Structured Interview-based Evaluation of Spoken Multimodal Conversation with H.C. Andersen

Niels Ole Bernsen and Laila Dybkjær

Natural Interactive Systems Laboratory, University of Southern Denmark
Campusvej 55, 5230 Odense M, Denmark
(nob,laila)@nis.sdu.dk

Abstract

This paper presents evaluation results on system performance and interaction from the user test of the first prototype of a multimodal conversational system. The system enables spoken and gestural interaction with life-like fairytale author Hans Christian Andersen about his fairytales, life, study, etc. The evaluation is based on structured interviews with 18 target users after their conversations with the system in a controlled laboratory setting. The obtained results are encouraging.

1. Introduction

One of the many exciting research challenges facing today's spoken dialogue systems (SDS) developers is to venture beyond the successful paradigm of *task-oriented SDSs.* A task-oriented SDS is a system which helps the user complete one or several tasks, such as getting train information, making flight ticket reservation, or contacting someone over the phone via an automated SDS switchboard service [2]. At this point, we lack adequate terminology for describing the huge space of potential non-task-oriented systems. To address this problem, we may distinguish between, on the one hand, the ultimate goal of developing members of the class of Turing test-compliant SDSs [6] and, on the other, the intermediate goal of developing *domain-oriented SDSs.* A domain-oriented SDS is defined solely by the domain(s) it has been designed to conduct spoken dialogue about. The developer no longer designs for any specific user task(s) nor can the developer assume the existence of shared goals among user and system. Rather, the system is like a human who has knowledge, beliefs, attitudes, etc. concerning one or several domains about which it is able to carry out unrestricted conversation. The user may address the system's domain(s) in any way and for any purpose and still expect the SDS to respond appropriately. This is why we tentatively call such systems *"real" conversational systems.*

If the domain-oriented SDS includes (an) embodied animated interface agent(s), we encounter an additional terminological issue because SDSs involving such agents are currently called embodied conversational agents, or ECAs, despite the fact that all or most of them are still task-oriented [7][8] and hence, as just argued, non-conversational.

In this paper, we first briefly describe the first running prototype of a domain-oriented SDS that allows users to have English speech-cum-2D-gesture conversation with 3D life-like fairytale author Hans Christian Andersen (HCA) in his 19th century study (Figure 1.1). The system was developed in the NICE (Natural Interactive Conversation for Edutainment) project [9]. We then evaluate the system's performance based on structured interviews with target users after their conversations with the system in a controlled laboratory setting.

Figure 1.1. HCA in his study.

2. The HCA system

The HCA system aims to enable edutaining conversation with 10-18 years old users in a public location, such as the HCA museum in Odense, Denmark, for an average duration of, say, 5-15 minutes. In generic terms, the system is a new kind of computer game which integrates spoken conversation into a professional computer game environment.

The architecture of the system is shown in Figure 2.1 and described in more detail in [1]. The speech recogniser is greyed out because it was not integrated in the first prototype. It still needs to be trained on 40-50 hours of speech data recorded with mostly non-native English speaking children and will be included in the second HCA prototype.

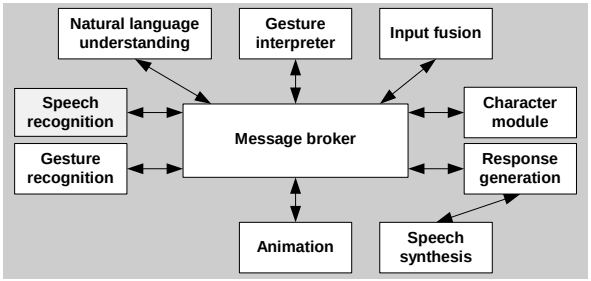

Figure 2.1. General NICE HCA system architecture.

NISLab has developed HCA's natural language understanding, character module [3] and response generation [5] components. The other components shown in Figure 2.1 have been developed by other NICE project partners or are (based on) freeware (gesture recognition, message broker, and speech synthesis). The project partners are: TeliaSonera, Sweden, Liquid Media, Sweden, Scansoft, Belgium, and LIMSI/CNRS, France.

HCA's domains of knowledge and discourse include his fairytales, his childhood in Odense, his persona and physical presence in his study, getting information about the user, his role as "gatekeeper" for the fairy tale games world (developed by project partner TeliaSonera and not described here), and the "meta" domain of resolving problems of miscommunication. These domains are probably among those which most users would expect anyway. HCA is *not* designed as a (task-oriented) Q&A machine for those domains but as a quasi-person who aspires to excel in conversation about them.

3. The user test

The HCA prototype was tested in January 2004 with 18 users (nine girls and nine boys) from the target user group of 10-18 year old kids and teenagers. The recogniser was replaced by a wizard who typed what the user said. The rest of the system was running. Users arrived in parallel, so there were two test rooms, two wizards, and two interviewers. In one room, the user had a mouse and a touch screen for gesture input while in the other room only a mouse was available as pointing device. In the room with the touch screen, the user could also watch HCA on a 42" flat-panel screen. An observer was present in this room as well, cf. Figure 3.1.

Figure 3.1. User in front of the touch screen.

Each user test session had a duration of 60-75 minutes. A session included conversation with HCA in two different conditions followed by a post-test interview. In the first, 15 minutes condition, the users only received basic instructions on how to operate the system, i.e. to speak using the headset, control HCA's movements, control the four virtual camera angles, and gesture using mouse or touch screen. As expected, most of the initiative was with HCA during the first session. In the second condition, the user received a set of 13 brief scenarios, such as "Find out if HCA has a preferred fairytale and which it is", "Make HCA tell you about two pictures and two other objects in his study", and "Tell HCA about games you like or know". The user fully decided on the order and number of scenarios to solve. The purpose of the scenarios was to increase user initiative in order to explore how the system would respond under the resulting input "pressure".

All interactions were logged, audio recorded, and video recorded. In the room with the touch-screen, a video camera pointed at the user and a second camera recorded the screen, cf. Figure 3.1. In the second room, a single camera recorded the user. In total, approximately 11 hours of interaction were recorded on audio, video, and logfile, respectively. In addition, 18 sets of structured interview notes were collected. Henceforth, we focus on the evaluation of the system based on the user interviews.

4. The user interviews

The structured post-session interviews took between 15 and 30 minutes per user. Each user was invited to simply report what came to mind when asked each of the following 20 questions:

1. User identity: Name, age, gender.
2. Occupancy.
3. How often do you play computer games: hours per week?
4. (If relevant) Which computer games do you like (types of game or concrete games)?
5. Did you ever talk to a computer before? If yes, which program did you use?
6. How well do you know HCA?
7. Was it easy or difficult to use the system? Why?
8. What do you think of HCA?
9. Could you understand what he said?
10. How did it feel to *talk* to HCA?
11. Could he follow what you wanted to talk to him about?
12. What do you think of his behaviour on the screen?
13. How did it feel to be able to use input gesture? (a) Did you use the mouse or point onto the screen? (b) How was it to do the gestures? (c) Would you like to be able to do more with gesture? If yes, what?
14. Was it fun to talk to HCA? If yes, what was fun? If no, can you imagine what could make it fun?
15. What did you learn from talking to with HCA?
16. What was bad about your interaction with HCA?
17. What was good about your interaction with HCA?
18. What do you think we should make better?
19. How interested would you be in playing computer games with speech and gesture?
20. Any other comments?

Structurally, questions (1) through (6) collect user information, questions (7) through (13) collect information on how the users experienced the interaction, questions (14) through (19) elicit information on the system's perceived usefulness and how it could be improved, and the final open question (20) invites any comments which were not elicited so far.

5. User information

Table 5.1 presents the basic user information gathered. The table shows gender balance, user average age at the centre of the target group of 10-18 year olds, and a one-year average age difference between girls and boys, suggesting a higher level of mastery of English among the girls, all users except one being in the process of learning English as second language. Except for one user, however, a 12-year old girl, all users courageously fought the English language and managed to conduct the intended conversations with HCA pretty well, grammatical errors and all notwithstanding. This girl only managed to input two spoken utterances in total and hardly understood what HCA said. Her (humorous) comments on the system, therefore, tended to be rather negative (U9 in Table 6.1). We shall disregard her generally atypical comments in Section 7 below as being largely irrelevant to the evaluation.

Table 5.1: Basic user information

Property	**Value**
No. girls	9
No. boys	9
Nationality	17 Danish, 1 Scottish
School pupil/student	all
Age range girls	12-17
Age range boys	10-18
Girls, average age	14.8
Boys, average age	13.8
All, average age	14.3
Girls, computer game playing range: hours per week:	0-21
Boys, computer game playing range: hours per week	0-24.5
Girls, average game hours/week	3.9
Boys, average game hours/week	11.1
All, average game hours/week	7.5
Talked to a computer before	3 (all girls)
Average knowledge of HCA	2.17

On average (Table 5.1), the boys do far more computer gaming than the girls. Only three users had spoken to a computer prior to their conversations with HCA. We quantified the users' expressed prior knowledge of HCA on a scale from (1)-top through (2)-good to (3)-poor, achieving a slightly-lower-than-good average of 2.17. Danish kids tend to grow up with HCA's stories both at home and at school.

6. Qualitative scoring of interviews

Applying the (1)-positive, (2)-medium, (3)-negative qualitative scoring principle introduced above, Table 6.1 offers a coarse overview of how the users perceived the system in terms of interview questions (7) through (19). The average score of 1.72 reflects a better-than-medium perception of the system. In our scoring, only 30 user judgments issued in a score of 3. Of these, 11 were due to user id=9, i.e. the girl who did not manage to have conversation with HCA at all. Another critical user is id=16 who, like several other users, judged the system from the point of view of its multi-hour use potential in the home. They rightly point out that the system is not yet rich enough to sustain multi-hour use. We did not inform the young users about the system's intended short-duration use in public environments as we did not want them to take abstract requirements into account when judging the system. Eight '3's occur in Column 13c, showing that many users did not want additional gesture functionality. Thus, Table 6.1 and the observations just made may be taken to show that the HCA system was judged quite favourably overall by its target users.

Table 6.1: Qualitative scoring of user perceptions. U is user id. Top-row numbers refer to interview questions (Section 4).

U	7	8	9	10	11	12	13a	13b	13c	14	15	16	17	18	19
1	2	2	2	1	2	1	Mouse	2	3	2	1	2	2	2	2
2	1	2	2	1	2	3	Mouse	1	3	2	2	2	1	2	2
3	1	1	2	2	2	1	Both	2	1	2	2	2	1	1	2
4	1	1	2	2	2	2	Touch	2	3	3	2	2	1	1	3
5	1	1	2	1	2	2	Mouse	2	3	1	2	1	1	1	2
6	2	1	2	1	2	2	Touch	2	3	1	2	2	1	2	2
7	2	1	1	1	2	1	Touch	1	1	1	2	2	2	2	1
8	1	1	2	2	1	1	Touch	2	2	1	2	2	1	1	1
9	3	3	3	3	3	1	Mouse	1	2	3	3	3	3	3	3
10	1	2	2	1	2	2	N/A	N/A	2	1	2	2	2	2	1
11	2	1	2	1	2	1	Touch	2	3	1	2	2	1	1	1
12	2	1	2	1	2	1	Touch	1	1	1	2	2	2	1	2
13	2	1	2	2	2	1	Touch	2	1	2	2	3	2	1	2
14	2	1	2	2	2	2	Mouse	1	3	1	3	2	2	1	2
15	1	1	1	1	2	1	Mouse	1	1	1	2	2	2	2	1
16	3	3	1	1	2	2	Mouse	1	2	3	3	2	2	3	2
17	1	1	1	1	3	2	Mouse	N/A	3	1	2	2	2	2	1
18	2	1	1	1	1	1	Touch	1	1	1	2	2	2	2	1

7. Evaluation of the interaction

We now look in detail at the users' evaluation of the interaction, i.e. questions (Qs) 7 through 13, cf. Section 4. Numbers in parentheses show how many users shared a comment.

The bulk (6) of the critical comments on how easy the system was to use (Q7) concern HCA's occasional difficulties in understanding what the users said, as evidenced by, e.g., irrelevant output and unnecessary repetitions. HCA himself (Q8) was generally received quite positively, being realistic, life-like, imaginative, and fun to watch (15). HCA's spoken intelligibility (Q9) was received surprisingly positively by his mostly non-native English-speaking interlocutors. The main criticism (6) was that the RealSpeak synthesiser sometimes "swallowed" or did not properly segment some syllables. Fifteen users had not spoken to a computer before (Q10, cf. Table 5.1). They found the experience strange, surprising (10), fun (6), or easy, like talking to a person (3). Three users found it embarrassing to talk to HCA while being observed.

One of the key interview questions (Q11) was if HCA could follow what the user wanted to talk to him about. One user was largely happy with HCA's conversational abilities and a single user (id=17) was rather dissatisfied (still ignoring id=9). The main criticisms were that HCA's output was sometimes irrelevant (15) or unnecessarily repetitive (3). Analysis of the transcribed conversations shows that these problems were aggravated in the second test condition (Section 3) in which the users put HCA under heavy-handed direct interrogation in order to quickly get through the scenarios. The scenarios had been designed to make this simple strategy fail. Two users observed that HCA stuck too much to some of his pet topics. Two users noted that he could understand one input formulation but not another, equivalent one.

HCA's on-screen behaviour, including movements, gestures, and facial expressions (Q12), was considered OK, fine, great, or good by the large majority of users (13). Five users noted the remaining bugs in the graphics, which made HCA able to stand in his furniture and go through walls. Only two users expressed the wish that he would be more lively.

Finally, concerning gesture interaction, we already noted that most users did not see a need for more of it (Q13c, Section 6). Otherwise, input gesturing (Q13b) was found to be easy to do (9) albeit slow in processing (5). Regarding Q13a on the gesture input device used, one user (id=3) who had the choice between touch screen and mouse, remarked that she preferred the mouse because her arm occluded the screen when reaching to touch it.

8. Usefulness and improvements

The question if it was fun to talk to HCA (Q14) addresses the system's entertainment qualities and received rich feedback. The bulk of the comments were that it was entertaining, fun, exciting, or great to talk to HCA (7), fine that he told long stories (4), and fun to get stories about objects by pointing to them (3). Several users (3) missed the multi-hour gaming perspective (Section 6), and a couple re-emphasised HCA's less-than-perfect conversational abilities.

The equally rich answers on the system's educational import (Q15) included several surprises. A minor surprise was that most Danish users considered HCA's fairytale knowledge as knowledge reminders rather than novelties. More surprisingly, most users (11) strongly valued HCA's stories about his life and said that they learned a lot from them. The real surprise was that five users pointed out the system's value for training their English skills, casting an entirely different light on the system's educational potential from what we had anticipated.

The system criticisms (Q16) centred on HCA's less-than-human linguistic and conversational skills, with 11 comments, and the system's less-than-perfect graphics, with six comments. Four users admitted their English language difficulties at this point. The system praise (Q17) may be summarised by quoting the user who said that the system is on the right track overall. Even the graphics bugs were praised by one user.

Essentially, the rich data on system improvement (Q18) expresses a wish for more of the same, with 14 comments, no bugs in the graphics (4), and better spoken input understanding (2). To the key question (Q19) on the users' interest in speech/gesture input computer gaming, no less than 12 users felt that spoken conversation might make games more entertaining, interesting, and immersive, many of them being quite precise as to the types of games which might benefit the most from spoken conversation. Finally, the any other comments question (Q20) did not add much to the above.

9. Discussion

Needless to say, the successful development of "real" conversational systems is a major challenge. Viewed in this light, the user test interviews following interaction with the first HCA prototype may be described as, even surprisingly, encouraging. Overall, the users found that the technology is on the right track and represents a first glimpse of entirely new spoken computer games technology which could significantly improve the entertainment and educational value of computer games as well as achieving a new level of user immersion.

Based on the collected data from the user test and data collected in an earlier, fully simulated Wizard of Oz setup of the system [4], the second HCA system prototype is now being designed and developed with particular emphasis on increased conversational smoothness and flexibility.

10. Acknowledgements

We gratefully acknowledge the support of the NICE project by the European Commission's Human Language Technologies (HLT) Programme.

11. 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", *Proc. of the International Working Conference on Advanced Visual Interfaces* (AVI 2004), Gallipoli, Italy, 2004.

[2] Bernsen, N. O., Dybkjær, H. and Dybkjær, L., *Designing Interactive Speech Systems. From First Ideas to User Testing,* Springer Verlag, London, 1998.
[3] Bernsen, N. O. and Dybkjær, L., "Domain-Oriented Conversation with H.C. Andersen", *Proc. of the Workshop on Affective Dialogue Systems*, Kloster Irsee, Germany, 2004.
[4] Bernsen, N. O., Dybkjær, L. and Kiilerich, S., "Evaluating Conversation with Hans Christian Andersen", *Proc. of the Fourth International Conference on Language Resources and Evaluation* (LREC), Lisbon, Portugal, 2004.
[5] Corradini, A., Fredriksson, M., Mehta, M., Königsmann, J., Bernsen, N. O., and Johannesson, L., "Towards Believable Behavior Generation for Embodied Conversational Agents", *Proc. of the Workshop on Interactive Visualisation and Interaction Technologies*, IV&IT, Krakow, Poland, 2004.
[6] Turing, A., "Computing Machinery and Intelligence", *Mind* 59, 1950, 433-60.
[7] Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (Eds.), *Embodied Conversational Agents.* MIT Press, Cambridge, MS, 2000.
[8] Rist, T., Aylet, R., Ballin, D. and Rickel, J. (Eds.), *Proc. of the 4th International Working Conference on Intelligent Virtual Agents (IVA),* Springer Verlag, Berlin, 2003.
[9] NICE, http://www.niceproject.com/