User Interview-Based Progress Evaluation of Two Successive Conversational Agent Prototypes

Niels Ole Bernsen and Laila Dybkjær

NISLab, University of Southern Denmark, Odense
{nob, laila}@nis.sdu.dk

Abstract. The H. C. Andersen system revives a famous character and makes it carry out natural interactive conversation for edutainment. We compare results of the structured user interviews from two subsequent user tests of the system.

1 Introduction

The Hans Christian Andersen (HCA) system has been developed in the European NICE project on Natural Interactive Communication for Edutainment (2002-2005). Computer games company Liquid Media, Sweden, did the graphics, Scansoft, Germany, trained the speech recogniser with children's speech, CNRS-LIMSI, France, did the 2D gesture modules and the input fusion, and NISLab developed natural language understanding, conversation management, and response generation.

3D animated fairytale author HCA is found in his study in Copenhagen where he wants to have edutaining conversation with children (target users are 10-18 years) about the domains he is familiar with or interested in, such as his life, fairytales, himself, his study, the user, and the user's favourite games. HCA. The intended use setting is in museums and other public locations where users from different countries can have English conversation with HCA for a duration of 5-15 minutes. The user communicates via spontaneous speech and 2D gesture while 3D animated HCA communicates through speech, gesture, facial expression, and body movement.

2 The HCA Prototype Systems

Two prototypes (PT1 [1] and PT2 [3]) of the HCA system were tested with representative users in January 2004 and February 2005, respectively, following similar protocols. In both tests, subjective user data was gathered in post-trial interviews.

Perhaps the most important difference between PT1 and PT2 is that PT2 uses automatic speech recognition. In PT1, speech recognition was emulated by human wizards. PT1 thus has near-perfect speech recognition whereas PT2 must deal with the additional technical difficulties of recognising the speech of children who, moreover, have English as their second language. Other important differences include that in PT2, the user can change the topic of conversation, backchannel comments on what

HCA said, or point to objects in his study at any time, and be responded to when appropriate. This yields a far more flexible conversation than was possible in PT1. Also, the handling of miscommunication has been improved in PT2. Although HCA's domain knowledge has been extended in PT2 as well, the major change is in the restructuring of his knowledge, i.e., in how the user can converse with HCA and get access to his knowledge, and in what HCA does when he has, or takes, the initiative.

Contrary to PT1, HCA can in PT2 display several gestures simultaneously and has semi-natural lip synchrony as well as some amount of face, arm and body movement. In PT1, HCA has a single output state, i.e., the one in which he produces conversational output. If no user is present, he does nothing but wait. In PT2, when alone, HCA walks around thinking, looks out his windows, etc. However, this new output state is not properly integrated with the conversational output state, and HCA's behaviour when alone is also sometimes rather weird. A problem in PT1 was that the gesture recogniser was always open for input. Those users who had a mouse and no touch screen tended to create large queues of gestures waiting to be processed, which generated internal system problems as well as some contextually inappropriate conversational contributions by HCA. The PT2 gesture recogniser does not "listen" while processing input. The same is true for the speech recogniser which does not have barge-in.

3 The User Tests

PT1 was tested with 18 users (17 Danes and 1 Scotsman, 9 girls and 9 boys), 10-18 years old. PT2 was tested with 13 Danish users (7 girls and 6 boys), 11-16 years old. Both tests included two test conditions and similar sets of user instructions for both conditions. Two test rooms were prepared with: a touch screen, except that for PT1 one of the rooms had a standard screen and a mouse for pointing; a keyboard for changing virtual camera angles and make HCA walk; a headset; and two cameras for recording user-system interaction. The software was running on two computers. The animation was on the computer connected to the user's screen and the rest of the system was on the second computer which, for PT1, was operated by the wizard and, for PT2, was being monitored by a developer out of sight of the user. User input, wizard input (PT1-only), system output, and interaction between modules was logged.

Each user test session took 60-75 minutes. Sessions began with a brief introduction to the input modalities available. The PT2 headset microphone was calibrated to the user's voice. The users were *not* instructed in how to speak to the system. In the PT1 test, this did not matter since the wizards would type in what the user said, ignoring contractions, disfluencies, etc., and only making few typos. We wanted to collect baseline data on how second-language speakers of English, most of whom had never spoken to a computer, talk to a conversational system with no prior instruction.

After the introduction followed 15 minutes of free-style interaction. It was entirely up to the user what to talk to HCA about. In the following break, the user was asked to study a handout listing 13 (PT1) and 11 (PT2) proposals, respectively, for what the user could try to find out about HCA's knowledge, make him do, or explain to him. It was stressed that the user did not have to follow all the proposals. The second session had a duration of approx. 20 minutes. In total, some 11 hours of interaction were recorded on audio, video, and logfiles for PT1, and some 8 hours in the PT2 test.

4 The PT1 and PT2 User Interviews

Users were interviewed immediately after their interaction with the system. The PT1 and PT2 interviews comprised 20 and 29 questions, respectively, see also [2, 3]. In both cases, the first six questions concerned the user's identity, background, computer gaming experience and experience in talking to computers. For PT2, we also asked about the user's experience in using a touch screen. Below, we focus on the other sections of the interviews which address system interaction and usefulness issues.

Due to increased functionality 14 PT2 versus only seven PT1 questions deal with the user's interaction with the system. Six PT1 and seven PT2 questions address system usefulness and suggested improvements. The questions are identical but for a PT2 question on overall system quality. In both interview series users were asked for any other comments. This question did not add any new information.

Each user's verbatim response to each question was scored independently on a three-point scale by two raters. Rating differences were negotiated until consensus was reached. An average score per question was then calculated (Figure 1). Grouping the issues raised in the interviews, the following picture emerges.

HCA's *spoken conversational abilities* have improved significantly in PT2. Conversation management problems do not enter into the PT2 replies on whether it was *fun to use* the system and if it was *easy to use*, and only rarely into the PT2 replies on what *was bad about the interaction,* but those problems figure prominently in the corresponding replies regarding PT1. Regarding PT1 users focused on slow gesture understanding and various problems in being understood. PT2 users focused on minor difficulties of manual control of camera angles and HCA's locomotion. Concerning the question of *what was bad about the interaction,* PT2 answers have much less of: did not change topic when the user wanted to, irrelevant replies, too much repetition, did not answer questions. For both PTs, the users want HCA to have more knowledge.

The answers to *whether HCA could understand what was said* and *what was good about the interaction,* support the conclusion that conversation has improved considerably. Despite the very significant decrease in speech recognition performance in PT2, the PT1 problems of: many unanswered questions, several unwanted repetitions, and HCA not following user-initiated topic change due to an overly inflexible conversation structure, are gone. Wrt. the question of *what was good about the interaction,* the PT1 and PT2 users agree that it was good to talk to HCA in English, get information about himself and his life, and point to objects and get stories about them. Criticism of HCA's conversational abilities surfaces in the question about *suggested improvements.* In both PT1 and PT2, there is a wish that HCA can understand more.

Conversely, the increase in animation articulation and expressiveness, and the reduction of the number of graphics bugs, in PT2 over PT1, is not rewarded by the users, cf. the questions on *naturalness of animation, what was bad about the interaction* and *what should be improved.* Despite PT1 users quite strong reaction to the presence of graphics bugs, the PT2 users react even more strongly to HCA's unnatural walk and antics. Other new functionality in PT2, such as lip synchrony, is appreciated, however.

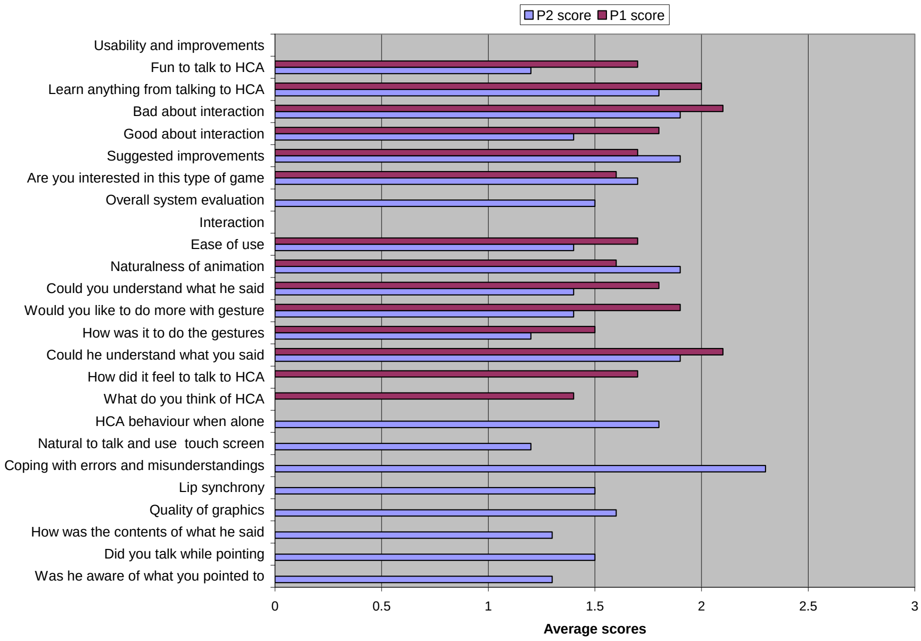

Figure 4. Average user ratings for each PT1 and PT2 question. Score 1 is the top score.

The intelligibility of PT2's *speech synthesis* is appreciated. The PT2 *touch-screen* is praised as giving more control than the mouse, even though this contradicts the PT1 scoring for mouse vs. touch screen. The use of *input gesture* is found more satisfactory in PT2 than in PT1. It is far more work to gesture using the touch screen than using the mouse and this may have reduced the PT2 users' wishes for more gesturable objects (the number of objects is the same for PT1 and PT2). The users' views on *learning from the system* are also better for PT2 than for PT1. The improved conversation may have affected the users' answers. Finally, the users' *interest in speech/gesture computer gaming* are similar for PT1 and PT2.

Two questions were only asked wrt. PT1 and nine only wrt. PT2. On the PT1-only question, what the user thinks of *the HCA character*, he is basically perceived as authentic. Given this positive feedback we included instead a more general question in PT2 about the *quality of the graphics*, which was rated rather good overall. Two more questions on the visual impression of PT2 were: one on the *lip synchrony* which users found quite good; and a question about *HCA's behaviour* when he is alone in his study which was evaluated quite negatively. The second PT1-only question addressed *how it feels to talk to HCA.* The answers mainly reflect that users took some time getting used to speaking to the system. The corresponding PT2 question asks *how natural it is to talk and use the touch screen.* Users replied very positively. A new, related PT2 question was *if the user talked while pointing and if it worked.* Half of the users did not talk while pointing while the rest did so occasionally. The score reflects that the multimodal input worked for almost all users who tried. The related question about *HCA's understanding of pointing input* was also answered very positively.

Of the three final PT2-only questions, one was about the *quality of the contents* of what HCA says which were generally felt to be fine though HCA tends to talk too much and is not sufficiently helpful in helping the user find something to ask him about. The question about *how easy it was to cope with errors and misunderstandings* received the harshest average score of all (2.3). This is where the system's imperfect speech recognition and limited vocabulary and domain knowledge take centre-stage. Finally, the users' *overall evaluation* was good with a majority of positive comments.

5 Conclusion

This paper has reported results from two similarly protocolled user tests with two research prototype generations of "the same" system. We have highlighted three major differences between PT1 and PT2, i.e., that only PT2 used automatic speech recognition, that conversation management was significantly improved in PT2 over PT1, and that PT2's animation was far more expressive and versatile. Whereas improvements in conversation management seem to have more than outweighed the adverse effects of a substantial amount of speech recognition failure in PT2, PT2's more expressive animation was not really perceived as natural or fun. The fact that users were neither instructed nor trained in how to speak to the system seems to have had a strong effect on their perception of the helpfulness of the system's meta-communication. Our next step is to correlate the subjective user evaluations with objective analysis of the conversations based on coding schemes for conversation robustness and success.

Acknowledgements

We gratefully acknowledge the support by the EU's HLT Programme, Grant IST-2001-35293. We would like to thank all colleagues at NISLab, Liquid Media, Scansoft, and LIMSI-CNRS who contributed to making the prototypes fit for user testing.

References

[1] Bernsen, N.O., Charfuelàn, M., Corradini, A., Dybkjær, L., Hansen, T., Kiilerich, S., Kolodnytsky, M., Kupkin, D., and Mehta, M.: First Prototype of Conversational H. C. Andersen. Proceedings of AVI 2004, Gallipoli, Italy, 2004, 458-461.

[2] Bernsen, N. O. and Dybkjær, L.: Evaluation of Spoken Multimodal Conversation. Proceedings of ICMI 2004, Penn State University, USA, 2004, 38-45.

[3] Bernsen, N.O. and Dybkjær, L.: User Evaluation of Conversational Agent H. C. Andersen. Proceedings of Eurospeech, Lisbon, Portugal, 2005.