The CUBE-G approach – Coaching culture-specific nonverbal behavior by virtual agents

Matthias Rehm, Elisabeth André, Nikolaus Bee, Birgit Endrass, Michael Wissner
Multimedia Concepts and Applications, University of Augsburg, Germany
Yukiko Nakano
Computer, Information, and Communication Sciences, Tokyo University of Agriculture and Technology, Japan
Toyoaki Nishida, Hung-Hsuan Huang
Intelligence Science and Technology, Kyoto University, Japan

Keywords: cross-cultural communication; virtual agent; role-playing

Abstract

Embodied conversational agents have been proven to be powerful tools for engaging users in interactions and thus are suitable for training scenarios that rely on a role-playing metaphor. In the CUBE-G project we propose an approach for culture-adaptive behavior generation of such agents, which can be employed in edutainment applications for increasing cultural awareness and for learning some of the appropriate behavior routines. In this paper we present the methodological approach of a standardized collection of multimodal behavioral corpora for different cultures to inform a parametrized model of cultural behavior generation.

Introduction

Imagine you are in Japan for the first time in your life. You looked at the first chapter of a Japanese language text book to learn some phrases beforehand. Now you know how to greet someone you meet for the first time:

A: Kon'nichi wa.
B: Kon'nichi wa.
A: Watashi wa Sakiko Honda desu. Hajimemashite.
B: Watashi wa Jeson Miraa desu. Hajimemashite. Doozo yoroshiku.

Although you know the phrases, you are still feeling a bit uncomfortable because it is not only the language that is different but also the nonverbal interaction habits. The language text book could not prepare you for the actual situational context. How do you behave in this situation? Do you shake hands? Where do you look? How close do you get to your conversational partner?
Embodied conversational agents (e.g. Cassell et al., 2000) have been proven to be powerful tools for engaging users in interactions (e.g. André and Rehm, 2003) and thus are suitable for training scenarios that rely on a role-playing metaphor (Core et al., 2006; Hubal et al., 2000; Watson et al., 2007). We propose an approach for culture-adaptive behavior generation of such agents, which can be employed in edutainment applications for increasing cultural awareness and for learning some of the appropriate behavior routines.

The CUBE-G Project

CUBE-G (**CU**lture-adaptive **BE**havior **G**eneration for interactions with embodied conversational agents) starts with a focus on the German and the Japanese culture and relies on a combination of a data-driven and a model-driven approach.

Data-driven approach

Data-driven in this context means collecting and analyzing video corpora of real interactions. The corpus-based approach is needed in order to find correlations between the cultural dimensions and communicative behaviors that cannot be directly inferred from the model and to validate the model. Corpus-based work allows to inform the verbal and nonverbal behavior of an embodied agent in an empirically sound way keeping the intuition of the researcher at bay, which otherwise is often the main source of information (Rehm and André, 2005). In this project, the analysis will focus on the nonverbal aspects of communication which has been neglected so far in comparison to verbal communication. But of course a multimodal corpus allows for unlimited analysis of aspects of face-to-face communication. The added value of the corpus collected in this project is the comparability of identical conversational interactions in two distinct cultures presenting a rich source of empirical data.

Model-driven approach

It is not feasible to manually specify culture-specific behaviors for all kinds of situations the agent may face. Thus, the data-driven approach is complemented by a model-driven approach, which utilizes Hofstede's theory of cultural dimensions (Hofstede, 2001). Hofstede defines culture along five dimensions:

1. Power distance: Power distance deals among other things, with superior's decision-making styles and with the decision-making style that subordinates prefer in their boss. Hofstede concludes that more coercive and referent power is used in high-H societies and more reward, legitimate, and expert power in low-H societies.
2. Individualism vs. collectivism: On the individualist side we find societies in which the ties between individuals are loose: everyone is expected to look after him/herself. On the collectivist side, we find societies in which people are integrated into strong, cohesive in-groups, often extended families which continue protecting them in exchange for unquestioning loyalty.
3. Masculinity vs. femininity: "refers to the distribution of roles between the genders." Hofstede's studies revealed, that women's values differ less among societies than men's values, and that men's values from one country to another contain a dimension from very assertive and competitive and maximally different from women's values on the one side, to modest and caring and similar to women's values on the other.
4. Uncertainty avoidance: It indicates to what extent a culture programs its members to feel either uncomfortable or comfortable in unstructured situations. Unstructured situations are novel, unknown, surprising, or different from usual.
5. Long-term vs. short-term orientation: Values associated with long term orientation are thrift and perseverance wheras values associated with short term orientation are respect for tradition, fulfilling social obligations, and protecting one's face.

What makes Hofstede's theory especially appealing is the fact that he shows tight correlations between verbal and nonverbal behavior and the proposed positions on the cultural dimensions. The predictive power of his model combined with a thorough analysis of specific communicative settings allows for developing a parameterized model of culture-specific behavior generation.

Towards cross-cultural embodied conversational agents

Whereas static presentations like e.g., web sites can be easily tailored to culture-specific demands during the design process (given that the designer recognizes the challenge), dynamic generations of multimodal presentations of information cannot so easily be dealt with, because they are tailored on the fly depending on situational and contextual factors. To make these dynamic presentations sensible to cultural differences, one needs a set of parameters or rules that allow for influencing the generation process in the same way as the situational and contextual factors. Embodied conversational agents (ECAs) can be regarded as a special case of multimodal dynamic interaction systems. They promote the idea that humans, rather than interacting with tools prefer to interact with an artefact that possesses some human-like. If it is true as Reeves and Nass' Media Equation suggests that people respond to computers as if they were humans (Reeves & Nass, 1996), then there are good chances that people are also willing to form social relationships with virtual personalities. As a consequence, it seems inevitable to take cultural aspects into account when creating such agents.
Rosis, Pelachaud, and Poggi (2004) illustrate this problem in a convincing manner. Their survey of the Microsoft Agents web site shows, that the look as well as the animations of the characters are all based on western cultural norms. They only found four non-western style agents, which moreover exhibited only a reduced set of animations. Sengers (2004) emphasizes this problem of a "McDonaldization" of agents, if culture-specific aspects are disregarded in the design and behavioral modelling of agents. Apart from imposing western cultural standards on all users, the danger lies in a very low acceptance of such agents by users with different cultural backgrounds. This fact can be attributed to such fuzzy aspects like globalisation but as Nass and colleagues have shown, the cultural background and behavioral consistency of an agent matter. In one of their studies, Korean subjects were confronted with either an American or a Korean agent. The subjects trusted the agent which corresponded to their own cultural identity more. Thus, in an e-commerce scenario, e.g., the appropriate agent should lead to more successful transactions. In another study, Takeuchi, Katagiri, Nass, and Fogg (1998) examine cultural effects for reciprocal behavior of American and Japanese subjects. In a variation of the desert survival task, they showed how subjects reacted towards a computer that helped them on the task. American subjects showed reciprocal behavior, if the same computer needed their help on a later task, the Japanese subjects on the other hand showed reciprocal behavior if the computer was a "member" of the same group as the previous computer. This group relation had to be established explicitly to invoke reciprocal behavior at all in the Japanese subjects.
Allbeck and Badler (2004) review culture-specific nonverbal communicative behavior taking into account facial expressions, gestures, body movements, posture, visual orientation (eye contact), physical contacts (handshakes, patting), spatial behavior (proximity, distance, positions), appearance, and nonverbal vocalizations. From a technical point of view, the problem arises of how to ensure consistency between verbal and nonverbal communicative behaviors. An agent that just stares at the interaction partner and does not show any appropriate eye movements or gaze behavior will create an awkward atmosphere which may well lead to a failure of the interaction. To prevent such failures of communication and make agents believable and consistent in their behavior, the EMOTE model by Allbeck and Badler seems to provide a promising starting point since it enables the generation of several variants for the same basic animation data depending on the settings of parameters, such as effort and shape. Even though they present a couple of interesting ideas regarding

the use of the EMOTE model to capture cultural aspects, there is not yet any implementation of the approach so far.
Noot and Ruttkay (2005) define a specific markup language called GESTYLE which allows the user to vary an ECA's style both for verbal and nonverbal modalities. The style specifies when and how an agent uses certain gestures and how the speech is modulated. An important component of their approach is a style dictionary which guides the choice of appropriate behaviors. In contrast to them, Martin et al. (2005) employ a copy-synthesis approach to specify expressivity dimensions for an embodied conversational agent. Based on annotated video recordings of human speakers, they manually define markups augmented by expressivity parameters which are then forwarded to an animation engine to generate individual behaviors. Johnson et al. (2004) describe a language tutoring system that also takes cultural differences in gesture usage into account. The users are confronted with some prototypical settings and apart from speech input, have to select gestures for their avatars, and have to interpret the gestures by the tutor agents to solve their tasks.
While the generation of individual behaviors is considered as an important prerequisite to realize culture-specific behaviors, none of the approaches so far is able to automatically extract the relevant parameters from a corpus. Even though there are a number of approaches to simulate culture-specific agents, a principled approach to the generation of cross-cultural behaviors is still missing. Furthermore, there is no empirical validated approach that maps cultural dimensions onto expressivity dimensions. In order to realize cross-cultural agents, we need to move away from generic behavior models and instead simulate individualized agents that portray idiosyncratic behaviors, taking into account the agent's cultural background.

Standardized corpus collection for prototypical cross-cultural situations

CUBE-G focuses on three social situations that are prototypical for cross-cultural encounters. They serve as a background for the corpus collection as well as for the system development (see Fig. 1).

1. Meeting someone for the first time: The user has to join a group of agents and get acquainted with them. This is a variation of the standard first chapter of every language textbook.
2. Negotiation: This is another prototypical situation where the user has to negotiate to reach a state which is satisfactory for both sides.
3. Conversation with high status individual: The user has to interact with an interlocutor of a higher social status. This can be indicated by the age or other attributions to the interaction partner.

Dyadic interactions between human subjects were recorded in the three scenarios mentioned above. Table 1 gives an overview of the design. One of the interaction partners in each scenario was an actor following a script for the specific situation. To control for gender effects, a male and a female actor is employed in each scenario interacting with the same number of male and female subjects. Thus, apart from the two male (MA1, MA2) and two female actors (FA1, FA2), ten male (MS1-MS10) and ten female subjects (FS1-FS10) were needed for this corpus study. The same design was used in Germany as well as in Japan. Subjects were told that they take part in a study by a well-known consulting company for the automobile industry which takes place at the same time in different countries. To attract their interest in the study, a monetary reward was granted depending on the outcome of the negotiation.

Figure 1: German actors during rehearsal for the three prototypical situations.

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

Table 1: Experimental design for the corpus collection.

After having met the student 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. The negotiation task is based on the standard "Lost at sea" scenario. Subjects are told that they survived a shipwreck and can choose three items out of 20 which they can take with them on the lifeboat. Each participant (subject and actor) has ten minutes to decide for three items. Then they start negotiating to come up with a single three item list ranked according to the relevance for survival. Because one of the participants is an actor, we could control that they initially agreed only on one item. Participants negotiated as long as it took to come to a conclusion which was between 8 and 12 minutes. Afterwards, they were debriefed by the high status actor who checked their list against the "official" list of the U.S. Navy. Debriefing was done individually, starting with the subject. The student actor was sent out of the room. The high status actor followed a script which ensured that only one item was ranked correctly, one was under the first three top items and one was completely wrong. Thus, we were able to first create a positive atmosphere and then to elicit whether students accept what the high status person said or whether they start arguing about the "official" list. At the end, each subject received a monetary reward of 15 Euro. Around ten hours of material were produced each for the Japanese and the German condition.

Deriving behavioral information for a parametrized model of culture-specific behavior

Annotation of the corpora is done in Anvil (Kipp, 2003). To allow for an efficient and reliable analysis, the recording setup was identical in Germany and in Japan. The setup consists of two video cameras, a webcam and a microphone.

- Gaze: Information about head pose allows estimating the gaze direction of the user during the interaction in realtime. Four different directions are distinguished: face of interlocutor, body of interlocutor, hands of interlocutor, elsewhere.
- Proxemics: A webcam captured the interaction space of the two participants allowing us to analyse their proxemics behavior. The proxemics information is

coded in categories corresponding to Hall's suggestions, i.e. intimate, personal, social, and public, each with two subcategories near and far.
- Sound: The microphone recorded the subject's utterances. The stream is used to categorize the volume of the subject's verbal interaction in three categories (soft, normal, loud).
- Gesture use: The video cameras focus the subject and the actor. The recordings are employed to analyse the use of gestures by the subject. The annotation scheme primarily focuses on gestural expressive features like speed or spatial extent (Pelachaud, 2005).
- Speech acts: A simplified DAMSL-scheme is employed to code speech acts (Core & Allen, 1997). The information gained from this annotation will inform the dialogue manager of the envisioned application.

Parameterizing the cultural dependent factors in social communications

The annotation of the multimodal corpus serves three different functions.

1. Statistical information about the time dependent nonverbal behavior is derived directly from the annotation and informs the generation of culture-specific i.e., German or Japanese, gaze, expressive, and proxemic behavior as well as volume of the speech synthesis.
2. Comparing the usage of nonverbal behaviors between German and Japanese among the different conversational situations will allow us to extract culture-specific behavior patterns like differences in volume or intensity of gestures that can be used to establish new parameters for the cross-cultural generation model allowing us to ground the model-driven in empirical data.
3. The annotation of human-human interactions serves as a benchmark for the system evaluation. Users are confronted with the same social situations as in the corpus study but the actor is replaced by an embodied conversational agent reacting to the user. Recording and analysing this interaction allows for a comparison between the human-human and the human-agent condition.

Role-playing with virtual agents

Isbister (2004) has convincingly argued for the use of agents to further cross-cultural communication skills between users. Although agents allow experiencing nonverbal behavior, there is no danger of social embarrassement when the user makes mistakes in his exercises. Agents have the additional advantage of not getting tired by repetitive exercises and of being able of replicating specific behavior without much deviation. One and the same agent can be used to exemplify the behavior of different cultures ("culture-hopping") making it possible to contrast the behavior of two cultures and point out the differences. And of course the agent enables personal feedback and even can be used to contrast the learner's behavior with the target culture's behavior if the learner's behavior has been tracked. It has to be noted that a role-playing system that features virtual agents cannot replace live experience but it presents an efficient and economical way of making such experience-based learing available to many users and can at least if nothing more be a beneficent addition to the classical language textbook training.
According to Hofstede (1991), learning cross-cultural communication skills always passes through three steps: awareness, knowledge, and skills. First the user must be aware that there is a difference in behavior and that this is not concerned with better or worse but just with different. Knowledge about behavior patterns has to follow, which can be interpreted as getting an intellectual grasp on where and how

one's own behavior differs. Skills at last is concerned with training specific behaviors that allow to get along better in the different culture like gaze behavior during negotiations.
The CUBE-G system concentrates on the following nonverbal communicative behaviors: proxemics, eye gaze, sound, gesture use. Users will enter the virtual meeting space (Rehm, André, and Nischt, 2005), where agents are able to show culture-specific behavior, and react appropriately. For instance, the user might have the task to join a group of agents and start a conversation with them. He will use his cultural patterns to do so, which might be not appropriate for the current group. The agents react accordingly in showing signs of embarrassment for example. The users nonverbal behavior is analysed and one the agents will be used to contrast the user's behaviors with the culturally appropriate ones. Speaking in a low voice e.g. might result in being ignored by a group of agents from a culture situated at the masculine end of the gender dimension. Eye gaze is another crucial feature of cross-cultural encounters because too much of it might be understood as impolite staring by those who use is sparingly whereas to few of it might be understood as disinterest by those who use it extensively. Following Hofstede (1991), the agents will ultimately have three different tasks in the coaching scenario.

1. Confronting: The agent or a group of agents confronts the user with a different cultural group. The confronting task is a test situation for the user where he gets feedback that increases his awareness about culture-specific behavior differences. The user's nonverbal interaction behavior might e.g. be considered rude by the group of agents. Thus, the user gets feedback that some of his behaviors might be interpreted as offensive in the given culture. Or vice versa, the user might experience the agents' behavior as offensive.
2. Contrasting: The agent or a group of agents contrast different cultural dependent behaviors for the user allowing to compare these behavior patterns. The contrasting task thus visualizes and illustrates potential communication problems to the user and increases the user's knowledge about specific behavior routines that differ between his own and the target culture.
3. Explicating: The agent or a group of agents explicitely focus on a specific nonverbal behavior. The explicating task allows the user to train specific aspects of cultural dependent nonverbal behaviors, e.g. frequency of gaze towards the interlocutor and thus increases his skills of nonverbal behavior in the target culture.

Conclusion

In this paper, we presented the project CUBE-G, which aims at generating nonverbal behavior for virtual agents based on their cultural background. This is a prerequisite to develop a virtual learning environment that allows users to experience culturally determined differences in communicative behaviors and which can ultimately serve as a training device for increasing cultural awareness, imparting cultural knowledge and training culture-specific behavior routines. To this end, an empirical data-driven approach is combined with a theoretical model-driven approach. So far, a standardized multimodal behavior corpora of prototypical social interactions has been collected in different cultures (Germany and Japan) and is analyzed at the moment to ground the theoretical model in empirical data.

Acknowledgment

The work described in this article is funded by the German Research Foundation (DFG) under research grant RE 2619/2-1.

List of references

Allbeck, J.M., Badler, N.I. (2004). Creating Embodied Agents With Cultural Context. In: Payr, S., Trappl, R. *Agent Culture: Human-Agent Interaction in a Multicultural World.* 107—126, Lawrence Erlbaum Associates.

André, E., Rehm, M. (2003): Künstliche Intelligenz, Special Issue on Embodied Conversational Agents, No.4.

Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (2000): *Embodied Conversational Agents.* MIT Press: Cambridge.

Core, M., Allan, J. (1997). Coding Dialogs with the DAMSL Annotation Scheme. Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines.

Core, M., Traum, D., Lane, H.C., Swartout, W., Gratch, J., Lent, M. van, Marsella, S. (2006). Teaching Negotiation Skills through Practice and Reflection with Virtual Humans. *SIMULATION*, Vol. 82, No. 11, 685-701.

Hofstede, G. (1991). Cultures and Organisations – Intercultural Cooperation and its Importance for Survival, Software of the Mind. Profile Books.

Hofstede, G. (2001). *Cultures Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations.* Sage Publications.

Hubal, R.C., Kizakevich, P.N., Guinn, C.I., Merino, K.D., West, S.L. (2000). The Virtual Standardized Patient – Simulated Patient-Practitioner Dialog for Patient Interview Training. In *Envisioning Healing: Interactive Technology and the Patient-Practitioner Dialogue*, 133-138. IOS Press: Amsterdam.

Isbister, K. (2004). Building Bridges Through the Unspoken: Embodied Agents to Facilitate Intercultural Communication. In: Payr, S., Trappl, R. Agent Culture: Human-Agent Interaction in a Multicultural World, Lawrence Erlbaum Associates: London.

Johnson, W.L., Choi, S., Marsella, S., Mote, N., Narayanan, S., Vilhjàlmsson, H. (2004). Tactical Language Training System: Supporting the Rapid Acquisition of Foreign Language and Cultural Skills. Proceedings of InSTIL/ICALL – NLP and Speech Technologies in Advanced Language Learning Systems.

Kipp, M. (2003). *Gesture Generation by Imitation – From Human Behavior to Computer Character Animation.*Universität des Saarlandes, PhD. Thesis.

Martin, J.C., Abrilian, S., Devillers, L., Lamolle, M, Mancini, M., Pelachaud, C. (2005). Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs. In: Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., Rist, T. *Intelligent Virtual Agents (IVA)*, 405—417, Springer.

Noot, H., Ruttkay, Z. (2005). Variations in Gesturing and Speech by GESTYLE. International Journal of Human-Computer Studies, Special Issue on Subtle Expressivity for Characters and Robots.

Pelachaud, C. (2005). Multimodal Expressive Embodied Conversational Agents. Proceedings of ACM Multimedia.

Reeves, B., Nass, C. (1996). *The Media Equation – How People Treat Computers, Television, and New Media Like Real People and Places.* Cambridge University Press

Rehm, M., André, E., Nischt, M. (2005). Let's Come Together – Social Navigation Behaviors of Virtual and Real Humans. In: Maybury, M., Stock, O., Wahlster, W. Intelligent Technologies for Interactive Entertainment, Springer: Berlin, Heidelberg.

Rehm, M., André, E. (2007). From Annotated Multimodal Corpora to Simulated Human-Like Behaviors. In: Wachsmuth, I., Knoblich, G. Titel. Springer: Berlin, Heidelberg.

Rosis, F. de, Pelachaud, C., Poggi, I. (2004). Transcultural Believability in Embodied Agents: A Matter of Consistent Adaptation. In: Payr, S., Trappl, R. *Agent Culture: Human-Agent Interaction in a Multicultural World.* 75—106, Lawrence Erlbaum Associates.

Sengers, P. (2004). The Agents of McDonaldization. In: Payr, S., Trappl, R. *Agent Culture: Human-Agent Interaction in a Multicultural World.* 3—20, Lawrence Erlbaum Associates.

Takeuchi, Y., Katagiri, Y., Nass, C.I., Fogg, B.J. (1998). Social Response and Cultural Dependency in Human-Computer Interaction. Proceedings of PRICAI, 114—123.

Watson, S., Vannini, N., Davis, M., Woods, S., Hall, M., Dautenhahn, K. (2007). FearNot! An Anti-Bullying Intervention: Evaluation of an Interactive Virtual Learning Environment. Proceedings of AISB 2007.