# Evaluating the Tangible Interface and Virtual Characters in the Interactive COHIBIT Exhibit

Michael Kipp[1], Kerstin H. Kipp[2], Alassane Ndiaye[1], and Patrick Gebhard[1]

[1] DFKI, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
{kipp, ndiaye, gebhard}@dfki.de
[2] University of the Saarland, Experimental Neuropsychology Unit,
66123 Saarbruecken, Germany
k.kipp@mx.uni-saarland.de

**Abstract.** When using virtual characters in the human-computer interface the question arises of how useful this kind of interface is: whether the human user accepts, enjoys and profits from this form of interaction. Thorough system evaluations, however, are rarely done. We propose a post-questionnaire evaluation for a virtual character system that we apply to COHIBIT, an interactive museum exhibit with virtual characters. The evaluation study investigates the subjects' experiences with the exhibit with regard to informativeness, entertainment and virtual character perception. Our subjects rated the exhibit both entertaining and informative and gave it a good overall mark. We discuss the detailed results and identify useful factors to consider when building and evaluating virtual character applications.

## 1 Introduction

Virtual characters are a versatile tool for conveying information or educational content in a playful and entertaining fashion [1]. The interactive edutainment system COHIBIT[1] is a good example for a system that pursues both entertainment and education with the help of virtual characters [2]. In this system, we have an unusual condition because the user interacts with the system only by manipulating tangible bits: moving around physical building blocks from a car the user elicits reactions from the two virtual characters who, through speech, give comments, helpful advise and educational background information. Hence, there is an asymmetrical distribution of communication channels. The characters can talk but not act, the user can act but not talk (or at least talk has no effect on the characters). Various questions arise in such a scenario: Can the characters capture the user's attention at all? Is the exhibit entertaining? Does the user find the information interesting? And how are the characters perceived?

These questions are interesting for most virtual character systems, no matter what modalities are employed. However, evaluations have rarely been done in the past, all too often they were only worth a side remark in the "Future Work" section. This is unfortunate because evaluations are a good way to find

[1] COnversational Helpers in an Immersive exhiBIt with a Tangible interface.

J. Gratch et al. (Eds.): IVA 2006, LNAI 4133, pp. 434–444, 2006.

out whether your objectives were achieved or not, and to discover some of the reasons for success and failure. In iterative development, evaluation is the key factor in the development cycle. We present such an evaluation for the COHIBIT system. Subjects performed an interaction session and filled in a subsequent questionnaire that addresses all of the mentioned issues. We show how we designed and analyzed the questionaire. We try to address some issues of general interest, e.g. how do you capture a virtual character's personality? We propose 5 personality dimensions and discuss why they might be useful. We look at the structure of the questionnaire and the role that question order plays. We illuminate the role of speech synthesis, confirming findings by [3] that speech plays a major role in how well the system is perceived. We propose to analyze dialogue diversity, a core quality for Eliza-type systems, which we assume might compensate for the users' impression that a system is too talkative.

Although this paper is far from being a cookbook for systems evaluation it might give researchers who just completed their own virtual character system a starting point for their own evaluation.

## 2 Related Work

The topic of evaluation, especially for virtual character systems, has been of growing interest in the community [4]. In this section, we focus on three relevant evaluation studies.

McBreen et al. [3] conducted a study to measure the effectiveness and user acceptibility of animated agents. The domain was a multimodal e-retail application. The 36 subjects only passively watched an interaction with different set-ups (voice only, 2D/3D talking head, 2D/3D full body agent) and did not participate themselves. Questionnaires were to capture the subjects impressions. Results for voice indicated that the voice of the agent may effect the participants' attitudes towards the appearance of the agent. A finding that our results support. For capturing agent personality the authors used four dimensions: politeness, friendliness, competence, and forcefulness. There were indications that gesturing may play a role in subjects' perceptions of politeness because embodied versions were found to be more polite. Gestures also contributed to perceived friendliness. They found that forcefulness can be off-putting for the participant. They suggested to design systems where the agent can make suggestions without being too forceful. We picked up these suggestions, and found that our agents were perceived helpful, polite and friendly. As opposed to McBreen et al. we found no gender differences. Some findings correspond to our design decisions: 3D agents (as well as 3D talking heads) were found to be preferred by subjects. Also, we tried to carefully select nonverbal behaviors to maximize believability but the somewhat indifferent results in this area indicate that further research is needed to pinpoint the factors of "good" and "appropriate" gesturing.

Hartmann et al. [5] presented a system that produces expressive gestures for embodied conversational agents. They identified six attributes from psychological

literature to model expressivity. In their user studies they asked 54 subjects to rank three different animations for preference (most appropriate to least appropriate with respect to the expressive intent). The three clips were: neutral, coherent (as generated by their system) and inconsistent. The results showed that participants preferred the coherent performance above neutral and inconsistent actions. This shows that principled coherent generation of gestures is perceivable and preferred by human observers.

Van Mulken et al. [6] pointed out that many virtual character systems do not exploit the affordances of the human bodies. In our system the presence of the embodied characters is an integral part of the experience. As the set-up in Figure 1 shows, the two life-size virtual characters can be seen as performers who react to user actions, act pro-actively if the user idles and even live on and talk with each other after the user has left the exhibit in order to attract new visitors. The attention the user gives to the agents while interacting but also when watching from a distance is largely due to their quality as life-like embodied beings.

## 3 COHIBIT System and Research Questions

COHIBIT is a mixed-reality museum exhibit which features tangible interaction and conversational virtual characters [2]. Figure 1 shows the spatial arrangement: The exhibit consists of a life-size projection of two virtual characters (VCs), a table in front of it and a large shelf on the side which houses 10 real car pieces (front piece, middle piece, various rear pieces). Museum visitors entering the exhibit are detected by cameras and welcomed by the two VCs. They point out the possibility of assembling a car using the real car pieces in the shelf and offer their (verbal) assistance. If the visitor starts putting pieces on the table the VCs engage in a dialogue to motivate, guide and inform the visitor: for instance, they comment on recognized actions ("you shifted the front piece to the left"), give corrections ("you have to turn the cockpit") or suggest further action ("if you place a middle piece between cockpit and rear, you're done"). Once a car is finished the VCs congratulate the visitor and tell him/her about the model just built. As the interaction unfolds, i.e. the user builds more and different models, the system shifts focus from assisting to conveying more and deeper information about cars (security systems, environmental aspects, technical data), reflect their own technology by talking about virtual character technology (speech synthesis, computer graphics, behavior control etc.), and weave in current context knowledge like daytime, number of finished cars and even the weather. If the visitor leaves, the VCs continue living and engage in smalltalk with each other to attract further potential visitors. The complex, varied and context-aware behavior is modeled by a 70-page "screenplay" of text chunks which is traversed using a so-called sceneflow. The sceneflow is a hierarchical state-based tool for modelling complex context-aware behavior [7][8].

The mixed-reality installation provides a tangible, multimodal, and immersive experience for a single visitor or a group of visitors. The ten tangible objects

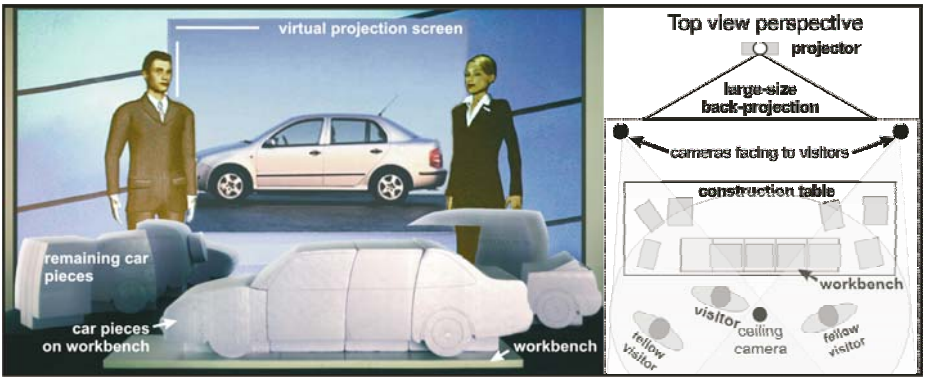


**Fig. 1.** Overview of the COHIBIT system. Left is a frontal view that shows the projection of the VCs, the car pieces and the table (workbench). Right is a top view showing projection, table, car pieces, visitors and the cameras for detecting visitor presence.

are instrumented with passive RFID tags and represent car-model pieces on the scale 1:5, the table (workbench) has five adjacent areas with RFID readers where car pieces can be placed. The back projection for the two virtual characters measures 3x3 meters. For speech output we use the commercial hi-end text-to-speech (TTS) synthesis system rVoice (Rhetorical) by Nuance. The VCs are animated with a real-time animation engine by Charamel[2] featuring 3D based keyframe animation and motion blending. The nonverbal behavior of our agents consists of a total repertoire of 28 actions (including idle-time behaviors) for each character. The gestures are in part authored and in part automatically generated from a set of rules [9].

## 4 Method

### 4.1 Participants

16 subjects (9 female) participated in the evaluation. Nine of them were 19-30 years old and the other seven 31-45 years old. All subjects were German native speakers and each subject was tested individually.

### 4.2 Procedure

The experiment consisted of letting each subject interact alone with the system for a duration of 15 minutes. Since our system also "lives on" during times when no user is present it was important to let the subject observe system behavior for some time before and after the interaction (4 minutes total).

All subjects were instructed that they would be watching the exhibit passively for some minutes before, upon a sign by the supervisor, they could "enter" the

[2] http://www.charamel.com

exhibit to interact for 15 minutes until, upon a second sign, they would watch the characters' final remarks, again passively. For the time of the interaction the subject was left alone with the system. Pre-evaluation studies showed that subjects often become nervous and self-conscious if a supervisor is present or a camera is visibly installed. Leaving subjects alone helped eliminating any kind of "examination fear" that they might have felt if being observed. We also told all subjects during instruction that it is the *system* that is being tested and not them, that they cannot make any mistakes and that they should feel free to experiment with the system.

After the experiment each subject filled in an anonymous, 2-page post-questionnaire.

### 4.3 Questionnaire

The post-questionnaire used 34 attitude statements with a 5-ary rating scale to capture how the subjects experienced the system (see Table 1 for an excerpt). The questionnaire had four major aspects: (1) general impression, (2) virtual characters, (3) dialogue, and (4) target age groups.

The questionnaire's first question was "Did you find the interaction entertaining?" aiming at a spontaneous reaction. In contrast, the questionnaire's final question asked for an overall school mark for the exhibit. The placement of the final question is meant to profit from the many previous questions that allowed the user to gain a differentiated view on her/his opinions of the system. As can be seen in Table 1, the rating for the first question (spontaneous impression) is very similar to the rating in the last question (differentiated judgment).

The "personality" of the characters was inquired using five dimensions: likability, competence, politeness, humanlikeness, talkativeness. We decided against

**Table 1.** Sample questions from the questionnaire (but in original order) with the 5-point scale and mean value over all subjects

| *question* | *scale* | *mean* |
|---|---|---|
| Did you find the interaction entertaining? | *not at all* – 1 2 3 4 5 – *very much* | 4.3 |
| Did the characters manage to get your attention? | *not at all* – 1 2 3 4 5 – *very much* | 3.9 |
| Did you find the system informative wrt. cars | *not at all* – 1 2 3 4 5 – *very much* | 3.8 |
| How did you find the dialogue variation? | *predictable* – 1 2 3 4 5 – *predictable* | 4.1 |
| The characters talked... | *too little* – 1 2 3 4 5 – *too much* | 3.4 |
| For the task the characters' comments were... | *distracting* – 1 2 3 4 5 – *helpful* | 3.9 |
| What overall mark would you give the system? | *very bad* – 1 2 3 4 5 – *very good* | 4.2 |

using the "Big Five" (openness, conscientiousness, extraversion, agreableness, and neuroticism) for two reasons. First, the range of human attributes in our system is limited. One could call our scenario a "service oriented" interaction, like a sales talk, a professional consulting session or a university lecture. In such interactions only a limited range of human behaviors occur. Extreme behaviors like screaming, crying or sulking silently are unlikely because the topics (car assembly, artificial intelligence, car research etc.) are relatively neutral and unemotional, plus the interaction protocol makes extreme behaviors a taboo. We therefore focused on those dimensions where we expect variations, "zooming in" on some of the Big Five. Thus, likability is an aspect of agreeableness, politeness may be considered a cross-product of agreeableness and conscientiousness, talkativeness is an aspect of extraversion and competence a cross-product of openness and conscientiousness. Also, for the questionaire, the notions we used should be very specific in how they are understood by the subjects. The Big Five notions might be somewhat unfamiliar to naive users. Finally, our characters are simply not human beings but artificial characters whose behavior is "designed". Therefore, we introduced the personality trait "humanlikeness" that you would not ask when rating humans.

## 5 Results

Most items of the questionnaire were unidirectional on a 5-ary scale (very bad to very good). For statistical analysis the data was transformed so that for all items the negative end of the scale was assigned "1" (e.g. not entertaining, very disturbing) and the positive end to "5" (e.g. very entertaining, not disturbing). Since rating scales can be treated as interval scales [10] we used parametrical tests for the statistical analysis. The *t-test for one sample* is a statistical significance test that proves whether a measured mean value of an observed group differs from an expected value. In our study, ratings were proven to be positive if the mean score significantly exceeded the neutral value of "3". Interaction between two factors (e.g. gender and questionnaire statement) were proved by an *analysis of variance* (ANOVA).

Some items were bi-polar (for instance, a 5-point scale ranging from "characters talk too much" to "talk too little"). In this case, both ends of the scale represent negative extremes. Therefore, the ideal value is "3". Tendencies to one of the negative sides were proved by *t-tests for one sample*.

### 5.1 Role of Gender and Age

Nine women and seven men participated in the evaluation. A 2 x 34 ANOVA with the factors gender and questionnaire statement revealed a significant effect of the factor questionnaire statement ($F(33, 462) = 5.80; p < .001$) which simply means that different questions were answered differently by the subjects. However, did gender display any visible answering pattern? The factor gender ($F(1, 14) = 1.34; p = .27$) as well as the interaction between gender and questionnaire statement ($F(33, 462) = 1.02; p = .45$) were not significant. This pattern of

effects does not change when analysing the four main issues of the questionnaire separately. There are no significant gender differences, men and women perceive the system in a similar way.

Since our subjects could be split into two age groups of similar size (9 subjects 19-30 years, 8 subjects 31-45 years) we could also compare age groups differences using analoguous means as for gender. However, we did not find any significant effect. Subject of age groups 19-30 and 31-45 perceived the system in a similar way.

### 5.2 General Impression

The subjects found the interaction entertaining ($t(15) = 7.46; p < .001$). At the same time, the interaction was informative in terms of cars ($t(15) = 3.00; p < .01$) but middle informative in terms of computer technology ($t(15) = .75; p = .47$).

The car construction task demanded the participants' attention ($t(15) = 3.16; p < .01$) and the characters' comments to the car construction task were perceived as helpful ($t(15) = 4.34; p < .001$). The participants rated the reactions of the characters towards their actions to be appropriate and neither very fast nor too slow ($t(15) = -1.25; p = .23$).

All in all, the participants marked the exhibit with 4.2 which is highly significantly above average ($t(15) = 8.73; p < .001$).

### 5.3 Characters

Despite the demands of the car construction task the virtual characters were able to catch participants' attention ($t(15) = 4.87; p < .001$).

Both characters were rated above mean in regard to liking ($t(15) = 5.66; p < .001$), competence ($t(15) = 3.81; p < .01$) and politeness ($t(15) = 12.85; p < .001$). Only concerning the impression of human-likeness the characters received mean rating ($t(15) = .99; p = .34$). Regarding talkativeness, the characters were rated to be rather too talkative ($t(15) = 2.35; p < .05$). See Figure 2 for a direct comparison of mean values for both characters.

A comparison of the personality profile of both characters calculating a 2 x 5 ANOVA with the factors character (male and female) and personality attribute (likable, competent, polite, human-like, talkative) did not reveal a main effect of the character ($F(1, 16) = 3.10; p = .10$). Overall, both characters received the same degree of positive perception. The factor personality attribute showed a significant effect ($F(4, 64) = 8.14; p < .001$). Moreover, the interaction between character and personality attribute was highly significant ($F(4, 64) = 5.52; p < .001$). This demonstrates that both characters are perceived as virtual agents with different personality profiles. Post-hoc comparisons using LSD-tests showed that the female character was rated as more likable ($p < .001$), more competent ($p < .01$) and more human-like ($p < .05$). With regard to politeness and talkativeness the characters did not differ.

A direct comparison between the male and female character was requested by two questions in the questionnaire concerning dominance and sympathy. A

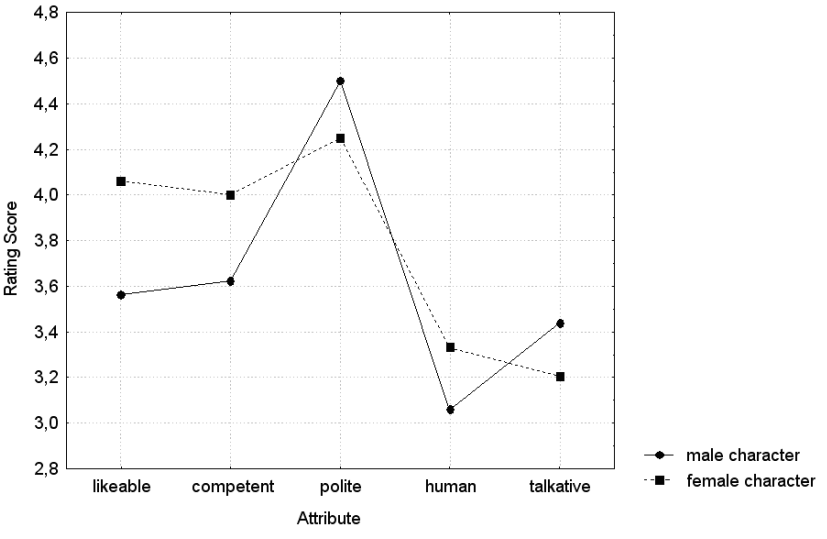


**Fig. 2.** Personality graphs of the two characters

value below 3 indicates an advantage for the male character and a value above 3 favours the female character. The female character was rated to be more dominant ($t(15) = 3.22; p < .01$) but in terms of sympathy the characters did not differ ($t(15) = 1.29; p = .22$).

### 5.4 Dialogue, Speech and Nonverbal Behavior

Subjects found the system too talkative ($t(15) = 3.42; p < .01$). However, at the same time the dialogue were perceived as being rich in variety ($t(15) = 5.51; p < .001$).

The intelligibility of speech was rated above average ($t(15) = 4.86; p < .001$) and the synthetic speech did not annoy the participants ($t(15) = 9.41; p < .001$). The coordination between speech and movements was rated average ($t(15) = 1.17; p = .26$).

### 5.5 Potential Target Age Groups

We asked the participants to estimate how enjoyable our exhibit would be for various age groups. Five disjoint age groups were presented. Except for the group "infants up to 6 years" ($t(15) = -.45; p = .66$) all age groups were expected to enjoy the exhibit (for all: $p < .01$). This is of course only an indicator for "appeal to different age groups". However, it is one way of approximation if time and/or budget do not allow to test a sufficient amount of subjects from the full range of age groups.

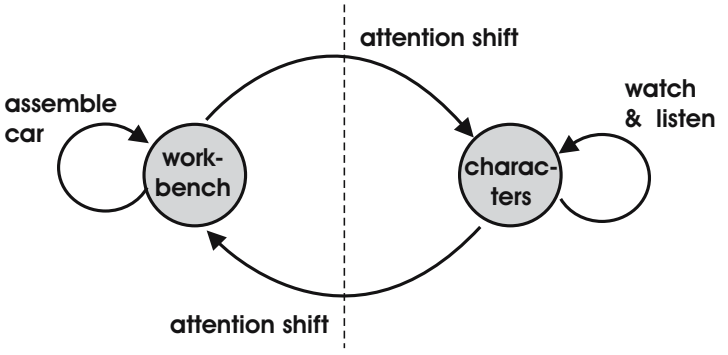


**Fig. 3.** The attention of the user repeatedly switches from assembling the car (active/play) and listening to the characters (passive/listen) and vice versa. The challenge is to find a good balance between entertainment (play) and education (listen).

## 6 Discussion

We presented an evaluation of a virtual character system that addressed important questions about the quality of the COHIBIT system in particular and about edutainment systems with virtual characters in general.

The major challenge when modeling VC behavior is to find a good balance between letting the user "play" with the car pieces and talking to him/her to convey some (possibly educating) information. This is a core problem in edutainment systems: you use entertainment to motivate, energize and get attention but you also need to get the educational content across – how do you balance the two? This pricipal dichotomy is even more pronounced in our system since the user cannot talk to the system, making the user either very active but unattentive (play) or very attentive but passive (listen) as depicted in Figure 3. Our questionaire focuses on this general problem. Since our tools allow to efficiently build behavior models, the flow of the conversion can be adapted very quickly after a number of exploratory tests. Thus, our iterative development has short and frequent test-adapt-compile loops. We used our questionaire to guide our development and, hopefully, our researchers can profit from it.

How entertaining the exhibit was found by the subjects was significantly correlated with the perception of the speech synthesis. The less annoying subjects found the virtual characters' speech, the more entertaining the whole exhibit was rated (Pearson Product-Moment Correlation: $r = .57; p < .05$). This confirms the finding by [3] that the quality of speech synthesis has a significant impact of how the system is perceived.

The fact that gender did not have an effect on perception is interesting, especially because the domain of cars would make you expect a slight bias towards male subjects. The number of subjects was too low to draw any hard conclusions but it appears as if a virtual character driven system does appeal to male and

female subjects in similar ways which is equally true for subjects differing in age, however restricted to two age groups of 19-30 and 31-45.

Concerning the characters' profiles we have carefully selected the personality traits to examine (Figure 2). We focused on those aspects that we wanted to design. They can be considered "quality criteria" of the characters' personality design. We wanted the characters to be likable, competent, polite, not too talkative and very human-like. We succeeded in most of these aspects. However, humanlikeness appears to be an important avenue for future improvement. The characters were also perceived as too talkative which might also be due to the quality of the speech synthesis. This result is balanced by the fact that subjects found the dialogue varied and informative. In terms of personality, we wanted the characters to differ so that they are perceived as distinct. The graph in Figure 2 shows that this was the case, especially with respect to liking and competence. Since there remains work to be done with respect to human-likeness, further evaluations should identify the factors that govern human-likeness. Synchrony of nonverbal behavior with speech was only rated average which might indicate that improvement might be necessary here to increase human-likeness, although again speech synthesis quality might have influenced this judgment negatively.

Our evaluation showed that our system was both informative and entertaining. The characters were able to catch the attention of the user without distracting him/her from the assembly task (Figure 3). On the contrary, the characters' comments were perceived as helpful in the task. Moreover, subjects thought that the exhibit would be enjoyed by people in a wide age range (everyone older than 6). Thus, the evaluation was helpful in confirming that our system was perceived as intended, potentially enjoyable to a wide range of people, and in identifying hotspots for future work.

## Acknowledgements

Thanks to our colleague Martin Rumpler and our students Gernot Gebhard and Thomas Schleiff for their excellent work on the COHIBIT system. This work is partially funded by the German Ministry for Education and Research (BMBF) as part of the VirtualHuman project under grant 01 IMB 01A.

## References

1. Rist, T., André, E., Baldes, S., Gebhard, P., Klesen, M., Kipp, M., Rist, P., Schmitt, M.: A review of the development of embodied presentation agents and their application fields. In Prendinger, H., Ishizuka, M., eds.: Life-Like Characters – Tools, Affective Functions, and Applications. Springer, Heidelberg (2003) 377–404
2. Ndiaye, A., Gebhard, P., Kipp, M., Klesen, M., Schneider, M., Wahlster, W.: Ambient intelligence in edutainment: Tangible interaction with life-like exhibit guides. In: Proceedings of the first Confernce on INtelligent TEchnologies for interactive enterTAINment (INTETAIN), Berlin, Heidelberg, Springer (2005) 104–113

3. McBreen, H., Anderson, J., Jack, M.: Evaluating 3D embodied conversational agents in contrasting VRML retail applications. In: Proceedings of the Workshop on "Multimodal Communication and Context in Embodied Agents" held in conjunction with the Fifth International Conference on Autonomous Agents (AGENTS), Montréal (2001) 83–87
4. Ruttkay, Z., Pelachaud, C., eds.: From Brows to Trust: Evaluating Embodied Conversational Agents. Kluwer (2004)
5. Hartmann, B., Mancini, M., Buisine, S., Pelachaud, C.: Design and evaluation of expressive gesture synthesis for embodied conversational agents. In: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems. ACM Press (2005)
6. van Mulken, S., André, E., Müller, J.: The Persona Effect: How Substantial is it? In: Proceedings of the British Computer Society Conference on Human Computer Interaction (HCI 98). (1998) 53–66
7. Klesen, M., Kipp, M., Gebhard, P., Rist, T.: Staging exhibitions: Methods and tools for modeling narrative structure to produce interactive performances with virtual actors. Virtual Reality. Special Issue on Storytelling in Virtual Environments **7**(1) (2003) 17–29
8. Gebhard, P., Kipp, M., Klesen, M., Rist, T.: Authoring scenes for adaptive, interactive performances. In: Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems. (2003) 725–732
9. Kipp, M.: Creativity meets Automation: Combining Nonverbal Action Authoring with Rules and Machine Learning. In: Proceedings of the 6th International Conference on Intelligent Virtual Agents (IVA 2006), Springer (2006)
10. Westermann, R.: Empirical tests of scale type for individual ratings. Applied Psychological Measurement **9** (1985) 265–274