Chapter 6
Combining a Data-Driven and a Theory-Based Approach to Generate Culture-Dependent Behaviours for Virtual Characters

Birgit Lugrin, Julian Frommel and Elisabeth André

Abstract To incorporate culture into intelligent systems, there are two approaches that are commonly proposed. Theory-based approaches that build computational models based on cultural theories to predict culture-dependent behaviours, and data-driven approaches that rely on multimodal recordings of existing cultures. Based on our former work, we present a hybrid approach of integrating culture into a Bayesian Network that aims at predicting culture-dependent non-verbal behaviours for a given conversation. While the model is structured based on cultural theories and theoretical knowledge on their influence on prototypical behaviour, the parameters of the model are learned from a multimodal corpus recorded in the German and Japanese cultures. The model is validated in two ways: With a cross-fold validation we estimate the power of the network by predicting behaviours for parts of the recorded data that were not used to train the network. Secondly we performed a perception study with virtual characters whose behaviour is driven by the calculations of the network and are rated by members of the German and Japanese cultures. With this chapter, we aim at giving guidance for other culture-specific generation approaches by providing a hybrid methodology to build culture-specific computational models as well as potential approaches for their evaluation.

B. Lugrin (✉)
Human-Computer Interaction, University of Wuerzburg, Am Hubland (M1),
97074 Würzburg, Germany
e-mail: birgit.lugrin@uni-wuerzburg.de

J. Frommel
Institute of Media Informatics, Ulm University, James-Franck-Ring,
89081 Ulm, Germany
e-mail: julian.frommel@uni-ulm.de

E. André
Human Centered Multimedia, Augsburg University, Universitätsstr. 6a,
86159 Augsburg, Germany
e-mail: andre@hcm-lab.de

C. Faucher (ed.), *Advances in Culturally-Aware Intelligent Systems and in Cross-Cultural Psychological Studies*, Intelligent Systems Reference Library 134, https://doi.org/10.1007/978-3-319-67024-9_6

Keywords Virtual Agents · Culture · Non-verbal behaviour
Bayesian Network · Hybrid approach · Evaluation

6.1 Motivation

Human behaviour is influenced by several personal and social factors such as culture and can be referred to as a mental program that drives peoples' behaviour [20]. Non-verbal behaviour as a part of human communication is, amongst others, also influenced by cultural background [45]. Cultural differences on the one hand manifest themselves on an outward level and are, on the other hand, also judged differently by observers of different cultural backgrounds [46].

Examples include the presentation and evaluation of facial expressions, gestures or body postures. In South Korea and Japan, for example, restrained facial expressions and postures indicate an influential person, while in the United States, relaxed expressions and postures give the impression of credibility [45].

As anthropomorphic user interfaces, such as virtual characters or humanoid robots, aim at realistically simulating human behaviour it seems likely that their behaviour conveys different impressions on observers from different cultural backgrounds as well. Taking potential cultural differences into account while designing an agent's non-verbal behaviour can thus help improve their acceptance by users of the targeted cultures.

Building models that determine culture-related differences in behaviour is challenging as the dependencies between culture and corresponding behaviour need to be simulated in a convincing and consistent manner. In the last decade numerous attempts have been made to face these challenges, mainly relying on either theoretical knowledge (theory-based) or empirical data (data-driven). While theory-based approaches model culture-specific behaviours based on findings from the social sciences, data-driven approaches aim to extract culture-specific behaviour patterns from human data to inform computational models. Data-driven approaches bear the advantage that they are based on empirically grounded models. However, a large amount of data is required to derive regularities from concrete instantiations of human behaviour. Theories from social sciences include information that may help us encode culture-specific behaviour profiles. But there are usually missing details that are required for a convincing realization of culture-specific virtual characters or humanoid robots. The objective of the present chapter is to combine the advantages of theory- and data-driven approaches. To this end, we illustrate and evaluate an integrated approach that coherently connects a theory-based and a data-driven approach.

The chapter is structured as follows: In the next section, we briefly introduce models of culture from the social sciences that inspired our work. Then we introduce related work on the computational integration of culture into intelligent systems. In Sect. 6.4, we outline our approach by referring to our own prior work that was used as a basis for the endeavor. Subsequently, we describe the design and implementation of a computational model that is built upon the combined approach.

Two potential ways of evaluating the resulting network are then presented in Sects. 6.6 and 6.7. Finally, we conclude our chapter by reflecting on potential contributions and tribulations of our approach.

6.2 Models of Culture

There is a wide variety of theories and models that explain the concept of culture and what drives people to feel that they are belonging to a certain culture. Many definitions of culture provided by the social sciences conceptually describe cultural differences but stay rather abstract. Some theories describe different levels of culture that address, among other things, that culture does not only determine differences on the surface but also works on the cognitive level. Trompenaars and Hampden-Turner [46], for example, distinguish implicit and explicit levels of culture that range from very concrete and observable differences to a subconscious level that is not necessarily visible to an observer.

Other definitions of culture use dimensions or categories to explain differences between certain groups. These approaches describe culture in a way that facilitates building computational models. Therefore, in this section, we selected cultural theories from the large pool provided by the social sciences which explain culture along dimensional models or dichotomies that help understand culture in a more descriptive manner.

A very well known theory that uses dichotomies was introduced by Hall [18], who classifies cultures using different categories such as their members' perception of space, time or context. Regarding haptics, for example, Hall [18] states that people from high-contact cultures tend to have higher tactile needs than members of low-contact cultures who, vice versa, have more visual needs. These needs can also show on the behavioural level. In some Arab countries, as an example for high-contact cultures, it is a common habit between two males to embrace for greeting or to link arms in a friendly way which can be very unusual behaviour in low-contact cultures.

An approach that describes cultures along dimensions was introduced by Kluckhohn and Strodtbeck [29], who formulate different value orientations in order to explain cultures. One dimension, for example, constitutes the relationship to other people, and describes how people prefer relationships and social organizations to be. Although in this theory a classification of values is provided, the impact on behaviour is described rather vaguely and is therefore hard to measure. Building a computational model with it is thus a demanding task and has not been attempted yet.

Another example of defining culture using dimensions is given by Hofstede [20], who categorized different cultures into a five dimensional model. For the model

more than 70 national cultures were categorized in an empirical survey.[1] The *Power Distance* dimension (PDI), describes the extent to which a different distribution of power is accepted by the less powerful members of a culture. The *Individualism* dimension (IDV) describes the degree to which individuals are integrated into a group. The *Masculinity* dimension (MAS) describes the distribution of roles between the genders. The *Uncertainty Avoidance* dimension (UAI) defines the tolerance for uncertainty and ambiguity. The *Long-Term Orientation* dimension (LTO) explains differences by the orientation towards sustainable values for the future. For each of these dimensions, clear mappings are available from national cultures to the cultural dimensions on normalized scales [21]. In [22], Hofstede and Pedersen introduce so-called synthetic cultures that are based on Hofstede's dimensional model. Each synthetic culture observes one of the extreme ends of each dimension in isolation, and conceptually describes stereotypical behaviour of its members.

6.3 Related Work

As pointed out earlier, basically two approaches have been proposed to simulate culture-specific behaviours for synthetic agents: *Theory-based approaches* and *Data-driven approaches*.

Theory-based approaches typically start from a theory of culture to predict how behaviours are expressed in a particular cultural context. A common approach to characterize a culture is to use dichotomies, which are particularly suitable for integration into a computer model [21].

Alternatively, cultures have been characterized by the prioritization of values within a society [42]. Approaches that aim to modulate behaviours based on culture-specific norms and values, typically start from existing agent mind architectures and extend them to allow for the culture-specific modulation of goals, beliefs, and plans.

One of the earliest and most well-known systems that models culture-specific behaviours within an agent mind architecture is the Tactical Language System (http://www.tacticallanguage.com/), which has formed the basis of a variety of products for language and culture training by Alelo Inc. Tactical Language is based on an architecture for social behaviour called Thespian that implements a version of theory of mind [44]. Thespian supports the creation of virtual characters that understand and follow culture-specific social norms when interacting with each other or with human users. While the user converses with the characters of a training scenario, Thespian tracks the affinity between the single characters and the

[1]Originally, Hofstede used a four-dimensional model. The fifth dimension, long term orientation, was added later in order to better model Asian cultures. Meanwhile a sixth dimension, indulgence, was added that described the subjective well-being that members of a culture experience.

human user, which depends on the appropriateness of the user's behaviour. For example, a violation of social norms would result in a decreased affinity value.

To simulate how an agent appraises events and actions and manages its emotions depending on its alleged culture, attempts have been made to enrich models of culture by models of appraisal, see [1] for a survey. An example includes the work by Mascarenhas et al. [31] who aim at the modelling of synthetic cultures that may be obtained by systematically varying particular behaviour determinants. To this end, they extend an agent mind architecture called FAtiMA that implements a cognitive model of appraisal [35] by representations of the Hofstede cultural scales. Based on the extended architecture, agents with distinct cultural background were modelled. In their model, an agent's alleged culture determines its decision processes (i.e., the selection of goals) and its appraisal processes (i.e., how an action is evaluated). For example, an action that is of benefit to others is the more praiseworthy, the more collectivistic the culture is. Using the extended FAtiMA architecture as a basis, the ORIENT [3], MIXER [2] and Traveller [26] applications simulate synthetic cultures with the overall aim to generate a greater amount of cultural awareness on the user's side.

Data-driven approaches rely on annotated multimodal recordings of existing cultures as a basis for computational models of culture. The recordings can be used directly for imitating the human behaviour or statistical patterns can be derived from the data which govern the behaviour planning process.

Such a cross-cultural corpus has, for example, been recorded for multi-party multi-modal dialogues in the Arab, American English and Mexican Spanish cultures [19]. The corpus has been coded with information on proxemics, gaze and turn taking behaviours to enable the extraction of culture-related differences in multi-party conversations. A statistical analysis of the corpus reveals that findings are not always in line with predictions from the literature and demonstrate the need to enhance theory-driven by data-driven approaches.

More recent work by Nouri and Traum [34] makes use of a data-driven approach to map statistical data onto culture-specific computational models for decision making. In particular, they simulate culture-specific decision making behaviour in the Ultimatum Game based on values, such as selfishness, held by Indian and US players collected through a survey. Their work aims to adapt decision making of virtual agents depending on culture-specific values, but does not consider culture-specific verbal and non-verbal behaviours.

While the theory-driven approach ensures a higher level of consistency than the data-driven approach, it is not grounded in empirical data and thus may not faithfully reflect the non-verbal behaviour of existing cultures. Another limitation is that it is difficult to decide which non-verbal behaviours to choose for externalizing the goals and needs generated in the agent minds. The advantage of data-driven computational models of culture lies in their empirical foundation. However, they are hard to adapt to settings different from the ones recorded, as the data cannot be generalized due to a lack of a causal model.

6.4 Approach

To implement a hybrid approach that combines a theory-based with a data-driven approach, we rely on our earlier work that includes implementations of both approaches, but did not yet integrate them in a synergistic manner (see Fig. 6.1).

In scope of the Cube-G project, we recorded a cross-cultural corpus for the German and Japanese cultures [37]. More than 20 participants were video-taped in each culture, each running through three scenarios. So far, only the first scenario, a first time meeting, was considered. For this scenario a student and a professional actor (acting as another student) were told to get acquainted with one another to be able to solve a task together later. Recording one participant and an actor at a time ensured a higher control over the recordings. This way, participants did not know each other in advance and the actor was able to control that the conversation lasted for approximately 5 min. Actors were told to be as passive as possible to allow the participant to lead the conversation and be active in cases where the conversation was going to stagnate.

For the corpora, statistical analyses were performed to identify differences between German and Japanese speakers in the use of gestures and postures, communication management, choice of topics, and the like (e.g. [14, 15, 40]).

For simulation we developed a social simulation environment [10] with virtual characters that portray typical Japanese and German behaviours by gestures and dialogue behaviours. The modelling of the behaviours was based on the observations made in the multimodal corpora.

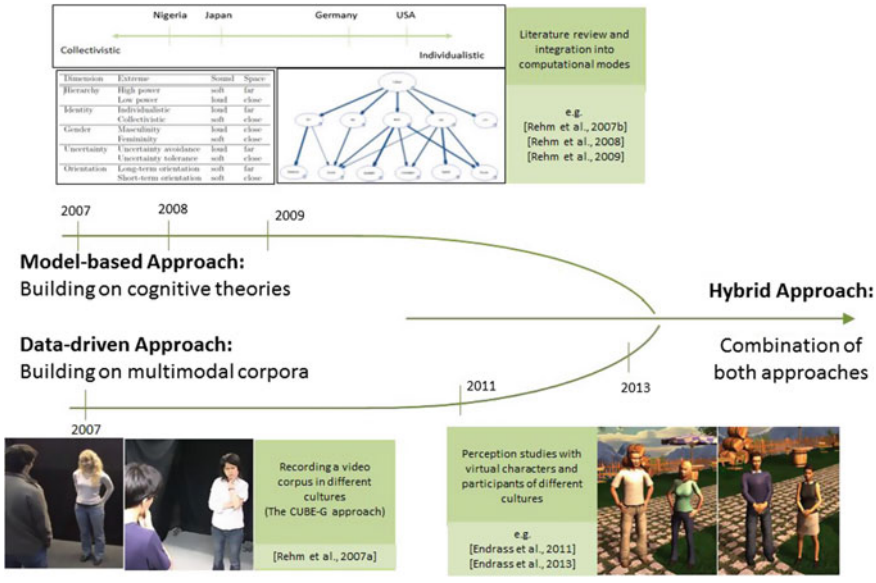

Fig. 6.1 Model-based and data-driven approaches as conducted in our former work

At a later stage, we made use of a data-driven approach and conducted perception studies with virtual characters that simulated the findings of the corpus analysis. Results suggest that users prefer virtual character behaviour that was designed to resemble their own cultural background, e.g. [13, 14]. At this point in time, the characters' behaviour was completely scripted to follow the statistical distribution of the corpus findings, and no computational model was built yet, while each of the studies looked at one behavioural aspect in isolation.

In parallel to the corpus recordings, we started a theory-based approach that developed a parameterized model of cultural variation based on Hofstede's dimensions [39]. To code dominant culture-specific patterns of behaviour, we made use of Bayesian Networks. Culture-specific behaviours were then generated following probability distributions inspired by theories found in the literature. The model was used to adapt the dynamics of gestures, proxemics behaviours as well as the intensity of speech of a group of virtual characters to the assumed cultural background of the user [38].

In this contribution, we combine the two approaches in a synergistic manner. In particular, we will make use of cultural theories to model dependencies between culture-related influential factors and behaviours and employ statistical methods to set the parameters of the resulting model; i.e. probabilities will be learned from recordings of human behaviour. A similar approach was presented by Bergmann and Kopp [4] to model co-verbal iconic gestures, however, without considering culture-specific aspects. We extend their work by adapting parameter settings of a Bayesian Network to a particular culture based on a data-driven approach.

The approach combines advantages of the commonly used theory-based and data-driven approaches, as it explains the causal relations of cultural background and resulting behaviour, and augments them by findings from empirical data. The resulting hybrid model, combines all previously considered behavioural aspects (verbal and non-verbal) in a complete model. Having such a model at hand, we are able to generate culture-specific dialogue behaviour automatically following the statistical distribution of the recorded data.

We are thus, extending our previous approaches in the following ways: (1) by integrating aspects of verbal and non-verbal behaviour into a complete model (2) by augmenting the model with empirical data (3) by validating the resulting model by testing its predictive qualities and performing a perception study with human observers.

6.5 Modeling Culture-Specific Behaviours with Bayesian Networks

We have chosen to model culture-specific behaviours by means of a Bayesian Network. The structure of a Bayesian Network is a directed, acyclic graph (DAG) in which the nodes represent random variables while the links or arrows

connecting nodes describe the relationship between the corresponding variables in terms of conditional probabilities [41]. The use of Bayesian Networks bears a number of advantages. In particular, we are able to make predictions based on conditional probabilities that model how likely a child variable is given the value of the parent variables. For example, we may model how likely it is that a person makes use of very tight gestures if the person belongs to a culture that is characterized as highly collectivistic. By using a probabilistic approach for behaviour generation, we mitigate the risk of overstereotyping cultures. For example, a character that is supposed to portray a particular culture would show culture-specific behaviour patterns without continuously repeating one and the same prototypical behavioural sequence. Furthermore, Bayesian Networks enable us to model the relationship between culture-related influencing factors and behaviour patterns in a rather intuitive manner. For example, it is rather straightforward how to model within a Bayesian Network that a member of a collectivistic culture tends to use less powerful gestures. Finally, Bayesian Networks support the realization of a hybrid approach. While the structure of the Bayesian Network—in particular the dependencies between cultural influencing factors and behaviour patterns—can be determined by relying on theories of culture, the exact probabilities of the Bayesian Network can be learnt from recordings of culture-specific human behaviours.

The aim of the Bayesian Network model described in this section, is to automatically generate culture-dependent non-verbal behaviour for a given agent dialogue in the domain of first time meetings. Therefore, the network is divided in two parts:

- influencing factors: those factors that are given for the specific conversational situation such as the cultural background of the interlocutors, and
- resulting behaviours: the specific settings of non-verbal behavioural aspects that are calculated by the network as a result of the given influencing factors.

The Bayesian Network was modeled using the GeNIe modeling environment [12]. Figure 6.2 shows the structure of the network with its influencing factors and resulting behaviours that will be further explained in the following subsections.

6.5.1 Influencing Factors

To be able to construct different agent dialogues that vary with cultural background, influencing factors in our model are further divided into cultural background and conversational verbal behaviour.

In our network model, we rely on Hofstede's dimensional model [21], that captures national cultures as a set of scores along dimensions, providing a quite complete and validated model, especially considering the fuzzy concept of culture (see Sect. 6.2). The model has widely been used as a basis to integrate culture for

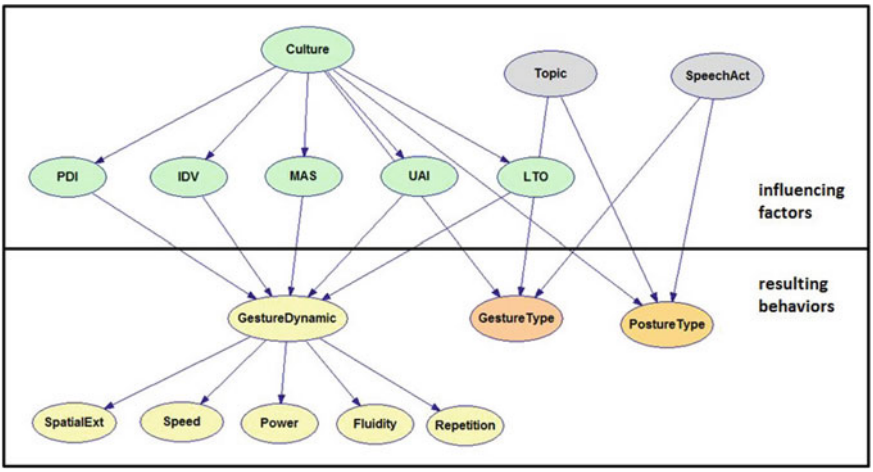

Fig. 6.2 Network model including influencing factors and resulting behaviours for culture-related behaviour generation

anthropomorphic interfaces, e.g. [2, 3, 32]. For our implementation, we categorized the scores on the cultural dimensions, provided by [21], into three discrete values (low, medium, high).

Regarding the content of dialogues, cultural differences can play a crucial role as well. According to Isbister and colleagues [24], for example, the categorization into safe and unsafe topics varies with cultural background. Therefore we expect variations in the semantics of first-time meeting conversations dependent on cultural background. Schneider [43] classifies topics that prototypically occur in first-time meetings as follows: The *immediate situation* holds topics that are elements of the so-called frame of the situation, such as the surrounding or the atmosphere of the conversation. The *external situation* describes topics that hold the larger context of the conversation, such as the news, politics, or recent movies. In the *communication situation* topics are focusing on the conversation partners, e.g., their hobbies, family or career. Potential topics for dialogues in our network model are based on this categorization.

Besides the semantics of speech, the function of each utterance is considered in our network. For computational models, verbal behaviour is often described by speech acts. Core and Allen [9], for example, provide a coding schema that categorizes speech acts along several layers. One layer of their schema, labels the communicative meaning of a speech act. As the whole schema is too complex for our purpose, we use the following subset of communicative functions that meet the requirements of first time meeting scenarios: *statement*, *answer*, *info request*,

agreement/disagreement (indicating the speaker's point of view), *understanding/misunderstanding* (without stating a point of view), *hold*, *laugh* and *other*.

6.5.2 Resulting Behaviours

The lower part of our network model consists of non-verbal behavioural traits, that are calculated by the model based on the influencing factors (see Fig. 6.2). There is a large number of non-verbal behaviours that could be dependent on cultural background. At this stage, we focus on the upper part of the torso and consider gesture types, their dynamic variation and arm postures, and do not look at other parts of the body yet, e.g. head movements. The generation of adequate postures, gestures, and their dynamics have been widely studied in the field of virtual agents (e.g. [5, 8, 36]) and seem to be good aspects to improve the characters' believability. Regarding cultural differences, for body postures, for example, it has been shown that different cultures perceive different emotions depending on body postures [28].

In [6], Bull provides a coding schema to distinguish different body postures, including prototypical positions of, for example, head, feet or arm. To distinguish arm postures in our model, we employ Bull's categorization for arm postures. In total, 32 different arm positions are described in the schema and included to the network. Table 6.1 shows an extraction of the schema containing postures that were frequently observed during our corpus study and are thus most relevant for our approach.

Gestures can be considered on two levels: which gesture is performed and how it is performed. The most well known categorization of gesture types has been provided by McNeill [33]. Although the categorization is not meant to be mutually exclusive, the types described are a helpful tool to distinguish gestures: *Deictic* gestures are pointing gestures; *Beat* gestures are rhythmic gestures that follow the prosody of speech; *Emblems* have a conventionalized meaning and do not need to be accompanied by speech; *Iconic* gestures explain the semantic content of speech, while *metaphoric* gestures accompany the semantic content of speech in an abstract manner by the use of metaphors; *Adaptors* are hand movements towards other parts

Table 6.1 Extraction of arm postures considered in our model

Arm posture	Description
PHIPt	Put hands to pocket
PHFE	Put hand to face
PHEW	Put hand to elbow
PHWr	Put hand to wrist
FAs	Fold arms
JHs	Join hands
PHB	Put hands back

of the body to satisfy bodily needs, such as scratching one's nose. In our network model, we classify gesture types as a subset of McNeill's categorization [33]: deictic, beat, iconic, and metaphoric. Adaptors were excluded from the network as we are focusing on gestures that accompany a dialogue. Emblems were excluded as they are not generalizable and might convey different meanings in different locations. An example includes the American OK-gesture (bringing the thumb and the index finger together to form a circle). While it means OK in the Northern American culture, it is considered an insult in Latin America, and can be interpreted as meaning homosexual in Turkey. Thus, even if our model could predict that a emblematic gesture type would be appropriate in a given situation, different concrete gestures needed to be selected based on cultural background.

Besides the choice of a gesture, its dynamics can differ and be dependent on cultural background. According to Isbister [23] "*what might seem like violent gesticulating to someone from Japan would seem quite normal and usual to someone from a Latin culture*". To include this phenomenon in our network, we added a node containing a gesture's dynamics which is divided into three discrete values (low, medium, high). The dynamic variation of a gesture can be further broken down into dimensions [17], each describing a different attribute of the movement. Following [30], who investigated gestural expressivity for virtual characters, we employ the parameters spatial extent (the arm's extent relative to the torso), power (acceleration), speed, fluidity (flow of movements) and repetition. Initial values in our network were set in a manner that a high dynamics is more likely to result in a higher value for each of the parameters.

6.5.3 Dependencies

So far we introduced the nodes and their parameters of the influencing factors and resulting behaviours of the network and stated why they were included. In this subsection we explain their dependencies that were modeled for the network.

Although there is a strong evidence that the types of gestures and arm postures are dependent on cultural background, e.g. [45], there are no clear statements in the literature on how exactly McNeill's gesture types or Bull's arm posture types would correlate with Hofstede's dimensions of culture. We thus connected the nodes holding gesture types and arm postures directly to the culture node instead of linking them via Hofstede's dimensions.

Regarding the dynamics of a gesture and its correlation to Hofstede's cultural dimensions, we rely on the concept of synthetic cultures that builds upon Hofstede's dimensions (see Sect. 6.2). For these abstract cultures prototypical behavioural traits are described. For example, the *extreme masculine culture* is described as being loud and verbal, liking physical contact, direct eye contact, and animated gestures. Members of a *extreme feminine culture*, on the other hand, are described as not raising their voices, liking agreement, not taking much room and

being warm and friendly. Furthermore, the position on Hofstede's dimensions determines the stereotypical movements of a synthetic culture. Thus, in our model, culture is connected to the gesture dynamic node via the cultural dimensions.

Please note that although it is known from the literature that aspects of verbal behaviour, e.g. the choice of conversational topic, can be dependent on cultural background, in our network model they are considered as influencing factors. This design choice was made, as we aim at generating culture-dependent non-verbal behaviour for a given dialogue and do not want to generate the dialogue itself. Therefore, nodes containing information about the selected verbal behaviour (speech act and conversational topic) are not connected to the culture node but linked directly to the nodes describing the resulting non-verbal behaviour types. The dialogue that is used as an input, can of course contain culture-specific content and thus influence the selected non-verbal behaviour.

6.5.4 Parameters of the Model

Using an automated learning process the network's model described in the previous subsection was augmented with empirical data. For that purpose the findings of the corpus (cf. Sect. 6.4) had to be processed before applying an automated learning process. The Anvil tool [25] allows to align self-defined attributes and parameters to moments in videos and was therefore used to annotate the videos for our former statistical analysis. Different behavioural attributes, as described in the previous subsection, were annotated for the videos. Conversational topics were annotated for the verbal behaviour, as well as speech acts. Further, non-verbal behaviour was annotated, i.e. arm postures, gesture types and dynamics of gestures. Afterwards the cultural background was added to the meta data of the annotations.

For further processing, the different modalities had to be aligned. Therefore, the annotated conversations were divided into conversational blocks (*dataset* in the following). These datasets are defined to refer to a specific speech utterance, specified by speech act and conversational topic. Following the categorization into speech acts, datasets are thus determined by a clause or sub-clause.

As non-verbal behaviour was defined to be determined by the verbal behaviour it accompanies, gestures and arm postures were added to datasets based on timely overlap. Thus, if a arm posture and/or gesture was annotated for the same time as a speech act, it was added to the corresponding dataset. If there was no non-verbal behaviour in the same time frame, an empty token was added to the dataset. In order to reflect that gestures and postures can be maintained for a longer time span, a gesture (or posture) was added to all datasets it overlapped with. Inversely, when multiple gestures or postures overlapped with a dataset's time period, the token with the longest overlap was chosen.

Due to data loss during the years of our endeavor that was caused by multiple annotation files and separate statistics, it was not possible to temporally align the gestures' dynamics with the corresponding speech acts any more, but to use them only quantitatively. Therefore, two different datasets were used for learning the

parameters of the network. The aligned dataset was used to learn the joint probability distributions of arm postures and gesture types subject to verbal behaviour and culture. The probabilities of the gestures' dynamics were learned from the unaligned dataset and thus based on culture only.

After the extraction, we had two datasets: the aligned dataset containing 2155 values and the non-aligned dataset containing 457 values. Parameters were learned with the EM-algorithm [11] that is provided by the implementation of the SMILE-Framework [12] underlying the modeling environment used to model our network. This algorithm is able to deal well with missing data and is thus suitable for our purpose as there were some aspects not annotated for every person recorded. In the Japanese part of the corpus, e.g., there were some annotations of speech act and topic missing due to absent translation. The learning process itself was performed in two steps. First the probabilities of the parameters for arm posture and gesture type were learned from the aligned dataset. Afterwards the parameters for gesture dynamics were determined by applying the EM-algorithm with the non-aligned dataset.

6.5.5 Resulting Network

Figure 6.3 exemplifies the calculations of the resulting network with the evidence of cultural background either being set to Japanese (upper) or German (lower). In case no other evidence than cultural background is set within the network, distributions reflect the findings of our former statistical analysis, where we have been looking at behavioural aspects depending on cultural background in isolation (see Sect. 6.4). With this setting, cultural variations in non-verbal behaviour can be reflected in a general manner based on culture only.

Having a model at hand that combines several behavioural aspects instead of looking at them in isolation, further observations can be done in an intuitive manner. By setting additional evidences in the network, e.g. for verbal behaviour, the trained Bayesian Network allows to explore further interdependencies in the data. For example, a correlation of chosen topic and non-verbal behaviour frequency stayed unnoticed in our earlier work. From previous analysis of verbal behaviour [14], we know that the topic distribution is different for the two cultures in the data. While in Japan significantly more topics covering the immediate situation occurred compared to Germany, in Germany significantly more topics covering the communication situation occurred compared to Japan. Setting evidences of the topic nodes and the cultural background, the network reveals that people in both cultures are more likely to perform gestures when talking about less common topics. In particular, the communication situation in the Japanese culture and the immediate situation in the German culture. This effect could be explained by the tendency that talking about a more uncommon topic might lead to a feeling of insecurity that results in an increased usage of gestures. Thus, the network also reveals how culture-related non-verbal is mediated by culture-specific variations in verbal behaviour.

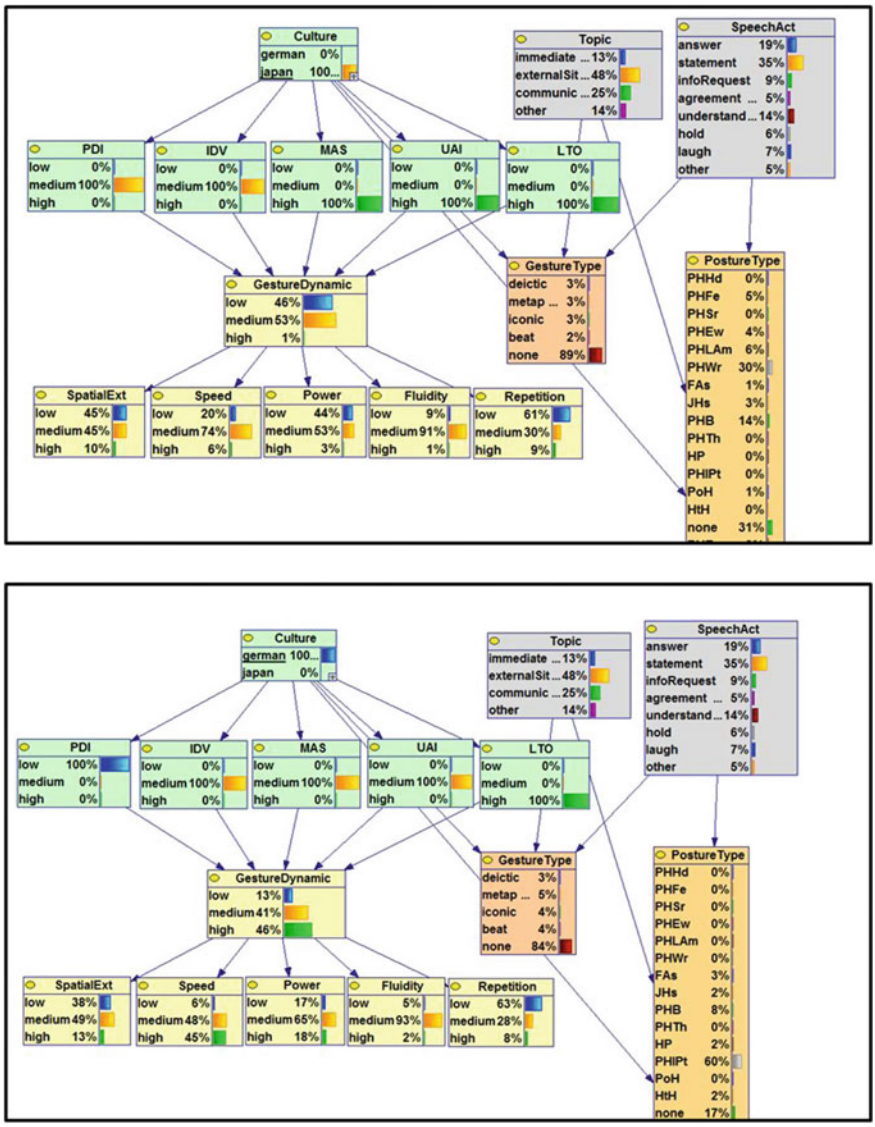

Fig. 6.3 Resulting Bayesian Network with parameters learned from empirical data, with cultural background set to Japanese (*upper*) and German (*lower*) respectively

6.6 Validation of the Network

An important question concerns the validation of the network. Basically, there are two possibilities. In this section, we will evaluate to what extent the network is able to predict characteristic behaviours of a person portraying a particular culture; i.e.

the generated behaviours are compared against the collected corpus. In Sect. 6.7, we will present a perception study in order to find out how human observers respond to the culture-specific behaviours of virtual characters generated with the network. That is the effect of the generated behaviours is evaluated from a user's perspective. We decided to rely on videos for the perception study to be able to evaluate our hypotheses in a controlled setting. As an alternative, we might have users interact with characters showing culture-specific behaviors. For example, in [27] we represented users by avatars which imitate their body movements and postures. System controlled characters respond to the users by dynamically adapting their behavior depending on their own assumed cultural background.

6.6.1 Measuring the Predictive Qualities of the Network

To validate the model, we investigate whether the network is able to predict appropriate culture-specific behaviours for new situations that are not included in the training corpus. To this end, a tenfold-cross-validation was performed. A dataset of 2155 values was used in which non-verbal behaviour (gesture types and arm postures) was aligned with the speech of the participants. The model was trained using 90% of the data while the remaining 10% were used as validation data. This validation process was performed ten times each time leaving out another part of the original dataset. For each dataset the cultural background (German or Japanese) and performed verbal behaviour (speech act type and topic) is given while gesture and arm posture are predicted by the network. Please note that the non-verbal dynamics cannot be validated using this approach due to the missing alignment. Predictions of the network were compared to the behaviour observed in the corpus data. Generally it appears quite unlikely for humans to behave the exact same way in a given situation several times. As a similar variety of behaviour is desirable for virtual characters we consider the network as suitable in case the observed gesture or posture is finding itself in the best three guesses of the network.

Figure 6.4 shows the prediction rates for gesture and arm posture types. Although results look quite promising, with an overall accuracy of 88% for gesture types and 56% for posture types, these results should not be overrated as for many of the observed speech acts no non-verbal behaviour was conducted (resulting in the gesture and posture type *none*). In particular, a gesture was performed only in 11% of our dataset, while a posture was performed for 71% of the values in the dataset.

As a result, another 10-fold-cross-validation was carried out on an adjusted dataset excluding data where no gesture or posture was observed. With it, prediction rates of the network are calculated assuming a gesture or posture should be performed by an agent. In total, 233 speech acts were accompanied by a gesture, 1551 by a posture. Figure 6.5 shows the results. Although a weak trend can be observed, only 34% of the performed gestures were predicted correctly by the

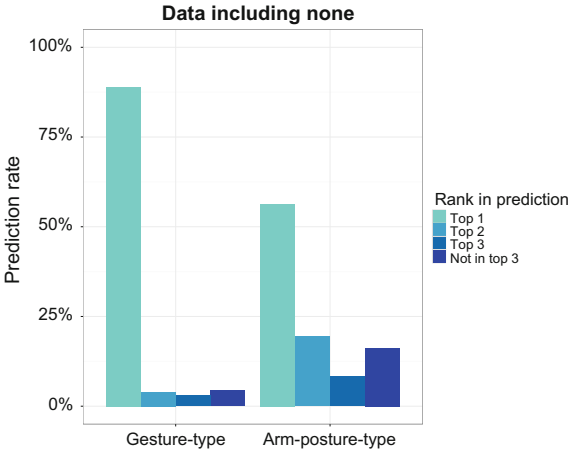

Fig. 6.4 Prediction rates of observed gesture types and posture types being in the first, second or third most likely gesture and posture type

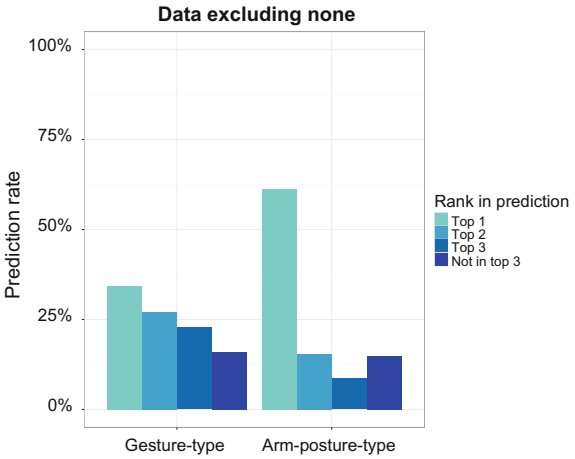

Fig. 6.5 Prediction rates of observed gesture type and posture type being in the first, second or third most likely gesture and posture type, excluding none-elements

network which seems not better than random. For posture types, however, the overall accuracy looks much more promising with 61% correct predictions.

In a further validation step cultural background was reversed to find out whether predicted gestures and arm postures reflect a prototypical cultural background. Thus, a 10-fold-cross-validation was performed with the cultural background set to German in the network while using the Japanese validation data, and vice versa. Including *none*-elements the accuracy of gesture types was still 80% as no gesture was the most likely option for both cultures. Therefore, *none*-elements were excluded again resulting in a drop of accuracy for gesture types to 24% (see Fig. 6.6), resulting in worse predictions of the network compared to the original data set (cf. Fig. 6.5).

For posture types, with reversed cultural background including *none*-elements accuracy drops to 5% while 75% of the observed postures did not even fall in the

top 3 predictions. Excluding none-elements, accuracy is less than 2% with reversed cultural background with 92% of the observed posture types not being in the top three guesses (see Fig. 6.6). Thus, for arm postures changing the cultural background leads to a very low predictive power of the network suggesting that postures can be predicted culture-dependently by the network.

For the parameters of the gesture dynamics a 10-fold-cross-validation was performed based on cultural background alone. Thus, probabilities for the levels of dynamics are calculated given that a gesture is performed. More dataset values could be used in this case, as missing translations of verbal behaviour could be ignored, resulting in a dataset containing 457 gestures. Results are shown in Fig. 6.7. Please note that in this case the levels did comprise only three categories. Still, results look promising, suggesting that trends can be predicted for culture-dependent gesture dynamics.

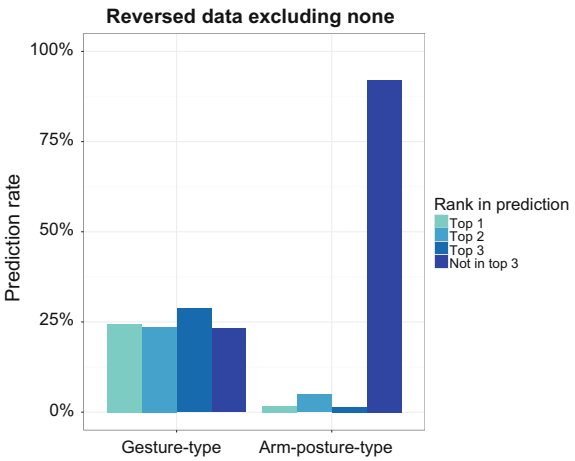

Fig. 6.6 Prediction rates of observed gesture types and posture types being in the first, second or third most likely gesture or posture type with the cultural background set to the reversed culture, excluding none-elements

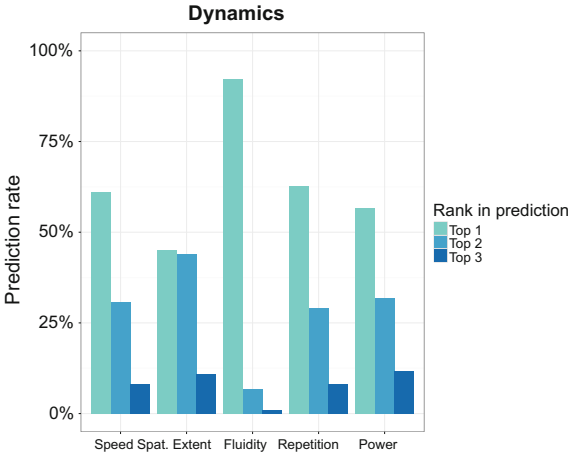

Fig. 6.7 Prediction rates of observed gestural dynamics being in the first, second or third most likely category

6.6.2 Discussion

With the cross-fold validation, we evaluated to what extent the network is able to predict culture-related behaviours of the underlying corpus. Regarding posture types and gestural expressivity, the presented network performed well. Our previous statistical analysis (e.g. [40]), also showed a strong correlation of cultural background and body posture with statistically significant differences between the cultures. In general, postures that regularly occurred in one culture barely occurred in the other culture. Similarly, the data revealed strong differences between the cultures regarding gestural expressivity suggesting that German participants gestured more expressively than Japanese participants.

Regarding gesture types, no reliable predictions could be made by our network. This is also reflected in the underlying data. The former statistical analysis showed that the overall number of gestures is similar in both cultures and no significant differences were found regarding the frequencies of McNeill's gesture types. As a result, the network cannot add to believably simulating culture-specific behaviours focusing on gesture types. This might have been caused by the abstraction of gestures to categories. Even if, for example, a deictic gestures is performed by people of different cultural backgrounds, the concrete execution can be very different. While a deictic gesture, for example, is typically performed using the index finger in Western cultures, this is considered rude in some Asian cultures, where deictic gestures are usually performed using the whole hand.

In sum, the resulting model can only be only as meaningful as the data being used to learn the probabilities. We thus believe that learning a Bayesian Network to enculturate non-verbal behaviours for simulated dialogues is a good approach in case the underlying data contains strong cultural differences. In our case, we believe the network can be used to add posture-types and levels of expressive behaviour to simulated dialogues in order to increase the culture-relatedness of the simulated non-verbal behaviours. For gesture types further research is needed such as going into more depth regarding the concrete performance of specific gestures or their correlation to the semantics of speech rather than speech acts.

6.7 Perception Study

A second approach of validating our network, is to investigate the resulting behaviours with virtual characters, by asking people of the targeted cultural groups to rate their perceptions of the characters' behaviours.

According to the similarity principle [7] interaction partners who perceive themselves as being similar are more likely to like each other. We therefore expect that participants of our study prefer agent conversations that resemble their own cultural background.

In our former work we performed perception studies with virtual characters that followed scripted behaviour based on the statistical analysis of our video corpus. Each behavioural aspect was tested in isolation to find out which of the implemented aspects of behaviour cause the desired effects. Results suggested that observers tended to prefer virtual agent behaviour that is in line with their own cultural background for some of the behavioural aspects (such as postures or prototypical topics), while we did not find significant differences for other aspects (such as gesture types) [13].

In comparison to the scripted perception studies, the present study uses the Bayesian Network described in Sect. 6.5.5 which is able to present culture as a non-deterministic concept and preserve a certain variety in the characters' behaviours. In addition, all behavioural aspects implemented in the network are generated in combination based on cultural background and the underlying dialogue.

6.7.1 Design

The perception study was conducted in a mixed-design with the participant's culture as independent between-subjects variable with two levels (German and Japanese), the agent-culture as within-subjects measure (German and Japanese) and the participants' subjective impression of the characters behaviour and the conversation in general as dependent measures. The participants consecutively watched four videos of conversations of virtual characters. After watching a video they rated their agreement to several statements regarding their impression of the conversation they saw in the video. Hence, the subjective impression per culture was designed as a within-subjects factor with two levels, each calculated as the average of two videos for every variable. The videos were presented to the participants based on a 4×4 Latin square to counterbalance order effects.

6.7.2 Apparatus

For the realization of the perception study we need a demonstrator containing a virtual environment with virtual characters, a conversational setting with verbal behaviour, as well as a simulation of the generated non-verbal behaviours.

6.7.2.1 Virtual Environment

We use our virtual character engine [10] that contains a virtual Beergarden scenario, in which virtual characters can be placed. Regarding the verbal behaviour of the characters, a text-to-speech engine with different voices and languages such as German, English or Japanese can be used.

To simulate a first time conversation similar to the ones recorded in our corpus, two characters were placed into the scenario facing each other. To be as culturally neutral as possible, characters were chosen that do not contain typical ethnically appearances such as blond hair. To avoid side effects evoked by gender, we chose a mixed gender combination for the agent conversations. Thus a female and a male character interact with each other. Figure 6.8 shows a screenshot of the setup.

To use the virtual agent engine with our network model, a dialogue component was added that allows to script dialogues in an XML structure where speech acts and topic categories can be tagged for verbal behaviour, and a cultural background can be set for each character. For dialogues that have been prepared in that manner, the Bayesian Network is able to generate non-verbal behaviours for a given cultural background.

Regarding non-verbal behaviour, over 40 different animations can be performed by the characters, including gestures and body postures. Body postures were modeled to match the arm-posture types included in our network. Figure 6.9 illustrates two typical arm postures that were observed regularly in our video corpus.

To select gestures, existing animations had to be labeled according to the gesture types used by our network. In addition, gestures can be customized by the animation engine to match different levels of expressivity [10]. The speed parameter is adapted by using a different frame rate. Animation blending is used for differences in the spatial extent (blending with a neutral body posture) and fluidity (blending over a shorter or longer period of time). Differences in the repetivity are archived by playing the stroke of a gesture several times.

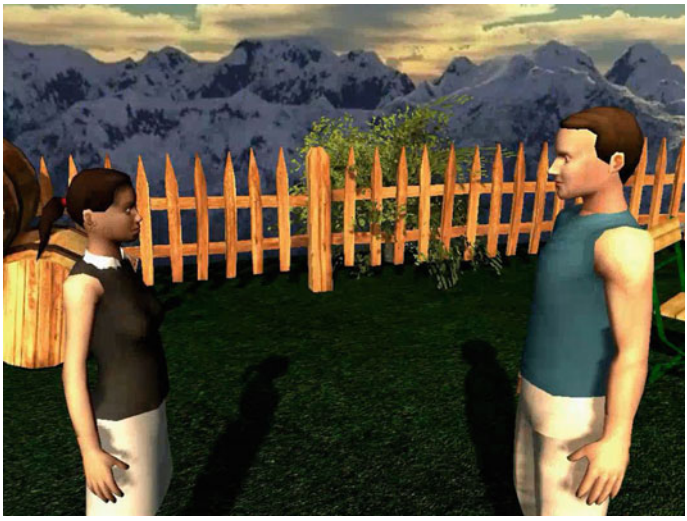

Fig. 6.8 Two virtual characters facing each other during a conversational setting

Fig. 6.9 Prototypical arm postures displayed by our male virtual character (*left*: The prototypical Japanese posture "*Put hands to wrist*"; *right*: the prototypical German posture "*Put hands into pocket*")

6.7.2.2 First-Time Meeting Dialogue

As the focus of the present chapter lies on non-verbal behaviour that accompanies first-time meeting dialogues, a dialogue needs to be written that contains the casual small talk of such a situation, whilst not holding culture-specific content. Creating such a dialogue is not trivial, since, as we pointed out earlier, dialogue behaviour can heavily depend on cultural background. In one of our previous studies [14], we addressed that issue, and analysed the first-time meetings of our video corpus regarding differences in topic selection. Following Schneider [43], topics have been classified into (1) Topics covering the *immediate situation* which describe elements of the so-called "frame" of the situation. The frame of a small talk conversation at a party, for example, holds topics such as the drinks, music, location or guests. (2) The second category, the *external situation* or "supersituation" includes the larger context of the immediate situation such as the latest news, politics, sports, movies or celebrities. (3) The *communication situation* contains topics that concentrate on the conversation partners. Thus, personal things such as hobbies, family or career are part of this category. The corpus analysis revealed that topics covering the *external situation* were the most common topics in both cultures. In the German conversations the *immediate situation* occurred significantly less compared to the *external situation*, while in the Japanese conversations the *communication situation* occurred significantly less compared to the *external situation*. In a perception study with virtual characters we found out that conversations with a typical German topic distribution were preferred by German participants, while conversations with a typical Japanese topic distribution were preferred by Japanese participants [14]. It is therefore crucial to avoid topic categories in the present study that are not common

in one of the targeted cultures. Another issue might be that a dialogue that is written by us might be influenced by our own cultural background and might thus not be general enough to be considered a normal casual small talk conversation in the other culture. To tackle this issue, in [14], we agreed on six English dialogues with our Japanese cooperation partners, that would in general be feasible in both cultures. Please note that for our former aim three of the dialogues contained a prototypical German topic distribution and three of the dialogues contained a prototypical Japanese topic distribution. Therefore, for the present study, we needed to cut down the dialogues in a sense that we only keep parts in which the *external situation* is discussed, while the *immediate situation* and the *communication situation* were avoided. This results in a dialogue that lasts for approximately 60 s, that could occur in both cultures, and that only contains topics that are common in both cultures. As pointed out in Sect. 6.7.2.1, dialogues need to be tagged with their speech acts, to be used by our network. Please see Table 6.2 (left part) for the resulting dialogue and the annotated speech acts. The table additionally contains an example of non-verbal behaviour that was generated by our network for the German and Japanese culture respectively. Please note, that a posture was maintained by the characters until a different animation was selected.

6.7.2.3 Video Generation

The probabilities for non-verbal behaviours are generated by the network depending on the current speech act, topic, and cultural background of the agents. For display with the virtual characters, our demonstrator allows two options. Either the most likely non-verbal behaviour is displayed, or the probability distribution of the network is reflected by displaying a non-verbal behaviour that is chosen based on an algorithm that follows the probabilities. While always choosing the most probable behaviour is very well suited for illustration, it lacks a certain variety in the characters' behaviours. Thus, to present culture as a non-deterministic concept and to reflect the generative power of our network, we use the second option for the videos that shall be shown in our perception study.

This approach can be quite risky, as it might result in a conversation that contains behaviours that are very unlikely for a given cultural background, or produce conversions that do not contain a reasonable amount of animations at all. While these effects would cancel out over a long time period of agent conversations, for an evaluation study with a dialogue of 60 s only the generated behaviour might not be representative enough. Therefore, we were running the network ten times for each culture and recorded videos of the resulting conversations. To prepare our perception study, we manually selected two videos for each culture, that contained a comparable amount of non-verbal behaviours and no animations that were very unlikely for the given cultural background. By selecting two videos per culture, we wanted to assure that we did not accidentally choose a video that would in general be rated better or worse due to the specific animation selection.

Table 6.2 Dialogue and generated prototypical non-verbal behaviour

Interlocutor	Utterance	Speech act	Jap. non-verb.	Ger. non-verb.
Agent A	Do you know Mary for a long time now?	InfoRequest	PHB	PHIPt
Agent B	No.	Answer	None	PHIPt
	Not too long.	Hold	None	PHIPt
	I met her last year at university.	Statement	PHWr	PHIPt
Agent A	Mary looks busy for her part-time job.	Statement	PHB	PHIPt
Agent B	Yes.	Agreement	PHWr	PHIPt
	I heard that she goes to the part-time job 3 times a week.	Statement	Beat	PHIPt
	But one of our friends is missing.	Statement	PHWr	Beat
	She is on a trip to Brazil.	Statement	PHWr	PHIPt
Agent A	Brazil?	Hold	PHB	PHIPt
	It is supposed to be very beautiful there.	Statement	PHWr	Metaphoric
Agent B	Yes.	Understanding	PHB	PHIPt
	And people there are very friendly.	Statement	PHB	PHIPt
Agent A	This is especially good for hiking as far as I know.	Statement	PHWr	PHIPt
Agent B	Right.	Agreement	PHB	PHIPt
	There are many good hiking trips.	Hold	PHB	PHIPt
	There are also the Olympic Games in Brazil this summer, aren't they?	InfoRequest	Deictic	PHIPt
Agent A	I think you are right.	Agreement	PHWr	PHIPt
	They are taking place in Brazil this year.	Statement	PHWr	Deictic

Please note, that the original dialogue had slightly to be modified, as the current location of the proceeding Olympic games had changed over the years

Please see Table 6.2 (right part) for the non-verbal behaviours that were generated by our network for two of the chosen videos of our perception study.

6.7.2.4 Questionnaires

For evaluation, a two-parted questionnaire was developed: Part A included questions focused on the observed conversation, part B requested demographical data.

In part A of the questionnaire, participants were asked to rate the characters in the video as well as their perception of the conversation on 7-point-Likert scales,

Table 6.3 Participants had to state their agreement regarding their perception of the characters and the conversation in general (Part A of the questionnaire)

Statements regarding the perception of the characters
The characters' behaviour was natural
The characters' behaviour was appropriate
The characters were getting along with each other well
Statements regarding the perception of the conversation
The characters' movements matched the conversation *I liked watching the conversation*
I would like to join the conversation
The conversation or a similar conversation would be realistic in my own life /my friends life

ranging from "strongly disagree" to "strongly agree" (see Table 6.3). At the end, participants were provided a comment box for further opinions.

In part B of the questionnaire, participants were asked to provide demographical data on:

- age
- gender
- the country they currently live in
- the country they have lived in mostly in the last 5 years
- their ability to understand spoken English well and
- their ability to understand the spoken English of the videos well (the latter two were rated on a 7-point-Likert skale)

6.7.3 Procedure

The study was embedded in an online survey.[2] At first it was explained that participation was totally voluntary, that they could withdraw at any time, and that they should turn on their speakers to be able to listen to the proceeding conversations. Then they were introduced to the scenario they were going to see. The imaginary scenario involved two virtual characters, that have just met each other for the first time in a social setting. They were introduced to each other by a common friend that has just left to pick up something to drink for all of them. To get participants acquainted with the virtual environment, the visual appearance of the characters, as well as their ability to conduct non-verbal behaviours they were shown a "neutral" video first. This video showed the two characters greeting each other by introducing themselves stating their names. In order to assure participants had their speakers

[2]The study was created with SoSci Survey (Leiner 2014) and made available to the participants on www.soscisurvey.com.

enabled and understood the language, they had to enter the characters' names as free-response items to be able to further proceed in the study.

After the introduction, participants were told that they were going to watch four videos of the same conversation. Videos could be watched several times if the participants wanted to do so. After each video, participants had to answer part A of our questionnaire.

After the last conversation, participants were asked to provide the demographic data requested in part B of our questionnaire. Finally, participants were thanked for their participation. Overall, completion of participation took participants 10–15 min.

6.7.4 Participants

There were 56 respondents taking part in the online-survey. After excluding four respondents that did not finish the questionnaire, there was a final sample of 52 participants (29 German and 23 Japanese). The culture of participants was assessed by asking where they lived mostly the last 5 years and where they currently live. Two participants did currently live in another country than the country they lived in mostly in the last five years. For these participants latter values were used as they fit the scope of the study better. 27 of the participants were male (11 Japanese-male, 16 German-male) and 25 female (12 Japanese-female, 13 German-female), with an age range between 19 and 41 ($M = 24.98$, $SD = 5.29$). All participants were volunteers, recruited via social networks or e-mail lists. On average the self-reported English skills for Japanese participants ($M = 4.22$, $SD = 1.20$) was significantly lower than for German participants ($M = 6.31$, $SD = 0.76$), $t(36) = 7.44$, $p < 0.001$, $r = 0.78$. As Levene's test indicated unequal variances ($F = 4.613$, $p = 0.037$) degrees of freedom were adjusted from 50 to 36. Regarding their understanding of the spoken English of the videos Japanese participants' ratings ($M = 4.87$, $SD = 1.25$) were also significantly lower than the ratings of the German participants ($M = 6.76$, $SD = 0.51$), $t(28) = 6.790$, $p < 0.001$, $r = 0.79$. Degrees of freedom were adjusted from 50 to 28 as Levene's test showed unequal variances ($F = 20.023$, $p < 0.001$). Since all self-reported English skills were high enough, no participant had to be excluded due to language skills.

6.7.5 Results

In order to draw conclusions based on the generated behaviour and not on a specific video, the mean of both videos was calculated for each agent-culture and every item of the subjective impression. For example, the subjective naturalness of the prototypical German behaviour was calculated as the mean of the subjective naturalness of both videos showing prototypical German behaviour. Figures 6.10 and 6.11

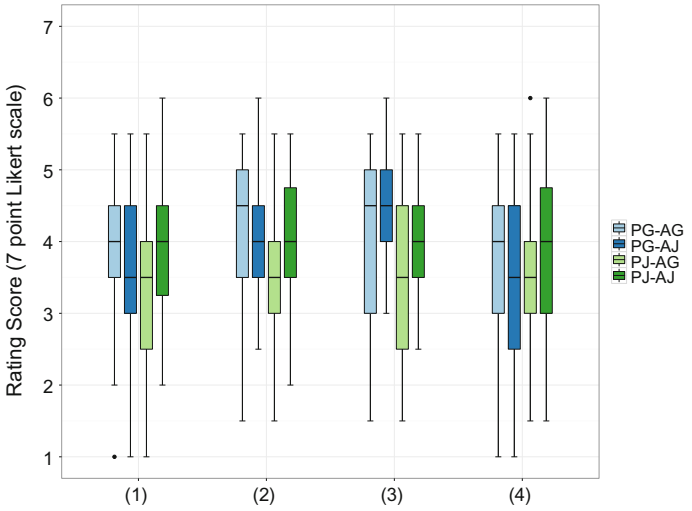

Fig. 6.10 Results of the subjective impression of the German agents (AG) and Japanese agents (AJ) for German participants (PG) and Japanese participants (PJ) with regard to the statements from Table 6.3: (*1*) "*The characters' behaviour was natural.*", (*2*) "*The characters' behaviour was appropriate.*", (*3*) "*The characters were getting along with each other well.*", (*4*) "*The characters' movements matched the conversation.*"

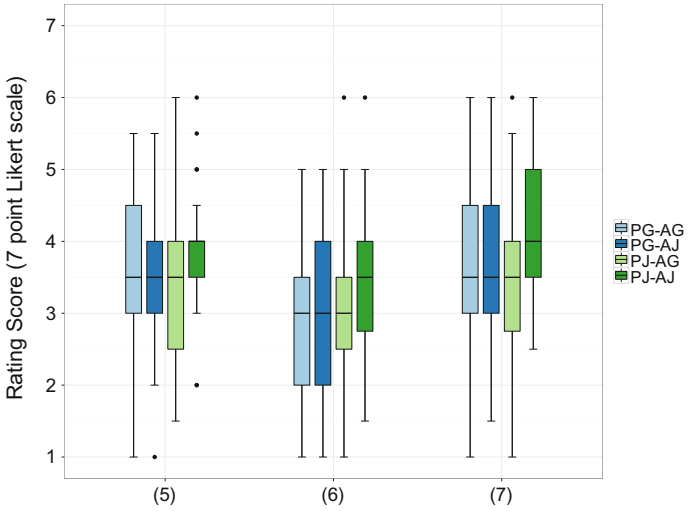

Fig. 6.11 Results of the subjective impression of the German agents (AG) and Japanese agents (AJ) for German participants (PG) and Japanese participants (PJ) with regard to the statements from Table 6.3: (*5*) "*I liked watching the conversation.*", (*6*) "*I would like to join the conversation.*", (*7*) "*The conversation or a similar conversation would be realistic in my own life/ my friends life.*"

show the ratings for German and Japanese agents. Mixed ANOVAs were conducted for every item, each with the agent-culture as independent within-subjects measure and the participants' culture as independent between-subjects variable.

The subjective naturalness of the characters was not significantly different between the agent-cultures, $F(1,50) = .727$, $p = .398$, as well as for the participants culture, $F(1,50) = .221$, $p = .640$. However, there was a significant interaction of the agent-culture and participant-culture, $F(1, 50) = 4.173$, $p = .039$. Participants did rate the agent-culture based on their own cultural background higher than the conversations of the other agent-culture. That means that Japanese participants did rate the Japanese conversations more natural than the German conversations while Germans' ratings did show the reverse effect.

There was no significant effect of agent-culture on subjective appropriateness, $F(1, 50) = 2.110$, $p = .153$. Participants' culture had no effect, $F(1, 50) = 1.723$, $p = .195$, and neither did the interaction, $F(1, 50) = 2.110$, $p = .153$.

Regarding the statement "*The characters were getting along with each other well*" there was a significant main effect of the agent-culture, $F(1, 50) = 5.744$, $p = .020$. The conversations of the Japanese agents were overall rated better. The participants' culture, however, did not show significant effects, $F(1, 50) = .3.431$, $p = .070$, and neither did the interaction, $F(1, 50) = .001$, $p = .981$.

The participants' ratings whether the characters' movements matched the conversation did not differ significantly depending on the participants' culture, $F(1, 50) = .303$, $p = .584$, on the agents' culture, $F(1, 50) = .051$, $p = .823$, or the interaction of participants' and agents' culture, $F(1, 50) = .787$, $p = .379$.

Participants did not like watching conversations of either agent-culture significantly more, $F(1, 50) = 1.425$, $p = .238$. There was also no effect of the participants' culture, $F(1, 50) = 1.002$, $p = .322$, and no effect of the interaction, $F(1, 50) = 1.958$, $p = .168$.

Concerning the statement "*I would like to join the conversation*" there was no significant effect on the participants' rating depending on the agents' culture, $F(1, 50) = 2.772$, $p = .102$. However, there was a significant effect of the participants' culture, $F(1, 50) = 4.196$, $p = .038$. Japanese participants did agree significantly more liking to join the conversation than Germans did. There was no interaction effect of agent-culture and participant-culture, $F(1, 50) = .518$, $p = .475$.

Participants' ratings whether such or a similar conversation could happen in their own life did significantly depend on agent-culture, $F(1, 50) = 6.327$, $p = .015$, with the Japanese conversations being rated more realistic. The effect of participants' culture, $F(1, 50) = .194$, $p = .661$, and the interaction effect, however, were non-significant, $F(1, 50) = 2.992$, $p = .090$.

6.7.6 Discussion

The survey was conducted to investigate whether the network can generate culture-specific non-verbal behaviour that is perceived differently by human observers of different cultures. This question comprises different hypotheses, i.e. if

(1) agent-culture, (2) participant-culture, and (3) the interaction thereof has an effect on the participants' subjective impression of the characters' conversations. The results show that the hypotheses can be confirmed in part.

In general, the ratings of the subjective impression of the Japanese agents were higher than for the German agents for all statements. However, significant effects were only found for "*The characters were getting along with each other well*" and "*The conversation or a similar conversation would be realistic in my own life /my friends' life*". This suggests that the network did indeed generate different behaviours for the characters as the videos with Japanese agents were rated better. However, this might also imply that the generated behaviour for the Japanese agents fits the conversation better than for the German agents. For example gestures might have been more suitable to the semantics of speech or their timing was better. This is in fact a limitation of the automatic generation of non-verbal behaviour. Therefore, it appears necessary to incorporate further techniques into an application using an automatic generation approach in order to validate that the selected non-verbal behaviour is suitable.

Participant culture did also have a significant effect on the subjective impression for one statement. Japanese participants were significantly more interested in joining the conversation than Germans did. Further research is necessary to explain whether this is a cultural difference or whether some design aspect of the conversations influenced the ratings.

The expectation that the participants of our study prefer agent conversations that resemble their own cultural background led to the hypothesis that there should be an interaction effect of participant-culture and agent-culture on the subjective impression. Figures 6.10 and 6.11 show that the participants' culture did mostly lead to higher ratings of agents of their own culture, albeit these effects were not significant except for the subjective naturalness confirming the hypothesis partially. As suggested, this might be due to the similarity principle [7] that states that interaction partners who perceive themselves as being similar are more likely to like each other. This principle could also apply in case participants perceive agents as being similar to themselves.

Although participants watched a neutral video first to get acquainted with the scenario, ratings still might have been influenced by details of our demonstrator. This impression is strengthened by some open-ended comments of the participants mentioning details such as graphics and sound of the videos, facial expressions or interpersonal distance of the characters—although these aspects were the same in all videos, including the neutral video.

6.8 Conclusion and Future Work

In this chapter, we presented an approach to generate culture-dependent non-verbal behaviour for virtual characters that is theory-based as well as data-driven. The approach combines advantages of procedures commonly used, as it explains the

causal relations of cultural background and resulting behaviour, and augments them by findings from empirical data. Therefore, we built on our former work where we have explored both approaches (theory-based as well as data-driven) separately.

To realize this endeavor, we relied on a Bayesian Network. While the structure of the network along with categorizations of behavioural aspects have been constructed based on existing theories and models, the parameters were learned from annotated data. The resulting network generates non-verbal behaviours based on observations for the German and Japanese cultures for given conversations. With the network we are able to keep a certain variability in behaviour by making predictions based on conditional probabilities.

Later in the chapter, we showed and performed two ways of validating the resulting model: regarding its predictive power and the perception of human observers of the generated behaviours.

The more technical validation of the network shows promising results for some of the behavioural aspects (postures and gestural dynamics), while it fails in predicting culture-related choices of gestures-types. Comparing these outcomes with our previous statistical analysis of the video data, it reflects that the resulting model can only be as meaningful as the underlying data. In those cases, where strong culture-related differences can be observed in the data, training a Bayesian Network seems a good approach to generate culture-related differences for simulated dialogues. In cases where no significant differences were found, e.g. gesture types, the network fails in predicting culture-specific behaviour. We therefore have to research the usage of gestures in more detail and analyse, for example, their concrete performance (e.g. handedness or hand-shape), their correlation to the semantics of speech, or their timely synchronization.

In the second evaluation we examined the perception of participants of the targeted cultures for different versions of behaviour generated by our network. We expected participants to prefer virtual characters with simulated culture-specific behaviour that is in line with prototypical culture-specific behaviour of the participant's culture. The survey partly confirmed our expectations. We therefore hope that the localization of the characters' behaviours can help improve their acceptance by users of the targeted cultures.

With the present chapter we want to provide guidance for other research aiming at integrating culture-specific behaviour into virtual character systems by describing a complete approach. The resulting model may be expanded by further aspects of culture-specific behaviours. In our future work, we aim at adding additional non-verbal behavioural traits that are known to be dependent on cultural background, such as head nods. In a similar way, other diversifying factors, such as gender or age, can be added.

References

1. André, E.: Preparing emotional agents for intercultural communication. In: Calvo, R., D'Mello, S., Gratch, J., Kappas, A. (eds.) The Oxford Handbook of Affective Computing. Oxford University Press (2014)
2. Aylett, R., Hall, L., Tazzymann, S., Endrass, B., André, E., Ritter, C., Nazir, A., Paiva, A., Hofstede, G.J., Kappas, A.: Werewolves, Cheats, and Cultural Sensitivity. In: Proc. of 13th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2014) (2014)
3. Aylett, R., Paiva, A., Vannini, N., Enz, S., André, E., Hall, L.: But that was in another country: agents and intercultural empathy. In: Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009) (2009)
4. Bergmann, K., Kopp, S.: Bayesian decision networks for iconic gesture generation. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) Proc. of 9th Int. Conf. on Intelligent Virtual Agents (IVA 2009), pp. 76–89. Springer (2009)
5. Buisine, S., Courgeon, M., Charles, A., Clavel, C., Martin, J.C., Tan, N., Grynszpan, O.: The role of body postures in the recognition of emotions in contextually rich scenarios. Int. J. Hum. Comput. Interac. **30**(1) (2014)
6. Bull, P.: Posture and Gesture. Pergamon Press, Oxford (1987)
7. Byrne, D.: The attraction paradigm. Academic Press, New York (1971)
8. Cassell, J., Vilhálmsson, H., Bickmore, T.: BEAT: The behaviour expression animation toolkit. In: Proc. of 28th Annual Conf. on Computer Graphics (SIGGRAPH 2001), pp. 477–486. ACM (2001)
9. Core, M., Allen, J.: Coding Dialogs with the DAMSL Annotation Scheme. In: Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines, pp. 28–35. Boston, MA (1997)
10. Damian, I., Endrass, B., Huber, P., Bee, N., André, E.: Individualized Agent Interactions. In: Proc. of 4th Int. Conf. on Motion in Games (MIG 2011) (2011)
11. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B (Methodological) pp. 1–38 (1977)
12. Druzdzel, M.J.: SMILE: Structural Modeling, Inference, and Learning Engine and GeNIe: A development environment for graphical decision-theoretic models (Intelligent Systems Demonstration). In: Proc. of the 16th National Conf. on Artificial Intelligence (AAAI-99), pp. 902–903. AAAI Press (1999)
13. Endrass, B., André, E., Rehm, M., Nakano, Y.: Investigating culture-related aspects of behavior for virtual characters. Auton. Agent. Multi-Agent Syst. **27**(2), 277–304 (2013). doi:10.1007/S10458-012-9218-5
14. Endrass, B., Nakano, Y., Lipi, A., Rehm, M., André, E.: Culture-related topic selection in SmallTalk conversations across Germany and Japan. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) Proc. of 11th Int. Conf. on Intelligent Virtual Agents (IVA 2011), pp. 1–13. Springer (2011)
15. Endrass, B., Rehm, M., Lipi, A.A., Nakano, Y., André, E.: Culture-related Differences in Aspects of Behavior for Virtual Characters across Germany and Japan. In: Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011), pp. 441–448 (2011)
16. Field, A.: How to Design and Report Experiments. Sage Publications, UK (2003)
17. Gallaher, P.E.: Individual differences in nonverbal behavior: Dimensions of style. J. Pers. Soc. Psychol. **63**(1), 133–145 (1992)
18. Hall, E.T.: The Hidden Dimension. Doubleday (1966)
19. Herrera, D., Novick, D.G., Jan, D., Traum, D.R.: Dialog behaviors across culture and group size. In: Stephanidis, C. (ed.) Universal Access in Human-Computer Interaction. Users Diversity—6th International Conference, UAHCI 2011, Held as Part of HCI International 2011, Orlando, FL, USA, July 9–14, 2011, Proceedings, Part II, *Lecture Notes in Computer Science*, vol. 6766, pp. 450–459. Springer (2011)

20. Hofstede, G.: Cultures Consequences—Comparing Values, Behaviours, Institutions, and Organizations Across Nations. Sage Publications, UK (2001)
21. Hofstede, G., Hofstede, G.J., Minkov, M.: Cultures and Organisations. SOFTWARE OF THE MIND. Intercultural Cooperation and its Importance for Survival. McGraw Hill (2010)
22. Hofstede, G.J., Pedersen, P.B., Hofstede, G.: Exploring Culture – Exercises, Stories and Synthetic Cultures. Intercultural Press, United States (2002)
23. Isbister, K.: Building bridges through the unspoken: Embodied agents to facilitate intercultural communication. In: Payr, S., Trappl, R. (eds.) Agent Culture: Human-Agent Interaction in a Multikultural World, pp. 233–244. Lawrence Erlbaum Associates (2004)
24. Isbister, K., Nakanishi, H., Ishida, T., Nass, C.: Helper agent: Designing an assistant for human-human interaction in a virtual meeting space. In: Turner, T., Szwillus, G. (eds.) Proc. of Int. Conf. on Human Factors in Computing Systems (CHI 2000), pp. 57–64. ACM (2000)
25. Kipp, M.: Anvil - A Generic Annotation Tool for Multimodal Dialogue. Eurospeech **2001**, 1367–1370 (2001)
26. Kistler, F., André, E., Mascarenhas, S., Silva, A., Paiva, A., Degens, N., Hofstede, G.J., Krumhuber, E., Kappas, A., Aylett, R.: Traveller: An interactive cultural training system controlled by user-defined body gestures. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) Human-Computer Interaction—INTERACT 2013—14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2–6, 2013, Proceedings, Part IV, *Lecture Notes in Computer Science*, vol. 8120, pp. 697–704. Springer (2013)
27. Kistler, F., Endrass, B., Damian, I., Dang, C.T., André, E.: Natural interaction with culturally adaptive virtual characters. J. Multimodal User Interfaces **6**(1–2), 39–47 (2012)
28. Kleinsmith, A., Silva, P.D., Bianchi-Berthouze, N.: Recognizing emotion from postures: Cross-cultural differences in user modeling. In: User Modeling 2005, no. 3538 in LNCS, pp. 50–59 (2005)
29. Kluckhohn, K., Strodtbeck, F.: Variations in value orientations. Row, Peterson, New York, United States (1961)
30. Martin, J.C., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., Pelachaud, C.: Levels of representation in the annotation of emotion for the specification of expressivity in ECAs. In: Proc. of 5th Int. Conf. on Intelligent Virtual Agents (IVA 2005), pp. 405–417. Springer (2005)
31. Mascarenhas, S., Dias, J., Prada, R., Paiva, A.: One for all or one for one? the influence of cultural dimensions in virtual agents' behaviour. In: Z. Ruttkay, M. Kipp, A. Nijholt, H.H. Vilhjálmsson (eds.) Intelligent Virtual Agents, 9th International Conference, IVA 2009, Amsterdam, The Netherlands, September 14–16, 2009, Proceedings, *Lecture Notes in Computer Science*, vol. 5773, pp. 272–286. Springer (2009)
32. Mascarenhas, S., Silva, A., Paiva, A., Aylett, R., Kistler, F., André, E., Deggens, N., Hofstede, G.J., Kappas, A.: Traveller: an intercultural training system with intelligent agents. In: Proc. of 12th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2013) (2013)
33. McNeill, D.: Hand and Mind—What Gestures Reveal about Thought. University of Chicago Press, Chicago, London (1992)
34. Nouri, E., Traum, D.R.: Generative models of cultural decision making for virtual agents based on user's reported values. In: Bickmore, T.W., Marsella, S., Sidner, C.L. (eds.) Intelligent Virtual Agents—14th International Conference, IVA 2014, Boston, MA, USA, August 27–29, 2014. Proceedings, *Lecture Notes in Computer Science*, vol. 8637, pp. 310–315. Springer (2014)
35. Ortony, A., Clore, G., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge, UK (1988)
36. Pelachaud, C.: Multimodal expressive embodied conversational agents. In: Proc. of 13th annual ACM Int. Conf. on Multimedia, pp. 683–689 (2005)
37. Rehm, M., André, E., Nakano, Y., Nishida, T., Bee, N., Endrass, B., Huan, H.H., Wissner, M.: The CUBE-G approach - Coaching culture-specific nonverbal behavior by virtual agents. In: Mayer, I., Mastik, H. (eds.) ISAGA 2007: Organizing and Learning through Gaming and Simulation (2007)

38. Rehm, M., Bee, N., André, E.: Wave like an egyptian: accelerometer based gesture recognition for culture specific interactions. BCS HCI **1**, 13–22 (2008)
39. Rehm, M., Bee, N., Endrass, B., Wissner, M., André, E.: Too close for comfort?: adapting to the user's cultural background. In: Proceedings of the international workshop on Human-centered multimedia, pp. 85–94. ACM (2007)
40. Rehm, M., Nakano, Y., André, E., Nishida, T., Bee, N., Endrass, B., Wissner, M., Lipi, A.A., Huang, H.H.: From observation to simulation: generating culture-specific behavior for interactive systems. AI & Soc. **24**(3), 267–280 (2009)
41. Russell, S.J., Norvig, P.: Artificial Intelligence—A modern approach (3. internat. ed.). Pearson Education (2010)
42. Sagiv, L., Schwartz, S.H.: Cultural values in organisations: Insights for Europe. European Journal of International Management **1**(3), 176–190 (2007)
43. Schneider, K.P.: Small talk: Analysing phatic discourse. Hitzeroth, Marburg (1988)
44. Si, M., Marsella, S., Pynadath, D.V.: Thespian: Modeling socially normative behavior in a decision-theoretic framework. In: Gratch, J., Young, R. Aylett, D. Ballin, P. Olivier (eds.) Intelligent Virtual Agents, 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21–23, 2006, Proceedings, *Lecture Notes in Computer Science*, vol. 4133, pp. 369–382. Springer (2006)
45. Ting-Toomey, S.: Communicating across cultures. The Guilford Press, New York (1999)
46. Trompenaars, F., Hampden-Turner, C.: Riding the waves of culture—Understanding Cultural Diversity in Business. Nicholas Brealey Publishing, London (1997)