Generating Culture-Specific Gestures for Virtual Agent Dialogs

Birgit Endrass[1], Ionut Damian[1], Peter Huber[1],
Matthias Rehm[2], and Elisabeth André[1]

[1] Multimedia Concepts and Applications, Augsburg University,
Universitätsstr. 6a, D-86159 Augsburg, Germany
{endrass,andre}@informatik.uni-augsburg.de
http://mm-werkstatt.informatik.uni-augsburg.de
[2] Department of Media Technology, Aalborg University,
Niels-Jernes Vej 14, DK-9220 Aalborg, Denmark
matthias@imi.aau.dk

Abstract. Integrating culture into the behavioral model of virtual agents has come into focus lately. When investigating verbal aspects of behavior, nonverbal behaviors are desirably added automatically, driven by the speech-act. In this paper, we present a corpus driven approach of generating gestures in a culture-specific way that accompany agent dialogs. The frequency of gestures and gesture-types, the correlation of gesture-types and speech-acts as well as the expressivity of gestures have been analyzed in the two cultures of Germany and Japan and integrated into a demonstrator.

1 Motivation

Virtual agents are used in a vast variety of applications. However, many researchers are only interested in certain aspects of behavior. In our interactive storytelling system [1], for example, we investigate dialog generation. But when focusing on verbal behavior, nonverbal behavior cannot be left aside. According to Kita [2], they are tightly linked systems, where "*the link is strong enough that speech-accompanying gestures do not disappear even when the addressee does not have a visual access to the gestures (e.g., on the telephone)*". Selecting gestures appropriate to a virtual agent dialog, however, can be a time consuming task. Thus, we aim at generating gestures automatically. In human conversations, gestures are not performed randomly and are not just a decorative feature. Often they serve a function, such as supporting a speech-act.

Verbal and nonverbal behavior of virtual agents became more sophisticated in recent years and social factors such as personality or culture came into focus. In this paper, we present an approach of generating gestures for virtual agents in a culture-specific way, driven by the speech-act generation of the system. To this end, we recorded a video corpus in the two cultures of Germany and Japan and integrated our findings into a multiagent system.

2 Related Work

Several approaches have focused on the challenge of automatically generating nonverbal behaviors. The most well known system, BEAT, was presented by Cassell et al. [3]. As input, it receives plain text and generates synchronized nonverbal behavior for a virtual character. In their work, the authors describe behavior selection according to filter functions that regulate how much nonverbal behavior is performed. Such filters can reflect the personality, affective state or energy level of an agent. We consider these filters as an inspiration for our work and therefore suggest culture as an aspect that effects the selection of nonverbal behaviors.

A nonverbal behavior generator that generates BML scripts containing nonverbal behaviors for a given input text is introduced in [4]. Nonverbal behavior is generated based on rules that were extracted from a set of video clips. Similar to in the work described in this paper, speech-utterances have been labeled by the authors and their co-occurrences with nonverbal behaviors have been analyzed. They focused on head movements, facial expressions and body gestures. However, they did not analyze different cultures in their approach.

Bergman and Kopp [5] introduce a system that generates iconic gestures to express spatial information. A corpus containing landmark descriptions was recorded and annotated for their purposes and a prototype that performs iconic gestures has been developed and evaluated. The authors state that the performance of iconic gestures varies across speakers. It would be interesting whether there are differences aroused by cultural background as well.

In [6], Ruttkay describes a markup language, where different aspects of styles are defined in a dictionary of meaning-to-gesture mappings. The style dictionary suggests appropriate verbal and nonverbal behaviors. Culture specific styles could be considered as well.

Integrating culture into the behavioral model of virtual agents has come into focus lately. Most other work either focuses on abstract cultures, is not very specific in modeling differences in certain aspects of behavior or is not corpus driven. In [7], for example, an educational application for inter-cultural empathy is introduced. To achieve cultural awareness, a group of users interacts as a team with a group of virtual agents. However, in their system no awareness for an existing culture or culture-specific behavior is trained, but an overall awareness of something that is different from one's own culture.

The tactical language training system [8] explores cultural differences in gesture usage. Users have to select gestures for their avatars along with speech input. In addition, they have to interpret the gestures made by other agents appropriately in order to solve their tasks. Another system that demonstrates cultural differences is presented in [9]. A group of characters performs differently in a situation, depending on their cultural background. However, abstract concepts of culture are used rather than differences extracted form a corpus.

We consider the automatically generation of gestures to agent dialogs along with a corpus driven approach to simulate culture-specific differences in behavior, as the new contribution of our work.

3 Theoretical Background

In order to generate gestures that accompany virtual agent dialogs in a culture-specific way, we have to explore the concepts of gestures, dialog-acts and different cultures.

McNeill [10] has introduced the most well known classification of gestures into gesture-types: *Deictic* gestures are pointing or indicative gestures. *Beat* gestures are rhythmic gestures that follow speech prosody. *Emblems* have a conventionalized meaning and do not need to be accompanied by speech. *Iconic* gestures explain the semantic content of speech. *Metaphoric* gestures accompany the semantic content of speech in an abstract manner by the use of metaphors. *Adaptors* are hand movements towards other parts of the body. In addition, McNeil [10] explores the temporal course of gestures according to the phases: preparation, hold, stroke and retraction. In the preparation phase, the hands are brought into the gesture space. A hold might occur when the gesture is not aligned with the corresponding utterance yet. The stroke phase carries the content of the gesture and can be categorized by the gesture-types described above. In the retraction phase, the hands are finally brought back into a resting position. Annotating gestures as suggested by McNeill [10] is already successfully used in behavior generation for virtual agents (see [11], [3] or [12] for examples). The dynamic variation of a gesture is another aspect to be considered. In [13], Pelachaud describes six parameters that characterize a gesture's expressivity, which can depend on individual factors, such as personality or emotional state. The *spatial extent* describes the arm's extent toward the torso. The *speed* and the *power* of a gesture can vary as well. The *fluidity* describes the continuity between consecutive gestures, while the *repetitivity* holds information about the repetition of the stroke. The *overall activation* explains the frequency of gestures.

To categorize dialog-acts, we use the annotation schema DAMSL (Dialog Act Markup in Several Layers) that was introduced by Core and Allen [14]. One layer of the schema, the communicative function, serves our purposes very well as it labels the communicative meaning of a speech-act. For the work described in this paper, we use the following subset of communicative functions: statement, info-request, influence on future, agreement/disagreement (indicates the speaker's point of view), hold, understanding/misunderstanding (without stating a point of view) and answer.

To generate nonverbal behaviors for prototypical German and Japanese agents, we need to distinguish these two cultures. Ting-Toomey [15] distinguishes high- and low-context communication cultures. In high-context communication little is encoded explicitly and the conversation relies mainly on physical context. Messages and symbols might seem relatively simple but contain a deep meaning. In contrast, low-context communication explicitly codes information; symbols and messages are direct and to the point. In [15], Germany is mentioned as one of the most extreme low-context cultures, while Japan is named to be on the extreme high-context side. We expect a more frequent use of direct gestures (deictic and iconic) in low-context cultures. Vice versa, we expect more metaphoric gestures in Japanese conversations.

4 Empirical Verification

In order to find statistical tendencies that describe what gesture-types are commonly used for which dialog-utterance, we analyzed the video corpus recorded for the CUBE-G project [16]. In total, more than 40 students from Germany and Japan participated in the study, where around 25 minutes of video data were recorded for each subject. To ensure a high control over the recordings, subjects interacted with actors whom they did not know in advance. At the beginning of the experiment, participants were asked to get acquainted with each other as a preparation for the task they had to solve later. During this time recording already started. For the work described in this paper, we analyzed this Small Talk scenario that lasted for approximately five minutes for each subject. For more information on the recordings, please see [16].

4.1 Quantitative Analysis

Using the Anvil tool [17], verbal and nonverbal behavior was annotated according to the subset of DAMSL dialog-utterances, McNeill's classification of gestures and the gestural expressivity parameters (see Section 3). So far, the videos of 21 German and 7 Japanese subjects were considered. As we focused on gestures that accompany speech and that are of a general nature, we did not consider adaptors and emblems yet.

Table 1 shows the frequencies of dialog-utterances per minute in the two cultures of Germany and Japan averaged over the number of subjects (left), as well as the probability that a gesture is performed during a given utterance (right). During the dialog-utterances "hold" and "influence on future" rarely any gestures occurred. We thus do not consider them for our model. Our analysis revealed that there are significantly more info-requests in the Japanese videos than in the German ones (with a p-value ≤ 0.003 using the two sided t-test). Interestingly, there are also more gestures occurring during info-requests in the Japanese corpus ($p \leq 0.075$). Regarding the frequency of understanding/misunderstanding utterances, we found significantly more of these dialog-acts in the Japanese conversations ($p \leq 0.02$). This is in line with expectations about the two cultures: giving verbal feedback without stating a personal opinion is supposed to be very

Table 1. Average occurrence of dialog-utterances per minute (left) and probabilities that a gesture occurs during the utterance (right) in German and Japanese videos

utterance per minute	Germany	Japan
info-request	1.10	2.09
answer	2.95	2.20
statement	4.96	5.74
agreement/ disagreement	0.87	0.74
understanding/ misunderstanding	1.64	2.74

gesture co-occurrence	Germany	Japan
info-request	**5**%	**11**%
answer	10%	10%
statement	24%	15%
agreement/ disagreement	2%	4%
understanding/ misunderstanding	0%	0%

Table 2. Probabilities that a certain gesture accompanies an utterance in the two cultures of Germany and Japan

	Germany				Japan			
utterance/gesture	beat	deictic	iconic	metaphoric	beat	deictic	iconic	metaphoric
info-request	0%	**67%**	**33%**	0%	0%	**75%**	0%	**25%**
answer	25%	25%	21%	29%	25%	12.5%	37.5%	25%
statement	27%	17%	27%	29%	16%	26%	35%	23%
agreement/disagreement	50%	50%	0%	0%	0%	100%	0%	0%

common in the Japanese culture, while stating an opinion is more common in Western cultures.

As a next step, we explored the gesture-types that accompany dialog-utterances. Table 2 shows the probabilities for a gesture-type during an utterance given that a gesture occurs. Our analysis revealed differences in the usage of gestures that occur with info-requests. Japanese subjects showed significantly more deictic and metaphoric gestures during info-requests than German subjects (both with p-values ≤ 0.01).

4.2 Qualitative Analysis

Besides the choice of gesture-type, the way a gesture is performed differs across cultures, too. A German deictic gesture, for example, is usually executed using the index finger for pointing, while in a typical Japanese deictic gesture the whole flattened hand is used. For other gestures, differences are not as simple to distinguish. To simulate these differences, we modeled different animations for the two cultures. Figure 1 shows two iconic gestures in Germany (1) and Japan (2). Animations for virtual agents (a) are presented next to the video-samples from our corpus (b), where the gestures were extracted from.

Our analysis of gestural expressivity revealed significant differences for all parameters. German subjects repeat gestures less, have more fluid motions, gesture more powerfully and faster and use more space in gesturing than Japanese subjects (see [16] for more details).

Fig. 1. Examples for iconic gestures in Germany (1b) and Japan (2b), imitated by virtual agents (a)

5 Integration into a Virtual Scenario

As a simulation platform, we are using the Virtual Beergarden scenario (see [18] for technical details). The process of action selection is realized by a hierarchical planning system. Verbal behavior is generated from a knowledge base and sent to the text-to-speech component. Nonverbal behavior is added, considering the agent's cultural background, taking into account the following questions: (1) Should the speech act be accompanied by a gesture? (2) Which gesture-type should be selected? (3) Are there culture-specific restrictions on the execution?

For the first decision, Table 1 (right) is used as a basis. If a gesture should be performed, the gesture-type is selected according to the distribution presented in Table 2. Finally, the animation is selected.

Gestures are stored in a nonverbal knowledge base inspired by [12]. Following their approach, we are using an XML structure that comprises a form, a function and restrictions for each gesture. In our version, a gesture can either be culture-specific or not. Culture-specific gestures can only be performed by agents of the specified cultural background. General gestures, e.g. a simple beat gesture, can be exhibited by every agent. The performance of these gestures, however, is realized in a culture-specific way, taking into account the expressivity parameters. Therefore every gesture is divided into phases: preparation, stroke and retraction. Preparation and retraction phases are used for animation blending. A gesture could, for example, be chosen while the agent does not stand in a neutral position. In this case, the preparation phase is used to blend into the gesture space. In the stroke phase, the actual gesture is performed. It can be customized to match different gestural expressivities. The parameter repetition, for example, can be varied by playing the stroke phase several times, while it can be played faster or slower to customize the speed parameter.

6 Conclusion

We recorded a video corpus in the two cultures of Germany and Japan where speech-acts and gestures as well as their correlation were annotated and analyzed. Findings were integrated into a demonstrator. The contribution of this work is to automatically generate gestures for virtual agents in a culture-specific way. By that means, the process of gesture selection is speech-act as well as corpus driven. Although we consider this integration as an important step towards enculturating our virtual agents, there is still a long way to go. Other nonverbal behaviors, e.g. head nods, have not been considered yet. We found differences in the usage of understanding utterances in our corpus. It would be interesting to know how these utterances correlate with head-nods.

Acknowledgments. The first author was supported by a grant from the Elitenetzwerk Bayern (Elite Network Bavaria). This work was also partly funded by the European Commission under grant agreement IRIS (FP7-ICT-231824).

References

1. Endrass, B., Rehm, M., André, E.: What Would You Do in their Shoes? Experiencing Different Perspectives in an Interactive Drama for Multiple Users. In: Iurgel, I.A., Zagalo, N., Petta, P. (eds.) ICIDS 2009. LNCS, vol. 5915, pp. 258–268. Springer, Heidelberg (2009)
2. Kita, S.: Corss-cultural variation of speech-accompanying gesture: A review. Language and Cognitive Process 24(2), 145–167 (2009)
3. Cassell, J., Vilhálmsson, H., Bickmore, T.: BEAT: The Behaviour Expression Animation Toolkit. In: SIGGRAPH 2001, pp. 477–486. ACM, New York (2001)
4. Lee, J., Marsella, S.: Nonverbal Behavior Generator for Embodied Conversational Agents. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 243–255. Springer, Heidelberg (2006)
5. Bergmann, K., Kopp, S.: Bayesian Decision Networks for Iconic Gesture Generation. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 76–89. Springer, Heidelberg (2009)
6. Ruttkay, Z.: Presenting in Style by Virtual Humans. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 22–36. Springer, Heidelberg (2007)
7. Aylett, R., Paiva, A., Vannini, N., Enz, S., André, E., Hall, L.: But that was in another country: agents and intercultural empathy. In: Decker, S., Sierra, C. (eds.) AAMAS 2009, pp. 329–336. ACM, New York (2009)
8. Johnson, W.L., Choi, S., Marsella, S., Mote, N., Narayanan, S., Vilhjálmsson., H.: Tactical Language Training System: Supporting the Rapid Acquisition of Foreign Language and Cultural Skills. In: InSTIL/ICALL 2004 (2004)
9. Mascarenhas, S., Dias, J., Afonso, N., Enz, S., Paiva, A.: Using rituals to express cultural differences in synthetic characters. In: Decker, Sichman, Sierra, Castelfranchi (eds.) AAMAS 2009, pp. 305–312. ACM, New York (2009)
10. McNeill, D.: Hand and Mind – What Gestures Reveal about Thought. University of Chicago Press, Chicago (1992)
11. 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 & Society 24(3), 267–280 (2009)
12. Krenn, B., Pirker, H.: Defining the Gesticon: Language and Gesture Coordination for Interacting Embodied Agents. In: AISB 2004, pp. 107–115 (2004)
13. Pelachaud, C.: Multimodal expressive embodied conversational agents. In: Zhang, Chua, Steinmetz, Kankanhalli, Wilcox (eds.) ACM Multimedia, pp. 683–689. ACM, New York (2005)
14. 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, Boston, MA (1997)
15. Ting-Toomey, S.: Communicating across Cultures. The Guilford Press, New York (1999)
16. 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, Mastik (eds.) ISAGA 2007 (2007)
17. Kipp, M.: Anvil - A Generic Annotation Tool for Multimodal Dialogue. In: Dalsgaard, Lindberg, Benner, Tan (eds.) Eurospeech 2001, pp. 1367–1370 (2001)
18. Damian, I., Huber, P., Endrass, B., Bee, N.: Advanced Agent Animation. In: IVA Gala 2010 (2010)