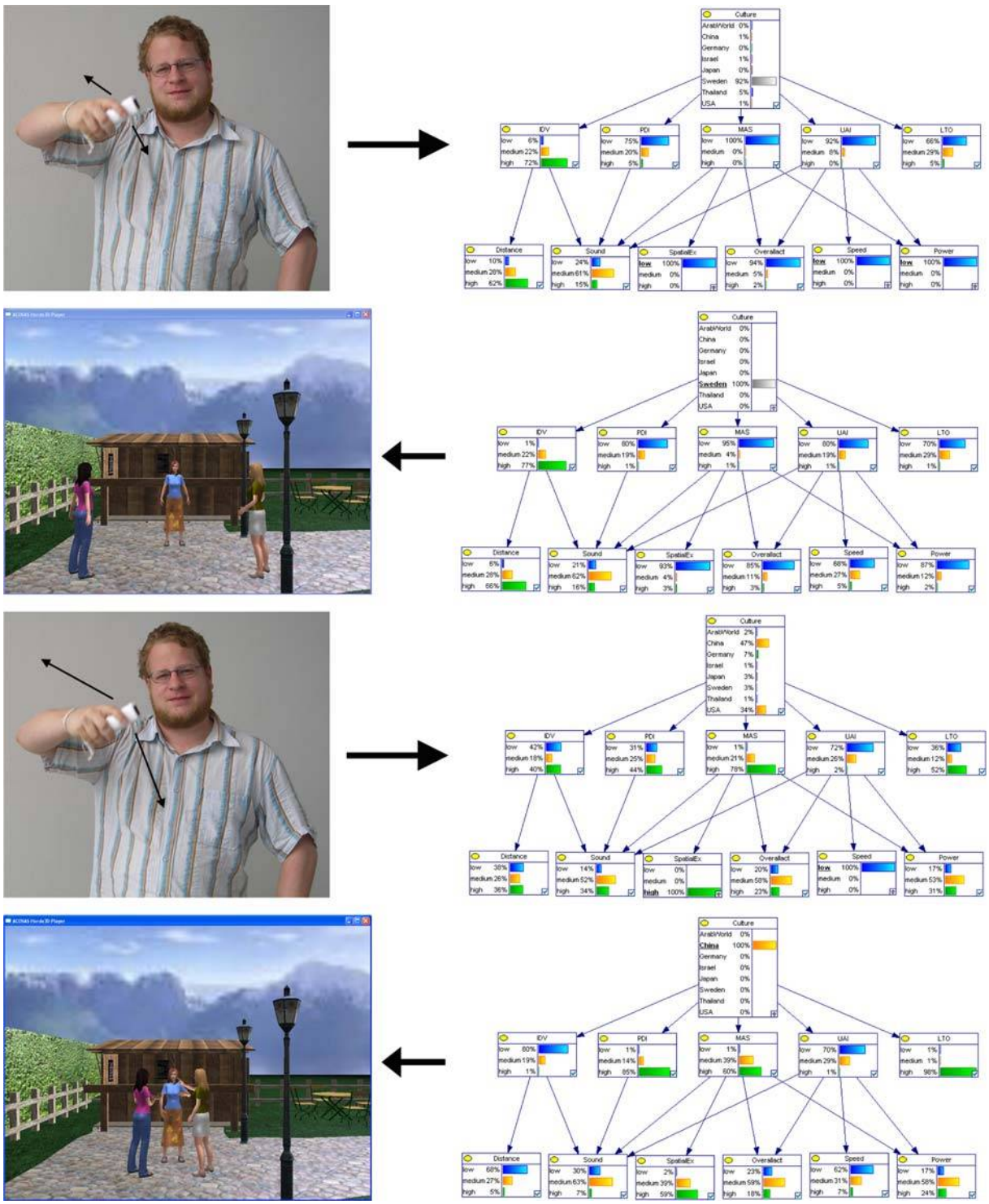

Fig. 6 Analyzing the user's cultural background (diagnostic inference) and setting culture-specific agent behavior (causal inference). The effects of the cultural mirror are exemplified for two cultures that differ on Hofstede's dimensions

PHFe is 65.49, which is judged as an appropriate Japanese-like posture.

The Generation module takes postures recommended by the Decision Module and looks for the animation file for that posture in the animation database. When it finds the animation file, it sends a request to the Horde3D animation engine to generate the animation file, while it also sends the text to Hitachi Hit Voice TTS to convert the text into a wav

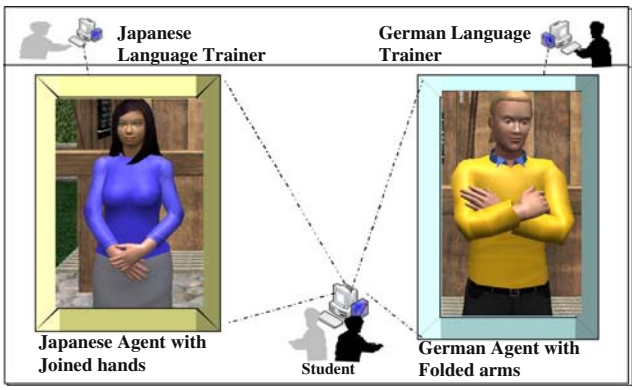

Fig. 7 Language trainer agent

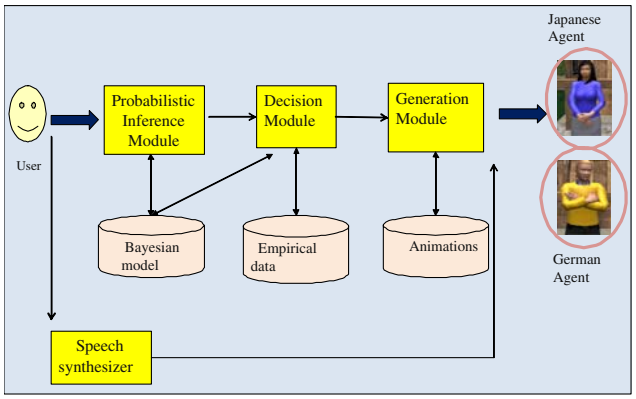

Fig. 8 A simplified architecture of the system

file. Finally, the speech sound and the culture specific posture animations are produced on the student's computer display shown in Fig. 7.

Thus, the system not only teaches language, but also makes the user familiar with the culture-specific non-verbal behaviors. We hope that this system can be used as a distance-learning system by which a user can train by herself how to smoothly communicate with people from other cultures.

6 Conclusion

In this article, we presented our approach of generating culture-specific behaviors in embodied agents that relies on a bottom-up empirical approach by collecting and analyzing data of human interactions and combines it with a top-down model-based approach that relies on a theory of cultural dimensions that has been proven successful in other areas. The corpus of multimodal behavior was collected under standardized conditions for three prototypical scenarios in two cultures, Germany and Japan. It was argued, that such a principled approach is needed to endow conversational agents with culture-specific verbal and non-verbal behavior which will further the successful use of such agent systems in the area of information presentation, persuasion, and edutainment. The analysis of the corpus data focused on specific non-verbal aspects of communication, body posture and gestural expressivity. For both aspects of behavior, differences between the cultures were found on different levels of granularity. The results have been integrated in a probabilistic model for generating agent behaviors and two sample applications have been developed that exemplify the use of this model.

Body posture as well as gestural expressivity is not only determined by one's cultural background. Indeed, the cultural background only gives general behavioral heuristics which might, e.g., result in preferring higher spatial extent. But such behaviors are also dependent on personality or personal style. This was not taken into account in the analysis presented here. To test for influences of personality on observed behavior, every participant had to do a NEO-FFI personality test (McCrae and John 1992). The results from these tests will allow us to analyze correlations between personality traits of our participants and behavior patterns.

Acknowledgments The work described in this article is funded by the German Research Foundation (DFG) under research Grant RE 2619/2-1 (CUBE-G) and the Japan Society for the Promotion of Science (JSPS) under a Grant-in-Aid for Scientific Research (C) (19500104).

References

Aylett R, Paiva A, Vannini N, Enz S, André E, Hall L (2009) But that was in another country: agents and intercultural empathy. In: Proceedings of AAMAS

Bennett MJ (1986) A developmental approach to training for intercultural sensitivity. Int J Intercult Relat 10(2):179–195

Bevacqua E, Raouzaiou A, Peters C, Caridakis G, Karpouzis K, Pelachaud C, Mancini M (2006) Multimodal sensing, interpretation and copying of movements by a virtual agent. In: PIT. pp 164–174

Bull PE (1987) Posture and gesture. Pergamon Press, Oxford

Caridakis G, Raouzaiou A, Bevacqua E, Mancini M, Karpouzis K, Malatesta L, Pelachaud C (2007) Virtual agent multimodal mimicry of humans. Lang Resour Eval 41:367–388

Cassell J, Sullivan J, Prevost S, Churchill E (eds) (2000) Embodied conversational agents. MIT Press, Cambridge

De Rosis F, Pelachaud C, Poggi I (2004) Transcultural believability in embodied agents: a matter of consistent adaptation. In: Payr S, Trappl R (eds) Agent culture: human-agent interaction in a multicultural world. Lawrence Erlbaum Associates, London, pp 75–106

Douglas-Cowie E, Cowie R, Cox C, Amir N, Heylen D (2008) The Sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In: Proceedings of LREC workshop on corpora for research on emotion and affect

Duncan S (1974) On the structure of speaker-auditor interaction during speaking turns. Lang Soc 3:161–180

Ekman P (1992) Telling lies—clues to deceit in the marketplace, politics, and marriage, 3rd edn. Norton and Co. Ltd, New York

Hofstede G (1986) Cultural differences in teaching in learning. Int J Intercult Relat 10:301–320

Hofstede G (1991) Cultures and organisations—intercultural cooperation and its importance for survival, software of the mind. Profile Books

Hofstede G (2001) Cultures consequences: comparing values, behaviors, institutions, and organizations across. Nations Sage Publications, Thousand Oaks, London

Hofstede GJ, Pedersen PB, Hofstede G (2002) Exploring culture: exercises, stories and synthetic cultures. Intercultural Press, Yarmouth

Jan D, Herrera D, Martinovski B, Novick D, Traum D (2007) A computational model of culture-specific conversational behavior. In: Pelachaud C, Martin J-C, André E, Chollet G, Karpouzis K, Pelé D (eds) Intelligent virtual agents (IVA'07). Springer, Berlin, pp 45–56

Jensen FV (2001) Bayesian networks and decicion graphs. Springer, New York

Johnson C (1994) Gender, legitimate authority, and leader-subordinate conversations. Am Sociol Rev 59:122–135

Johnson WL, Choi S, Marsella S, Mote N, Narayanan S, Vilhjálmsson H (2004) Tactical language training system: supporting the rapid acquisition of foreign language and cultural skills. In: Proceedings of InSTIL/ICALL—NLP and speech technologies in advanced language learning systems

Jovanovic N, Akker ROD, Nijholt A (2006) A corpus for studying addressing behavior in multi-party dialogues. Lang Resour Eval 40(1):5–23

Khaled R, Biddle R, Noble J, Barr P, Fischer R (2006) Persuasive interaction for collectivist cultures. In: Piekarski W (ed) The seventh Australasian user interface conference (AUIC 2006). pp 73–80

Kipp M, Neff M, Kipp KH, Albrecht I (2007) Towards natural gesture synthesis: evaluating gesture units in a data-driven approach to gesture synthesis. In: Pelachaud C et al (eds) Intelligent virtual agents (IVA), Berlin, Springer, pp 15–28

Lee J, Marsella S (2006) Nonverbal behavior generator for embodied conversational agents. In: Gratch J et al (eds) Intelligent virtual agents, Berlin, Springer, pp 243–255

Leffler A, Gillespie DL, Conaty JC (1982) The effects of status differentiation on nonverbal behavior. Soc Psychol Q 45(3):153–161

Marcus A (2000) Cultural dimensions and global web user-interface design: what? So What? Now What? In: HFWEB2000 conference proceedings

McCrae RR, John OP (1992) An introduction to the five factor model and its applications. J Personality 60:175–215

McNeill D (1992) Hand and mind—what gestures reveal about thought. The University of Chicago Press, Chicago, London

Nakano YI, Reinstein G, Stocky T, Cassell J (2003) Towards a model of face-to-face grounding. In: Proceedings of the association for computational linguistics. pp 553–561

Payr S, Trappl R (eds) (2004) Agent culture: human-agent interaction in a multicultural world. Lawrence Erlbaum Associates, London

Pelachaud C (2005) Multimodal expressive embodied conversational agents. In: Proceedings of ACM Multimedia. pp 683–689

Reeves B, Nass C (1996) The media equation—how people treat computers, television and new media like real people and places. Cambridge University Press, Cambridge

Rehm M, André E (2007) More than just a friendly phrase—multimodal aspects of polite behaviors in agents. In: Nishida T (ed) Conversational informatics. Wiley, Chichester, pp 69–84

Rehm M, André E, Nakano Y, Nishida T, Bee N, Endrass B, Huang H-H, Wissner M (2007) The CUBE-G approach—coaching culture-specific nonverbal behavior by virtual agents. In: Mayer I, Mastik H (eds) Proceedings of ISAGA 2007

Rehm M, Bee N, André E (2008) Wave like an Egyptian—acceleration based gesture recognition for culture-specific interactions. In: Proceedings of HCI 2008 culture, creativity, interaction. pp 13–22

Rehm M, André E, Bee N, Endrass B, Wissner M, Nakano Y, Lipi AA, Nishida T, Huang H-H (2009) Creating standardized video recordings of multimodal interactions across cultures. In: Kipp M et al (eds) Multimodal copora. Berlin, Springer

Schmidt SM, Yeh RS (1992) The structure of leader influence: a cross-national comparison. J Cross-Cult Psychol 23:251–264

Teng JTC, Calhoun KJ, Cheon MJ, Raeburn S, Wong W (1999) Is the east really different from the west: a cross-cultural study on information technology and decision making. In: Proceedings of the 20th international conference on Information Systems. pp 40–46

Ting-Toomey S (1999) Communicating across cultures. The Guilford Press, New York

Wahlster W (ed) (2006) SmartKom: foundations of multimodal dialogue systems. Springer, Berlin

Warren R, Diller DE, Leung A, Ferguson W, Sutton JL (2005) Simulating scenarios for research on culture and cognition using a commercial role-play game. In: Kuhl ME, Steiger NM, Armstrong FB, Joines JA (eds) Proceedings of the 2005 winter simulation conference