Creating standardized video recordings of multimodal interactions across cultures

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

Angaben zur Veröffentlichung / Publication details:

Rehm, Matthias, Elisabeth André, Nikolaus Bee, Birgit Endrass, Michael Wißner, Yukiko Nakano, Afia Akhter Lipi, Toyoaki Nishida, and Hung-Hsuan Huang. 2009. "Creating standardized video recordings of multimodal interactions across cultures." In *Multimodal corpora: from models of natural interaction to systems and applications*, edited by Michael Kipp, Jean-Claude Martin, Patrizia Paggio, and Dirk Heylen, 138–59. Berlin [u.a.]: Springer. https://doi.org/10.1007/978-3-642-04793-0_9.

Creating Standardized Video Recordings of Multimodal Interactions across Cultures

Matthias Rehm[1], Elisabeth André[1], Nikolaus Bee[1], Birgit Endrass[1], Michael Wissner[1], Yukiko Nakano[2], Afia Akhter Lipi[2], Toyoaki Nishida[3], and Hung-Hsuan Huang[3]

[1] Augsburg University, Institute of Computer Science, 86159 Augsburg, Germany
rehm@informatik.uni-augsburg.de
[2] Dept. of Computer and Information Science, Faculty of Science and Technology, Seikei University, Japan
y.nakano@st.seikei.ac.jp
[3] Dept. of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Japan
nishida@i.kyoto-u.ac.jp

Abstract. Trying to adapt the behavior of an interactive system to the cultural background of the user requires information on how relevant behaviors differ as a function of the user's cultural background. To gain such insights in the interrelation of culture and behavior patterns, the information from the literature is often too anecdotal to serve as the basis for modeling a system's behavior, making it necessary to collect multimodal corpora in a standardized fashion in different cultures. In this chapter, the challenges of such an endeavor are introduced and solutions are presented by examples from a German-Japanese project that aims at modeling culture-specific behaviors for Embodied Conversational Agents.

1 Introduction

The acquisition of corpora as a basis for cross-cultural studies is a challenging endeavor. First of all, an enormous amount of data is required in order to shed light on culture-specific differences in behavior. Even when concentrating on one culture only, it is hard to assess how behavior is influenced by the context in which recordings are taken. Cross cultural studies introduce a further behavior determinant which needs to be separated out from other contextual variables, such as the situation in which the recordings are taken and the personality of the videotaped subjects. As a consequence, the acquisition of corpora needs to be carefully planned requiring cooperation between researchers across different countries. In the following, we discuss some general recommendations for the acquisition of cross cultural corpora, and detail this recommendation in later sections with examples from a large corpus study, which was conducted in Germany and Japan.

1.1 Application Scenario Match

The recorded scenario should match the envisioned application scenario as closely as possible. For example, when implementing an application for cultural training, it is advisable to record prototypical scenarios that constitute situations a tourist or any business canvasser is likely to encounter. Furthermore, we need to take into account how the user will interact with the application. If the user is just listening to an agent, the recording of monologues might suffice. In case the user is expected to engage in a dialogue with an agent, a sufficient amount of samples for dialogue and turn taking acts should be represented in the data as well. Culture is not only reflected by specific speaking behaviors. Listening behaviors may significantly differ from culture to culture as well. If a corpus is to feature typical interaction patterns, recordings of at least two people engaging in a conversation are required per culture. Further conversational phenomena arise in multi-party conversations. In a three-party conversation one person may not just listen to another person, but overhear the dialogue between two other people. Four-party conversations enable the participants to form sub groups and engage in multi-threaded conversations. Usually, a tourist or business canvasser encounters situations that include more than two conversational partners. Nevertheless, research has mostly concentrated on the recordings of two-party conversations so far.

1.2 Definition of a Common Protocol

Corpus studies have been conducted for a large variety of languages and cultures. However, since previous studies focused on different settings and phenomena, the results are difficult to compare. For example, it is hard to identify characteristic culture-specific interaction patterns if the recordings feature conversing friends in one culture and conversing foreigners in another. To conduct a comparative cross-cultural study, a common protocol needs to be defined which guides corpus collection across different cultures. On the one hand, we have to make sure that the protocol is explicit enough to enable the experimenters to replicate studies in different countries and get comparable results. On the other hand, the protocol should still allow for a sufficient amount of culture-specific variations. In particular, we have to make sure that culture-specific behaviors, such as greeting rituals, are not implicitly included in the protocol. As a minimum requirement, experimenters need to agree on the application scenarios (e.g. asking for directions or making a hotel reservation) and the conversational setting (monologue, dialogue or multi-threaded conversation). Another issue is to decide on which behavior determinants apart from culture to vary and which to keep constant. For instance, different gender constellations might be considered in order to investigate how gender affects the style of conversation in different cultures. Many factors might, however, have to be kept constant out of practical considerations in order to limit the amount of data to be recorded.

1.3 Phenomena Occurrences

We aim at recording multimodal communicative behaviors of naturally acting people, but at the same time a sufficient amount of data covering the phenomena we are interested in should occur. To ensure a high control over the recordings, the employment of actors who are not known to the subjects in advance might be a promising option. Actors may help keep the conversation going and provoke a sufficient amount and variety of behaviors from the subjects. Furthermore, we have to decide on which cultural dimensions to focus. For instance, to analyze the influence of the power distance index on communication, scenarios with people that represent a different status might yield more interesting results than scenarios with people of equal status. Again, actors may help ensure that differences in status are reflected in a controlled manner. Naturalness of the data would still be guaranteed by just considering the non-acted behaviors in the later analysis.

1.4 Separating Culture from Other Behavior Determinants

Contextual factors that might have an impact on behavior need to be controlled and to be separated out from cultural-specific behavior determinants. Such factors include the participants' gender, their personality, and their social status. For instance, it might be observed that the spatial extension of gestures in a corpus is unusually high. In Germany, this might easily be attributable to the recorded persons' extrovert personality trait. But what if the data were from a Southern European country like Italy? Then the observed expressive behavior might well just be the standard behavior pattern in this culture (e.g., Ting-Toomey 1999). In order to identify the potential influence of the subjects' personalities on their behaviors, it is highly recommended to record the personality traits of the participating subjects. In the ideal case, potential gender effects on dialogue may be taken into account by taking recordings of all possible gender combinations (female-female, female-male, and male-male) which is, however, not feasible when investigating multi-party conversations due to the many combinations to be considered.

1.5 Technical Requirements

The subjects need to be recorded from various viewing angles in order to ensure that their gestures and facial expressions are visible. When aiming at studying typical interaction patterns, not only the speakers, but also the addressees need to be videotaped. In order to allow for comparing cross-cultural studies, the recordings should be taken from similar angles in all participating countries. For example, a comparison might become difficult if gestures are better visible in one culture than in another or if a person appears more dominant in one culture due to the viewing angle chosen. For the same reason, the audio quality needs to be kept constant across cultures. Another question concerns the choice of an appropriate setting for the recordings. To ensure that the recordings match the application as closely as possible, they might be taken in a similar situation.

Here, we have to take into account, however, that the context introduces a new variable which might be difficult to separate out from culture-specific factors. A neutral room helps avoid any bias due to context, but people might behave differently than in the application situation.

2 Culture

Aiming at simulating culture-specific behavior in interactive systems, it is necessary to clarify first what is meant by culture and second which observable phenomena are influenced by culture. We will not discuss different cultural theories here, but present one theoretical school that defines culture as norms and values that guide the behavior of people from the given culture. Culture in this approach provides a set of heuristics that structure the behavior and its interpretation by the members of a culture. One representative of this line of thinking is Hofstede [15], who develops a dimensional model of culture. His theory is based on a broad empirical survey that gives detailed insights in differences of value orientations and norms. Hofstede defines five dimensions on which cultures vary, the best known is perhaps his identity dimension that distinguishes between collectivistic and individualistic cultures and defines to what degree individuals are integrated into a group. The other dimensions are hierarchy, gender, uncertainty, and orientation. Hierarchy denotes if a culture accepts unequal power distance between members of the culture or not. Identity defines to what degree individuals are integrated into a group. Gender describes the distribution of roles between the genders. In feminine cultures for instance roles differ less than in more masculine cultures. Uncertainty assesses the tolerance for uncertainty and ambiguity in a culture. Those with a low uncertainty tolerance are likely to have fixed rules to deal with unstructured situations. Orientation distinguishes long and short term orientation, where values associated with short term orientation are for instance respect for tradition, fulfilling social obligations, and saving one's face often resulting in elaborate systems of politeness. Following this approach, a given culture is defined as a point in a five-dimensional space.

According to Hofstede, nonverbal behavior is strongly affected by cultural affordances. The identity dimension for instance is tightly related to the expression of emotions and the acceptable emotional displays in a culture. Thus, it is more acceptable in individualistic cultures like the US to publicly display negative emotions like anger or fear than it is in collectivistic cultures like Japan. Based on Hofstede's dimensions, Hofstede, Pedersen, and Hofstede [16] define synthetic cultures as representations of the end points of the dimensions and show how specific behavior patterns differ in a principled way depending on where a culture is located. Table 1 presents a summary for the acoustic and spatial behavior of these synthetic cultures, which can serve as a starting point for modeling culture-specific behavior [27]. Using this information as a predictor for the acoustic and spatial behavior of the German and Japanese culture leads to some problems as is exemplified with the following example. High power distance

Table 1. Synthetic cultures and corresponding patterns of behavior for low and high values. For illustrative purposes, the positions of the German (G) and the Japanese (JP) culture are given. If they fall into the same category (e.g. high values on the gender dimension) the relative position is indicated.

Dimension	Synthetic Culture	Sound	Space	Example
Hierarchy	Low: Low Power	Loud	Close	German
	High: High Power	Soft	Far	Japanese
Identity	Low: Collectivistic	Soft	Close	Japanese
	High: Individualistic	Loud	Far	German
Gender	Low: Femininity	Soft	Close	
	High: Masculinity	Loud	Close	G < JP
Uncertainty	Low: Tolerance	Soft	Close	
	High: Avoidance	Loud	Far	G < JP
Orientation	Low: Short-Term	Soft	Close	German
	High: Long-Term	Soft	Far	Japanese

(hierarchy dimension) results standing further apart in face-to-face encounters whereas collectivism (identity dimension) generally means standing closer together in the same situation. Both attributions hold true for the Japanese culture. Thus, what will be the result of these correlations if they are combined? The most sensible solution would be to consider the semantics of the dimensional position. If a culture has a high power distance then there could be differences in proxemics behavior that are related to social status, for instance resulting in standing further away from high status individuals but closer together with peers. Another obvious problem with this information is that the actual behavior of an existing culture might be completely different from this stereotypical behavior. To realistically model culture-specific patterns of behavior for interactive systems, it becomes thus necessary to capture the necessary empirical data for a given culture.

3 Application Scenario Match

The verbal and non-verbal behavior of embodied conversational agents (ECAs) becomes more and more sophisticated. But this behavior is primarily based on a Western cultural background due to the available agent systems and their predefined animation sequences. But according to [31] the most profound misunderstandings in face-to-face communication arise due to misinterpretations of non-verbal cues. Thus, culture-adaptive behavior in embodied agents can further cross-cultural communication in two ways. (i) Employing an agent that adheres to culturally determined behavior programs will enhance the efficiency of information delivery. (ii) Agents capable of changing their "cultural programs" can serve as embarrassment-free coaching devices of culture-specific behaviors. Based on Hofstede's theory of cultural dimensions [15], the CUBE-G[1] project

[1] CUlture-adaptive BEhavior Generation for embodied conversational agents http://mm-werkstatt.informatik.uni-augsburg.de/projects/cube-g/

investigates whether and how the non-verbal behavior of agents can be generated from a parameterized computational model, The project combines a top-down model-based approach with a bottom-up corpus-based approach which allows to empirically ground the model in the specific behavior of two cultures (Japanese and German).

One of the central goals of the project is to develop a role-playing scenario to increase cultural awareness following generally accepted learning steps. Bennett [2] as well as Hofstede [16] describe similar approaches that are widely used in real-world trainings and that will be adapted for the use in CUBE-G. The focus lies on scenarios that every tourist or ex-patriate is likely to encounter. A first meeting between strangers, a negotiation process, and an interaction of individuals with different social status have been identified to serve this purpose due to their prototypical nature, i.e. they can be found in every culture and they constitute situations a tourist or ex-patriate is likely to encounter. Thus, they present the three scenarios for collecting multimodal data on face-to-face interactions in the German and the Japanese culture.

3.1 Scenario 1: First Meeting

There are several specific reasons for including the first-meeting scenario. According to Kendon [18], it is not only found in all cultures but it also plays an important role for managing personal relations by signaling for instance social status, degree of familiarity, or degree of liking. There is also a practical reason for this scenario because it is the standard first chapter of every language textbook and thus known to everybody who ever learned a foreign language revealing a potential application of the results in a role-play for first meeting scenarios. For Argyle [1], a first meeting is a ritual that follows pre-defined scripts. Ting-Toomey [31] follows his analysis by denoting a first meeting as a ceremony with a specific chain of actions. Knapp and Vangelisti [19] emphasize a first meeting as a step into the life of someone else, which is critical for a number of reasons like face-keeping or developing a network of social relation. Thus, the ritualistic nature of a first meeting makes sense in order "to be on the safe side" by establishing such a new relationship in a satisfactorily, i.e. facekeeping, manner for both sides.

Previous studies established some differences in nonverbal behavior for the German and the Japanese cultures. According to [31], 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. On the other hand, gesture usage should be more frequent in an individualistic country like Germany. It has been shown [11] that more body contact can be expected in Germany due to ritualistic handshakes during the greeting. The corpus analysis in CUBE-G could confirm these findings (only partially for the body contact claim, see [28]). Additional results include differences in posture and in gestural expressivity.

3.2 Scenario 2: Negotiation

Whereas the main focus in the first meeting scenario was on the nonverbal behavior of the interlocutors, which we expect to be qualitatively similar in both scenarios, the negotiation scenario adds an additional layer concerning differences in the styles and strategies for such a negotiation. Following [32] we categorize such strategies into three main classes:

- Avoidance: Using the avoidance strategy, a negotiation partner tries to avoid the negotiation which is undesirable for him in some manner. This might be performed by shifting the focus of the conversation to something different or trying to get completely out of the conversation.
- Integrative: Following the integrative strategy, the negotiation partner tries to find a solution for the given problem that is satisfying for all participants. This includes for example the aim to understand the other's perspective of the situation.
- Distributive: Being in a distributive strategy the conversation partner wants to carry out his point and to "win" the negotiation. This might either be conducted in an offensive way as in criticizing the negotiation partner or in a more defensive way such as referring to prior commitments.

This categorization originates from [26], who relate these categories to cultural differences in decision conferencing. They state that the higher a culture scores on Hofstede's hierarchy dimension, the higher the probability for choosing an avoidance strategy gets. The position on the uncertainty avoidance dimension also influences the style of a negotiation. The higher the value on this dimension gets the more emotion and aggression is to be expected in a negotiation as well as more interpersonal conflict. Thus, the urge to find a solution is very strong and with it the probability for an integrative strategy increases and the probability for an avoidance strategy decreases. But as in cultures that score low on this dimension everyone's opinion is taken into account, here too the integrative strategy is the most probable. The identity dimension suggests that people from individualistic cultures tend to stand behind their attitudes. As a consequence, for them the task itself, i.e. the negotiation, is always in the main focus of the conversation. For people from collectivistic cultures, harmony is more important and thus less interpersonal conflict arises in a negotiation. The same holds true for the gender dimension, where the probability for interpersonal conflict increases with increasing masculinity.

An alternative differentiation of negotiation styles is introduced by [36] focusing on how an argument is created:

- Authority: Referring to a well known person that holds the same view.
- Norms and laws: Claiming that one is stating a generally accepted norm.
- Facts: The argument is based on objective and provable facts.
- Experience: Claiming own experience to exemplify the point.
- Analogies: Putting things in perspective by comparisons.
- Logic: Drawing conclusions from a chain of logical arguments.

Still another categorization takes the communicative functions into account that can be observed during a negotiation [6]:

- Task: The conversation partners are busy with solving the task (negotiating with each other)
- Task-management: The conversation partners talk about the task (the negotiation or situation itself)
- Communication-management: The conversation partners aim at maintaining the conversation, focusing on contact, perception, and understanding.
- Other level: Here all other possible parts of a negotiation are summarized, such as small talk or jokes.

The frequency of these functions and their distribution in an actual negotiation depends on the cultural background of the interlocutors. For instance, acknowledging understanding (communication management) is more likely to occur in collectivistic cultures like the Japanese [31]. As Western cultures are on the individualistic and masculine sides of Hofstede's dimensions they are expected to take an aggressive approach to reach a solution exhibiting a structured and analytical approach in their negotiation [30]. This is supported by the fact that western cultures are short term cultures, which suggests that they tend to solve problems quickly and in a sequential manner focusing mainly on the task itself. Eastern cultures on the other hand are found on the collectivistic and long-term sides of the dimension resulting in a slower and more exhaustive way of problem solving where every opinion is taken into account and harmony is at stake resulting in an increased frequency of the communication management and other related contributions.

The negotiation scenario is a variant of the standard lost at sea scenario [37]. Subjects have to assume that they are shipwrecked in the Pacific Ocean. They have only time to take three items with them that could help them in surviving. On the boat there are fifteen items and thus they have to choose among those fifteen. Every subject has to choose his top three items from this list of fifteen and then they had to negotiate to come up with a single three item list ranked in order of importance for surviving. This is afterwards compared with the "official" list by the U.S. Coast Guard (see below). This scenario has the advantage of forcing the subjects to come to a consensus about the relevant items and as their monetary award depends on how close they come to the official list they have an intrinsic motivation to argue for their choices.

3.3 Scenario 3: Status Difference

The different positions of the German and the Japanese culture on Hofstede's hierarchy dimension lead to some obvious differences in behavior. The hierarchy dimension describes how cultures deal with differences in power among the members of the culture. Cultures with high power distance accept this difference as a natural phenomenon, accepting decisions by persons with higher power. In cultures with low power distance, power is often only given temporarily and does

not constitute a merit in itself. Thus, decisions tend to be challenged regardless of the actual position of the interlocutors.

For the task of planning video recordings of such interactions, it becomes very relevant how status is created in the experimental condition. A literature review revealed that it is quite difficult to create a natural situation with status differences and that most studies rely on a task-oriented design, where the status of the interlocutors is ascribed by their roles in a role-play (e.g. one subject is teacher, the other is student). The difficulties start even earlier with defining what is meant by status. Berger and colleagues [3] distinguish between diffuse and (task-)specific status, where diffuse status denotes characteristics that are not relevant to the task at hand, e.g. gender, race, age, occupation or education, and (task-)specific status denotes characteristics and abilities that are relevant to solving the task at hand, for instance mechanical skills to solve a problem with a car. A similar suggestion is made by Trompenaars and Hampden-Turner [34], who distinguish between ascribed and achieved status. Ascribed status summarizes characteristics that describe what a person is, e.g. age, gender, social connections, education, or profession and thus correspond to the diffuse status by Berger and colleagues. Achieved status on the other hand is constituted by characteristics that describe what a person does, e.g. a person's track record and correspond to the specific status as one's track record can only establish a higher status if it is relevant to the task at hand.

The next challenge to be solved is how to assign status to the participants of the video recordings. Different suggestions have been made to deal with this problem. Hall and Friedman [14] choose subjects among members of the same company, thus ensuring that a natural status difference is in place. On the other hand, this didn't allow for controlling how status is created and thus it became difficult to pinpoint behavior differences to specific aspects of status. Others have deliberately created status differences in experimental situations. Leffler and colleagues [22] assigned different roles to their subjects like teacher and student and observed how interaction changed due to this role assignment. The problem here is that students had to act like teachers without being ones. This means that they acted the stereotypes they know about teachers and it cannot be claimed that the result resembles an actual teacher student interaction. Knoterus and Greenstein [20] create status difference by seemingly objective manners. For instance, two subjects had to take an IQ test before an intelligence-based experiment. Task-specific status was faked by telling each subject that they scored average while the other scored extremely high or extremely low. Ungar [35] suggests to concentrate on diffuse status by relying on superficial features like clothes. In his studies he observed how subjects interacted with confederates, whose status was faked by clothing, for instance suit, tie, topcoat vs. overall and construction helmet.

For the corpus study we were interested in how subjects interact with interlocutors of seemingly higher status. We focused on diffuse status to establish this difference by exploiting the cover story of the study. Subjects were recruited by flyers telling them that a large consulting company for the automobile industry

is conducting a study on negotiation styles in different countries. Thus, there was a representative from this consulting company present as the main leader of the experiment and he was the one doing the debriefing session after subjects had finished their negotiation task. The representative was of course dressed accordingly, he was equipped with some articles bearing the logo of the consulting company, and he had the "official" list from the U.S. coast guard to match the subject's result against. Although such a list does exist, it is modified dynamically according to the result of the negotiation process in order to create the following specific situations:

1. Positive: The first answer of the subjects is rated as being among the top three items and indeed ranked as the number one item ensuring the subject 10 Euro.
2. Neutral: The second answer of the subject is rated as being among the top three items but not ranked as the second important item. Thus, the monetary reward only increased by 5 Euro.
3. Negative: The third answer of the subject is rated as completely wrong resulting in no additional monetary reward.

For the third scenario we expected different behaviors between the two cultures as well as between this scenario and the preceding two in the same culture. Leffler and colleagues [22] describe three different categories of status dependent nonverbal behavior:

- Proxemics: High status individuals take more space and are less invaded by others.
- Vocalic: High status individuals talk more, longer, and louder (see also [5]), interrupt more frequently and successfully, and laugh less frequently.
- Symbolically intrusive: High status individuals may point at others, direct them or shut them up with a gesture.

Ellyson and colleagues [8] add that high status individuals have a higher visual dominance ratio, i.e. they look at others more when speaking than listening. Johnson [17] observed that low status individuals use more verbal facilitators like "yeah" or "mhmm".

These findings suggest that our subjects will stand further away from the interlocutor in the third scenario than in the other two scenarios, that the subjects might talk in a softer voice, and that the subjects will interrupt the interlocutor less frequently in the third scenario. Moreover, we expect differences between the cultures based on their position on the hierarchy dimension. The Japanese subjects are expected to accept the statements of the high status interlocutor more readily than the German subjects that are expected to question the "official" answer or at least to demand an explanation for the "official" ranking.

4 Definition of a Common Protocol

To ensure the replication of conditions in all cultures, a common protocol has to be established on how to conduct the study with detailed instructions to be

followed at every step. These instructions have to cover recruiting of subjects and actors, the timeline of each recording, scripts for the people conducting the experiment as well as detailed information about the necessary materials and the setup of the equipment. The CUBE-G protocol is detailed here as an example of such a protocol. Another necessary prerequisite is a common language for the researcher, which is English in the case of CUBE-G.

4.1 Recruiting

To recruit subjects, a believable cover story had to be manufactured that produces plausible answers for what is going on during the recordings. A flyer was produced stating that a large consulting company for the automobile industry is conducting a study simultaneously in different countries and that Augsburg (Kyoto) has been chosen as one of the locations in Germany (Japan). The objective of the study was to investigate negotiation styles in the participating countries. A monetary award was promised, the amount of this award depending on the outcome of the negotiation process. This was to ensure that subjects had an intrinsic motivation in the negotiation. The negotiation partner of each subject is played by an actor to ensure comparable conditions for each subject (see Section 5 for more details). The first meeting was introduced as a prerequisite to the negotiation process, ensuring a minimal acquaintance of the interlocutors, the high status scenario was introduced as the debriefing from the negotiation process with the representative of the consulting company, which was also played by an actor.

4.2 Timeline

To illustrate the timeline, the study was piloted several times in detail and an audiovisual protocol of this pilot was created. It turned out that this material is well suited to serve as a kind of "storyboard" illustrating the general timeline of the study. Figure 1 gives an impression of such a "storyboard". The subject is welcomed and asked to fill out the personality questionnaire as well as some additional information concerning data protection issues before it is let to the recording room. There the actor playing the second subject is already waiting. The subject is welcomed again this time by the representative of the consulting company and another lab person, who also introduces the goal of the study. Together with the actor playing the second student the subject is led to the recording area where they have time to get acquainted for the actual experiment. After five minutes, another lab person disrupts the conversation and leads the two subjects back to their tables where they have time to read the experimental instructions and prepare the negotiation task. After ten minutes, subjects are led back to the recording area, where they start negotiating. Discussions continue until they reach a consensus (10 to 15 minutes). Then the actor playing the role of the second subject is led out of the room to allow for separate debriefings. The representative of the consulting company enters the recording area with the "official" list and the subject has to report and defend their choices.

Fig. 1. “Storyboard” to illustrate the timeline of the study

Fig. 2. Snapshots from application with virtual characters

As can be seen in Figure 1, subjects and actors are recorded standing. Although sitting at a table would seem to be a more natural setting for the task, this would prevent us from eliciting most of the non-verbal behaviors we are interested in like proxemics, gestures and postures. Additionally, the envisioned application will take place in a virtual meeting place where groups of agents and users stand together and interact (see Figure 2). Thus, doing the recordings as shown in Figure 1 increases the application scenario match significantly.

4.3 Instructions

Along with the information about the timeline of the experiment, detailed instructions are necessary for each participant in the study. In the case of the CUBE-G project, these are two lab people welcoming the subject and introducing the task to the subjects, and two actors, one acting as another student and one acting as the principal investigator, i.e. the representative from the consulting company.

Lab people. To ensure the same conditions and especially the same amount of information for every subject, it is important to define in detail what is said and done by the people running the study. The welcoming people are only allowed to recite the cover story and to instruct subjects about the personality test and the data protection issues (formal protocol that has to be approved by the local data protection officer). The lab people in the recording room need a standardized text for a short introduction based on the cover story and detailed experimental instructions in written form for the subjects. They are entitled to answer questions that might arise concerning the mechanics of the study but otherwise have to stick to the cover story.

Actors. Instructions for the actors differ depending on their roles. The student actor is supposed to assume a passive role letting the subject lead the discussion and only take the initiative when this strategy fails. For the first meeting, the

actor can rely on his own background story for everything that does not concern his university life. For this part of his biography he is supplied with a cover story that ensures that he does not study the same subject as the participant to prevent detailed discussion about specific courses or teachers, which might easily blow the cover. For the negotiation scenario, the actor has to ensure that at the beginning he agrees only on one item with the subject, ensuring that they have to discuss at least four items. For the discussion, actors have been supplied with pro and contra arguments for each item to ensure that they (i) have enough arguments and (ii) that every subject has to face the same arguments for a given object on the list.

The high status actor acts as the representative of the consulting company from the back story. He follows a more constrained script to induce certain affective states in the subjects in order to analyze the potential differences in reacting to these situations. The first item from the subject's list has to be the top item of the "official" list ensuring a high monetary award and thus inducing a positive state in the subject. The second item had to be the third on the list, thus being just not right and ensuring less money. The third item had to be completely wrong, resulting in no money for this item and thus inducing a negative state in the subject. The hypothesis is that German subjects start to question the "official" list more often than the Japanese subjects.

4.4 Material and Technical Setup

Apart from the above mentioned instructions, material for the study includes personality questionnaires in German and Japanese and additional instructions for filling out the questionnaires in German and Japanese (see Section 5). Instructions for the subjects include detailed information for the negotiation task, i.e. the variation of the lost at sea scenario, and an informed consent concerning the data protection issues. How this issue is dealt with depends on local regulations in the countries that participate in the study. For the technical setup, materials include schematic figures on the configuration of the cameras, microphones, and other equipment, which is accompanied by audio-visual material from the piloting sessions from which the "storyboards" described above have been extracted. Most of the material has to be translated into the target language of the study (here German and Japanese).

Apart from the material presented, the recording sessions for the CUBE-G project were followed in each country by a representative from the organizing team to ensure that conditions match each other as close as possible.

5 Phenomena Occurrences

This challenge is not specific for capturing cultural-specific behavior but is of a more general kind. It concerns the decision which phenomena will be analyzed and how it can be ensured that enough occurrences of these phenomena are elicited during the recordings. In CUBE-G, we decided to make use of actors

as interaction partners for our subjects based on two considerations. First, we wanted to make sure that all subjects are confronted with the same conditions. By using actors, we were able to create scripts for each scenario that defined (sometimes strictly, sometimes loosely) how the interaction should take place. The second reason was to ensure that enough material is elicited. Our fear was that subjects would agree very soon on three items during the negotiation to get done with it. Using actors we could ensure that subjects really had to argue to make their point.

The next challenge concerns the question how the material is processed for the analysis. Multimodal corpora have been in the center of research in computer science for over a decade now. Thus, a large amount of annotation schemes have been proposed to cover the analysis of such data on which one can draw. Additionally, standards have been introduced for evaluating the validity of a scheme (e.g. [21]). The CUBE-G corpus allows for analyzing verbal as well as nonverbal behavior. Currently, the analysis is focusing on nonverbal behavior employing the following annotation schemes:

- Verbal channel: Currently, the dialogue is transcribed in the language of the interlocutors and an English translation of this transcript is provided.
- Gesture channel: Gestures are analyzed on two different levels of granularity. Following a well-established coding scheme by McNeill [24], the type of gestures is analyzed focusing on conversational gestures and distinguishing between emblematic, deictic, iconic, and metaphoric gestures. Additionally, it is analyzed how a gesture is performed relying on features of expressivity that have been described by Efron [7] (for examining cultural differences) as well as Gallaher [10] (for examining differences in personal style) and taking aspects like the spatial extent of a gesture and its speed into account.
- Posture channel: Postures are annotated based on Bull's suggestions [4], who defined coding schemes for head, trunk, arm, and leg movements.
- Proxemics channel: This channel describes the overall spatial behavior of the interlocutors. Coding of this channel follows Hall's [13] definition of spatial areas that trigger different behavioral routines, distinguishing between intimate, personal, social, and public spaces.
- Speech volume: As is evident from Table 1, volume depends on a culture's position on Hofstede's dimensions and is coded here in terms of three values (low, med, high). Preliminary observations showed that volume of speech seem to be gender related as well.

An in-depth description of the annotation schemes for gestures and postures as well as some results from the corpus study can be found in [28]. As we expect culture-specific differences in the observed behavior, we also have to assume culture-specific differences in the analysis of this behavior. For instance, if it is typical in a given culture to use a lot of space to perform gestures, then this might be regarded as the baseline against which the observed behavior is matched. Thus, what is interpreted as a large gesture in Germany might just be medium in Italy. This is one of the main reasons why for instance Northern Europeans are perceived as cold and distant by Southern Europeans. Consequently,

standardized instructions for the analysis of the data are necessary, defining in detail what constitutes for instance high, medium, and low spatial extent in general to avoid culture-specific interpretations.

6 Separating Culture from Other Behavior Determinants

Although definitions of culture state that norms and values lie at the bottom of heuristics of behavior and interpretation of behavior (see Section 2), it is not obvious if a specific behavior that is observed in an interaction is determined or at least influenced by the interlocutor's cultural background or has quite different origins, for instance rooting in the interlocutor's personality. Other influencing factors can be the interlocutor's gender, his age or even his current emotional and motivational state. For instance, it might be the case that mixed-gender pairs show different interaction patterns than same-gender pairs, and that male pairs again show differences compared to female pairs. In the CUBE-G corpus, we observed for instance that female Japanese subjects spoke in a quieter voice when interacting with males compared to their interactions with females.

Preparing for the video recordings in the CUBE-G project, we decided to control for gender and personality effects and to keep other factors like age and educational background as constant as possible by recruiting students of a specific age group at the universities in Augsburg and Kyoto.

6.1 Controlling for Gender

To control for gender, we took all possible pairings into account for the video recordings. For combinations of more than two interlocutors, this control mechanism soon becomes unfeasible due to the number of combinations that have to be considered. One of the interaction partners in each scenario was an actor (see previous section) following a script for the specific situation. After having met the actor for the first time, subjects negotiate with the same actor. Afterwards they interact with a person of seemingly higher status who is played by a different actor. Table 2 summarizes the design of the recordings. 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. Actors in scenarios (i) and (ii) played the other students, whereas the actors in scenario (iii) played

Table 2. Design of the corpus study to control for gender effects

First Meeting		Negotiation		Status Difference	
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 roles of representatives of the consulting company that conducted the study. Thus, apart from the two male (M_{A1}, M_{A2}) and two female actors (F_{A1}, F_{A2}), ten male (M_{S1}-M_{S10}) and ten female subjects (F_{S1}-F_{S10}) were needed for this corpus study.

Due to some over recruiting we recorded 21 pairs in Germany and 26 pairs in Japan. The recordings took place over four days in each country, where the first day was used for rehearsals with the actors.

6.2 Controlling for Personality

The main focus in CUBE-G is on culture-specific differences in nonverbal behavior. In Section 3 it was detailed what kind of differences can be expected deriving from different cultural backgrounds. For instance, gestural expressivity is often linked to cultural variables. For instance, Southern Europeans tend to use gestures more frequently than Northern Europeans [31], and Italian immigrants to the US show more expansive gestures than Jewish immigrants [7]. Others have shown that similar effects can be seen when the focus is laid on personality instead of culture. Gallaher describes gestural expressivity as a function of personal style and links it to different personality traits. Thus, the extensive gesture use of an interlocutor is not necessarily attributable to his Southern European origin; instead it could just be a trait of his extrovert personality. As a consequence, it seems inevitable to control for personality when investigating cultural differences in nonverbal behavior.

The best strategy would be to test subjects beforehand and design the experiment in order to capture all combinations of personality traits x gender. This is just not possible due to the large amount of material that would be produced and that would just not be possible to analyze. Thus, we opted for an a posteriori solution. All subjects had to take part in a personality assessment, which allows us to take the subjects' personality into account for the later analysis and perhaps even get indications on correlations between personality profiles and cultural dimensions. It is a question of some dispute if such correlations exist. Triandis and Suh [33] for instance review work on cultural influences on personality and culture and give an excellent overview of their interrelations. Thus, in the long run, an integrated model is needed that combines cultural variables and other influence factors. Nazir and colleagues [25] e.g. propose a first model that relates culture and personality in a cognitive architecture. A number of standardized tests exist for which it has been shown that they are applicable in different cultures. We chose the NEO-FFI assessment, which defines personality as a five dimensional space. It has been shown to work well in different cultures [23].

7 Technical Requirements

To produce comparable data sets it is indispensable to define technical requirements for the video recording sessions. This includes the specifications for the

recording equipment as well as the layout of the recording area to be able to reproduce the recording conditions. Both aspects are determined by the phenomena that are captured. For instance, if the focus is on the automatic analysis of facial expressions, it becomes necessary to capture the subject's face with the camera face-to-face with the subject. As the camera itself can have an influence on the subject's behavior, such a setting might prove too obtrusive to capture the relevant phenomena. Thus, the dilemma has to be solved to provide a situation that is as natural as possible while still allowing capturing the relevant phenomena. Another decision concerns the recording area itself. In CUBE-G, we opted for a neutral room that does not provide any distractions from the task. To be able to establish similar conditions in both countries, the specification for the recording area as well as for the technical equipment has to provide a high level of detail.

7.1 Specifying the Recording Area

The room was designed in order to focus subjects on the task at hand, i.e. interacting with the interlocutor, providing as little contextual cues as possible that could attract the attention of the participants. Additionally, the actual recording equipment was integrated in such a way that it remained as far in the background as possible. Figures 3 gives an impression of the recording rooms in Germany and Japan. The most obvious piece of equipment in each case is the microphone that had to be placed in a prominent place to capture the audio information. An alternative would have been to use headsets but it was decided that the use of headsets would have been even more obtrusive than placing the microphone in the recording area. The area itself was around 3 x 5m with a single entrance that allowed for controlling the location of the subject relative to the actor by leading the subject into the room first, which resulted naturally in the subject choosing the position opposite of the entrance. Additionally, in the third scenario, it was a natural move to lead the actor subject out of the recording area because his location was near the entrance, allowing us starting the debriefing session with the actual subject.

7.2 Specifying the Technical Equipment

Figure 3 gives an impression of the technical equipment that was used. Two video cameras capture the actions by the actor and the subject, one focusing on the actor, the second one focusing on the subject. A webcam was installed on the floor outside the recording area that captured the feet of the interlocutors allowing for analyzing the proxemics behavior. Audio was captured by an external microphone that was connected to the camera capturing the subject. Because our focus was primarily on the subject's behavior, we refrained from synchronizing the cameras. Although layout and equipment had been tightly specified, it turned out that some uncontrollable effects made it necessary to adjust this setting on the fly. Some of the Japanese subjects sat down on the floor for the interaction. In hindsight this is not a surprising effect but neither

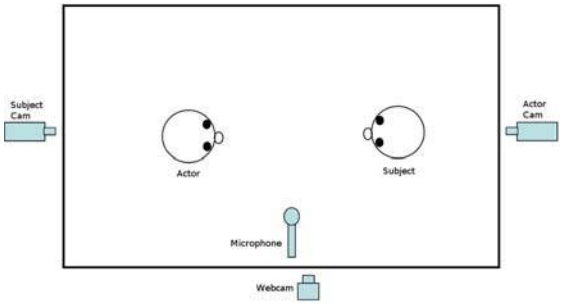

Fig. 3. Layout of the recording area

the Western nor the Japanese researchers took this as an option into account. Thus, the camera positions had to be adjusted during the recordings in two cases. Additionally, the placing of the microphone was also tailored to standing interactions, resulting in bad audio quality for these pairs.

8 Conclusion

The work presented in this article is part of a larger research endeavor that aims at enculturating interactive systems acknowledging the fact that the user's cultural profile provides heuristics of behavior and interpretation that influence his interactions. Non-verbal behavior, which was the focus in this article, is only one aspect of this endeavor. Other aspects include verbal behaviors like small talk, emotional facial displays, cognitive abilities like the appraisal of incoming data, or the appearance of the agent. Research in this area is in its infancy and one of the current challenges is to determine the importance of the different factors on the interaction. Complexity arises from the ill-defined nature of the domain which makes it difficult to reliably specify the links between cultural profiles and behavior traits. In this article, we gave an introduction into the specific challenges that arise by recording multimodal corpora in different cultures in order to capture behavior patterns that are attributable to the cultural background of the interlocutors with the aim of explicating some of these links. These challenges include the decision for appropriate scenarios, the establishment of common protocols for the recordings, the elicitation of relevant behavior, the identification of influencing variables, and the technical requirements for the actual recordings. Solutions to these challenges have been proposed by examples from the CUBE-G project, which aims at modeling culture-specific behaviors for Embodied Conversational Agents.

Following the suggestions given in this chapter produces a large amount of data that has to be analyzed. So far, purely statistical analyses have been conducted on the annotations of posture and gestural expressivity data. The results reveal significant differences between German and Japanese samples in preferred hand and arm postures, in most of the dimensions for gestural expressivity as

well as in communication management and small talk behavior. Details of the analyses can be found in [28] and [9]. As has been noted earlier, it can be expected that such a purely statistical analysis does only reveal some general trends and that the semantics of the dimensions and the interaction context have to be taken into account. For this purpose, it is indespensable to take the semantics of Hofstede's dimensions into account. To this end, we controlled for gender effects and status differences in order to be able to compare differences in behavior and link these differences to Hofstede's dimension like gender or hierarchy.

The question has been raised if the use of students is a valid approach for capturing relevant data of culture-specific interactions as they represent at best a subgroup of the culture. Relying on Hofstede's theory of culture, we define culture as national culture and from this perspective the German students can be expected to adhere in general to the cultural heuristics for the German national culture and the Japanese students to the heuristics for the Japanese national culture. Thus, the question is more fundamental in asking if the cultural theory of Hofstede is a good choice. The undeniable advantage of Hofstede's approach (and similar ones like Hall [12] or Schwartz and Sagiv [29]) is the very clear level of abstraction, i.e. national cultures. A popular counter argument runs like this: National cultures are all very well but our systems do not deal with THE German user but with a bank accountant from Berlin or a farmer from the heart of Bavaria. We would like to draw an analogy here to the treatment of different languages in HCI. Although there is without question a difference in how the Bavarian farmer and the bank accountant from Berlin speak, localizing a system for the German market nevertheless assumes something like a standard German language, which is as fictional as the standard German culture. Nevertheless, this abstraction works very well for the time being. Thus, it remains to be shown if a system that equally idealizes cultural heuristics on a national level does not have its merits in serving as a starting point for enculturating interactive systems.

The CUBE-G corpus presents a rich resource of cross cultural data. By and by, starting with the end of 2009, chunks of the corpus will be made available to the research community. For up-to-date information and condition please check the project website[2].

Acknowledgements

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

1. Argyle, M.: Bodily Communication. Methuen & Co. Ltd., London (1975)
2. Bennett, M.J.: A developmental approach to training for intercultural sensitivity. International Journal for Intercultural Relations 10(2), 179–195 (1986)

[2] http://mm-werkstatt.informatik.uni-augsburg.de/projects/cube-g/

3. Berger, J., Fisek, M.H., Norman, R.Z., Zelditch Jr., M.: Status Characteristics and Social Interaction: An Expectation-States Approach. Elsevier, Amsterdam (1977)
4. Bull, P.E.: Posture and Gesture. Pergamon Press, Oxford (1987)
5. Burgoon, J.K., Buller, D.B., Woodall, W.G.: Nonverbal Communication: The Unspoken Dialogue. Harper and Row, New York (1989)
6. Core, M., Allen, J.: Coding Dialogs with the DAMSL Annotation Scheme. In: Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines, Boston, MA (1997)
7. Efron, D.: Gesture, Race and Culture. Mouton and Co., Netherlands (1972)
8. Ellyson, S.L., Dovidio, J.F., Fehr, B.J.: Visual behavior and dominance in women and men. In: Mayo, C., Henely, N.M. (eds.) Gender and nonverbal behavior. Springer, Heidelberg (1981)
9. Endrass, B., Rehm, M., André, E.: Culture-specific communication management for virtual agents. In: Proceedings of AAMAS (2009)
10. Gallaher, P.E.: Individual Differences in Nonverbal Behavior: Dimensions of Style. Journal of Personality and Social Psychology 63(1), 133–145 (1992)
11. Greenbaum, P.E., Rosenfeld, H.M.: Varieties of touching in greeting: Sequential structure and sex-related differences. Journal of Nonverbal Behavior 5, 13–25 (1980)
12. Hall, E.T.: The Silent Language. Doubleday (1959)
13. Hall, E.T.: The Hidden Dimension. Doubleday (1966)
14. Hall, J.A., Friedman, G.B.: Status, gender, and nonverbal behavior: A study of structured interactions between employees of a company. Personality and Social Psychology Bulletin 25(9), 1082–1091 (1999)
15. Hofstede, G.: Cultures Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations. Sage Publications, Thousand Oaks (2001)
16. Hofstede, G.J., Pedersen, P.B., Hofstede, G.: Exploring Culture: Exercises, Stories, and Synthetic Cultures. Intercultural Press, Yarmouth (2002)
17. Johnson, C.: Gender, legitimate authority, and leader-subordinate conversations. American Sociological Review 59, 122–135 (1994)
18. Kendon, A.: Conducting Interaction: Patterns of Behavior in Focused Encounters. Cambridge Univ. Press, Cambridge (1991)
19. Knapp, M.L., Vangelisti, A.L.: Interpersonal Communication and Human Relationships. Pearson Education, Inc., Boston (1984)
20. Knoterus, J.D., Greenstein, T.N.: Status and performance characteristics in social interaction: A theory of status validation. Social Psychology Quarterly 44(4), 338–349 (1981)
21. Knudsen, M.W., Martin, J.-C., Dybkjær, L., Ayuso, M.J.M., Bernsen, N.O., Carletta, J., Heid, U., Kita, S., Llisterri, J., Pelachaud, C., Poggi, I., Reithinger, N., van Elswijk, G., Wittenburg, P.: ISLE Natural Interactivity and Multimodality Working Group Deliverable D9.1: Survey of Multimodal Coding Schemes and Best Practice (2002), `http://isle.nis.sdu.dk/reports/wp9/D9.1-7.3.2002-F.pdf` (07.02.07)
22. Leffler, A., Gillespie, D.L., Conaty, J.C.: The effects of status differentiation on nonverbal behavior. Social Psychology Quarterly 45(3), 153–161 (1982)
23. McCrae, R.R., Allik, J. (eds.): The Five-Factor Model of Personality Across Cultures. Kluwer Academics, Dordrecht (2002)
24. McNeill, D.: Hand and Mind — What Gestures Reveal about Thought. The University of Chicago Press, Chicago (1992)

25. Nazir, A., Lim, M.Y., Kriegel, M., Aylett, R., Cawsey, A., Enz, S., Zoll, C.: Culture-personality based affective model. In: Proceedings of the IUI workshop on Enculturating Conversational Interfaces, Gran Canaria (2008)
26. Quaddus, M.A., Tung, L.L.: Explaining cultural differences in decision conferencing. Communications of the ACM 45, 93–98 (2002)
27. Rehm, M., Nakano, Y., André, E., Nishida, T.: Culture-specific first meeting encounters between virtual agents. In: Prendinger, H., Lester, J.C., Ishizuka, M., et al. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 223–236. Springer, Heidelberg (2008)
28. Rehm, M., Nakano, Y., André, E., Nishida, T., Bee, N., Endrass, B., Huang, H.-H., Lipi, A.A., Wissner, M.: From Observation to Simulation — Generating Culture Specific Behavior for Interactive Systems. In: AI & Society (in press)
29. Schwartz, S.H., Sagiv, L.: Identifying culture-specifics in the content and structure of values. Journal of Cross-Cultural Psychology 26(1), 92–116 (1995)
30. Teng, J.T.C., Calhoun, K.J., Cheon, M.J., Raeburn, S., Wong, W.: 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 (1999)
31. Ting-Toomey, S.: Communicating Across Cultures. The Guilford Press, New York (1999)
32. Traum, D., Swartout, W., Marsella, S., Gratch, J.: Fight, Flight, or Negotiate: Believable Strategies for Conversing Under Crisis. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 52–64. Springer, Heidelberg (2005)
33. Triandis, H.C., Suh, E.M.: Cultural influences on personality. Annual Review of Psychology 53, 133–160 (2002)
34. Trompenaars, F., Hampden-Turner, C.: Riding the Waves of Culture. McGraw-Hill, New York (1998)
35. Ungar, S.: The effects of status and excuse on interpersonal reactions to deviant behavior. Social Psychology Quarterly 44(3), 260–263 (1981)
36. `http://www.jobware.de/ra/fue/vhs/41.html` (last visited: November 25, 2008)
37. Warwick, T.: Function analysis for team problem solving. In: SAVE Annual Proceedings (1994)