University of Brighton

ITRI-03-09

The effect of gestures on the perception of a dialogue between two embodied conversational agents: a pilot study

Paul Piwek

July, 2003

EC Project NECA IST-2000-28580

Information Technology Research Institute Technical Report Series

ITRI, Univ. of Brighton, Lewes Road, Brighton BN2 4GJ, UK
TEL: +44 1273 642900 EMAIL: `firstname.lastname@itri.brighton.ac.uk`
FAX: +44 1273 642908 NET: `http://www.itri.brighton.ac.uk`

NECA

NET ENVIRONMENT FOR EMBODIED EMOTIONAL
CONVERSATIONAL AGENTS

The effect of gestures on the perception of a dialogue between two embodied conversational agents: A pilot study

Paul Piwek

ist
information
society
technologies

Version: ***FINAL***

Date: July 14, 2003

Project ref. no.	*IST-2000-28580*
Project title	**NECA: A Net Environment for Embodied Emotional Conversational Agents**
Deliverable status	**N.A.**
Contractual date of delivery	*N.A.*
Actual date of delivery	*July 14, 2003*
Deliverable number	*N.A.*
Deliverable title	N.A.
Type	Report
Status & version	
Number of pages	12
WP contributing to the deliverable	WP9
Task responsible	ÖFAI
Author(s)	Paul Piwek
EC Project Officer	Patrice Husson
Keywords	Gestures, Empirical Study, Experiment, Evaluation
Abstract (for dissemination)	In this report, we describe an experiment in which the effect of the presence (+G) and absence of gestures (–G) in an automatically generated dialogue between two embodied conversational agents was examined. In particular, we focused on correlations between +/–G and 1. the ability of subjects who witnessed the dialogue to correctly answer questions regarding the content of the dialogue and 2. judgments on how engaging a dialogue was. We also tested whether subjects would judge the politeness of the characters in line with the system settings for this parameter. Finally, we asked our subjects for any comments. This yielded some interesting results regarding the relation between +/–G and judgments about the quality of speech synthesis.

Paul Piwek

ITRI – University of Brighton

Watts Building

Moulsecoomb, Brighton BN2 4GJ]

Phone: 00 44 (0)1273 64 29 16

Fax: 00 44 (0)1273 64 29 08

Email: Paul. Piwek@itri.bton.ac.uk

Contents

Executive Summary

In this report, we describe an experiment in which the effect of the presence (+G) and absence of gestures (–G) in an automatically generated dialogue between two embodied conversational agents was examined. In particular, we focused on correlations between +/–G and 1. the ability of subjects who witnessed the dialogue to correctly answer questions regarding the content of the dialogue and 2. judgments on how engaging a dialogue was. We also tested whether subjects would judge the politeness of the characters in line with the system settings for this parameter. Finally, we asked our subjects for any comments. This yielded some interesting results regarding the relation between +/–G and judgments about the quality of speech synthesis.

Introduction

The aim of this paper is to investigate the influence of gesturing on the perception of embodied conversational agents. In particular, we focus on the role of the gestures as they are generated in the NECA system (Krenn et al., 2002).

The modules of the NECA system are arranged in a pipeline. The input to the system is a database of facts, value judgements, role assignments to characters and personality descriptions of characters. The first module, the Scene Generator (André et al., 2000; Krenn et al., 2002), takes this information and turns it into a scene description. This scene description contains the dialogue acts that are to be performed in the scene, their temporal ordering, a specification of the interlocutors and their common ground at the outset of the dialogue. This information is passed on to the Multimodal Natural Language Generator (MNLG; Piwek, 2003). The MNLG determines the linguistic realization of the dialogue acts, taking into account the evolving common ground of the interlocutors. It also, after linguistic realization, adds certain gestures. The version of the MNLG that we discuss in this paper inserts two types of gestures:

1. *Turn-taking signals*: when a speaker has finished a turn, s/he looks at the other interlocutor and continues to do so whilst the other interlocutor speaks. When a speaker begins speaking, s/he looks slightly away from the other interlocutor.

2. *Discourse function signals*: these gestures are associated with the dialogue act type of an utterance. A distinction is made between, for instance, *inform* and *request* dialogue acts. The former cause the speaker to extend his/her hand to the hearer in a downward movement. The latter can cause the speaker to place their hands on their hips or raise a finger in the air. For a particular dialogue act, the generator selects at random a gesture from a set of suitable gestures. This approach is aimed at introducing some variation into the dialogue.

The output of the MNLG is sent for realization to the MS agents player (which includes the L&H TruVoice text-to-speech system). The version of the system we discuss here – eShowroom– generates car sales dialogues between a seller and a customer. In another version of the system (Socialite), the dialogues consist of small talk/gossip between two students. The latter system also differs from the version of eShowroom discussed here in that there are two further stages between the rendering by the player and the MNLG: gesture alignment and concept-to-speech synthesis (Schröder & Trouvain, 2001; Schröder & Trouvain, to appear).

The main aim of the current study was to examine the effects of gestures on the audience of a dialogue. We wanted to find out whether the presence/absence of gestures influences 1. the ability of the audience to remember the information that is exchanged in a dialogue and 2. the extent to which the audience enjoys watching and listening to a dialogue. Two hypotheses were put forward.

> **I** Subjects who watch a dialogue with gestures will do better on a memory test than those who watch the same dialogue, but with all gestures removed.

The rationale for this hypothesis is that gestures provide additional information on the structure of the dialogue, e.g., on whether an utterance is an inform, a request or an evaluation and on the relations between utterances. The system would, for instance, render the following three utterances "This car is safe. It has airbags. It has anti-lock brakes." by having no gesture with the first evaluative utterance, but an arm movement (hand moving from vicinity of shoulder downward and forward with the palm directed upwards) for the second and third utterances, emphazing these as points in an argument for the initial evaluative statement.

> **II** Subjects who watch a dialogue with gestures will give a higher rating for how engaging the dialogue is than subjects who watch the same dialogue, but with all gestures removed.

The motivation for this second hypothesis is that a dialogue with gestures will come across as more lively and natural, due to its multimodality.

Finally, we also wanted to investigate a third quite separate issue. The NECA system generates dialogues whose characters can be polite or impolite. We wanted to see whether the audience is able to recognize whether a character is polite or impolite. Additionally, we investigated whether the presence/absence of gestures had any influence on this.

Method

Research Participants

The subjects of the study were University of Brighton undergraduates (Level 3 final year students on undergraduate computing courses, taking a module on Adaptive Interactive Systems). This sample consisted of 28 subjects, both men (22) and women (6). The materials were administered during a class meeting. Permission from the instructor and the students' consent was obtained.

Materials

Two different computer-generated movies of a dialogue between embodied characters were used. One dialogue was automatically generated with the eShowroom NECA system. A screen capture of the dialogue with gestures can be found in figure 1. The other dialogue was obtained by omitting all the gestures from the automatically generated dialogue. The dialogues were between a salesman, on the left, and his customer. The technology for playing the dialogues on the screen was Microsoft Agents. The voices for the male and female characters were realized by L&H TruVoice.

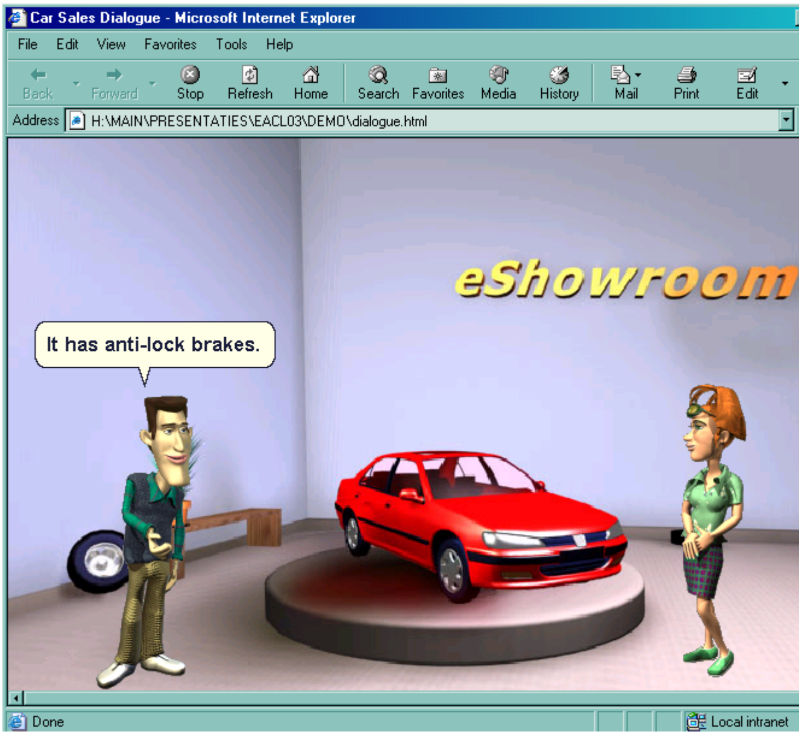
Car Sales Dialogue - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit
Address H:\MAIN\PRESENTATIES\EACL03\DEMO\dialogue.html
eShowroom
It has anti-lock brakes.
Done
Local intranet

Figure 1 Screen capture of eShowroom Dialogue with gestures between salesman and customer generated by the NECA system.

Procedure

Subjects were assigned randomly to the *with Gestures* (+G; *n* = 17) and *without Gestures* (–G ; *n* = 11) conditions. The groups were separately presented with a computer-generated movie of a dialogue between embodied characters: for one group the dialogue contained gesturing by the embodied characters, for the other it did not. Otherwise, the appearance and content of dialogue movies was the same. Before the start of the dialogue movie, the subjects were presented with the following instruction: ‘The dialogue which we are going to show you takes place between a salesman and his customer in a car sales showroom. Listen and watch closely.’ Both the instructions and the movie were presented using a data projector and a pair of loud speakers.

After the movie, answer sheets were handed out to the subjects and the subjects were presented with the following instructions on a slide: ‘If you look at the answer sheet, you will see a box at the top. Read the text and answer the two questions in the box.’ In the box, the subjects were asked to provide their sex and age. After a pause, a slide with the following instructions was presented: ‘Now you will be shown a number of questions, one at a time. You are supposed to answer the questions by ticking the box with the right answer on the answer sheet. You have 10 seconds for answering each question.’ After this slide, the following questions were presented:

1. Does the car have power windows?
2. Is the luggage compartment spacious?
3. Does the car have anti-lock brakes?

4. Does the car have airbags?

5. Does the salesman think that the car is safe?

6. Does the customer think that the car is prestigious?

7. Is the salesman polite?

8. Is the customer polite?

9. How engaging did you think the dialogue was?

For the first eight questions the subjects were asked to tick yes or no on the answer sheet. For the last question, they were asked to circle a number on a scale from 1 to 9, with 1 being marked as ‘very boring’ and 9 as ‘very engaging’.

Results

Gestures and memory

In the –G condition, the answers to the questions 1. to 6. contained 6 incorrect answers. Out of 11 subjects, 1 subject provided 3 incorrect answers and 3 subjects provided 1 incorrect answer each. In the +G condition, the answers to the questions 1. to 6. contained 4 incorrect answers. Out of 17 subjects, 4 subjects provided one incorrect answer each. Additionally, one of these subjects provided an illegimate answer: instead of ticking a box (with either ‘yes’ or ‘no’), this subject wrote ‘DK’, probably meaning ‘Don’t Know’.

If we exclude the illegimate answer from our results, the mean of the number incorrect answers for +G is 0.24 and for –G it is 0.55. In words, the direction of the correlation between +/–G and the ability to recall the content of the dialogue is as predicted in our hypothesis I. However, the computation of a *t* test on these data did not yield a statistically significant result ($t = 1.19$, $df = 26$, $p < 0.25$, $r = 0.23$ two-tailed). Note that for $r = 0.20$, we would need a sample of approximately 195 subjects to obtain a statistically significant result (power = 0.8, at 0.5 two-tailed), assuming our hypothesis is correct.

Gestures and level of engagement

In the +G condition, the mean for the level of engagement was 4.41, whereas in the –G condition it was 3.73. Again the direction of the correlation is as predicted in our hypothesis II: a dialogue with gestures is ranked higher with respect to how engaging it is, than the same dialogue without gestures. But again the result is statistically not significant ($t = 1.05$, $df = 26$, $p < 0.25$, $r = 0.20$ two-tailed). Note again that for $r = 0.20$, we would need a sample of approximately 195 subjects to obtain a statistically significant result (power = 0.8, at 0.5 two-tailed) , assuming our hypothesis is correct.

Discerning Politeness

Overall, in answer to questions 7 and 8, there were 49 correct answers and 7 incorrect ones. In the latter case, the subjects judged the politeness of the character different from the system setting for politeness. This result is highly significant ($\chi^2 = 30.02$, $df = 1$, $p < 0.001$, $r = 0.73$).

When we split the answers over +G and –G we obtain the following. For +G, 31 correct answers and 3 incorrect ones. For –G, 18 correct and 4 incorrect answers. Although the results are again in the predicted direction, they are statistically not significant ($\chi^2 = 1.07$, $df = 1$, $p = 0.30$, $r = 0.14$).

Comments from subjects: problems with speech synthesis

The last question we put to the subjects was to provide any comments. These have been listed in the appendix. One thing stands out: 7 subjects out of 17 in the +G condition complained about the quality of the speech synthesis (L&H TruVoice). For the –G condition, only 1 subject out of 11 complained about this.

Discussion

Although our results were consistent with hypotheses I (gestures improve recall/memory of dialogue content) and II (gestures improve the appreciation of the dialogue in terms of how engaging it is), i.e., the direction of the correlation was as predicted, the results were not statistically significant. We would need bigger sample sizes to find out whether the size effects that were observed are not random.

The judgements of the subjects regarding the politeness of the characters corresponded closely with the system settings for this parameter. However, the presence/absence of gestures did not significantly influence this judgement.

Quite surprisingly, subjects who watched the dialogue with gestures had much more complaints about the quality of the speech than those who watched the dialogue without gestures. Various explanations are possible. One might be that the contrast between characters with gestures and the unnatural speech was bigger than that between characters with no gestures and the same speech. The bigger contrast in the former situation might have led to more comments on the quality of the speech. In fact, one subject explicitly commented on the perceived contrast between the quality of the animations and the speech.

An alternative explanation could be that because the dialogues with gestures had more pauses for the onset and endings of gestures, these gave subjects more time to notice the deficiencies in the speech. Although the speech for both dialogues was identical in speed, the extra pauses caused a significant difference in the duration of the +G and –G dialogue. The former took about 1min 40s, whereas the latter took only 50s.

In follow-up experiments this difference in duration will need to be taken into account when testing hypotheses I and II again on bigger samples. It would be advisable to make sure that in the –G condition all pauses have the same duration as those in the +G condition. Otherwise, it would be difficult to tell whether any results were due to the presence/absence of gestures or to the fact that in one condition the dialogue contained significantly longer pauses than in the other.

Acknowledgements

Thanks are due to Judith Masthoff for providing the opportunity to carry out this study and the help with executing the experiment. This research is supported by the EC Project NECA IST-2000-28580. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

Appendix: List of Comments from Subjects

Note that comments on the speech pertain to the speech as synthesized by the L&H TruVoice TTS.

Subjects in –Gestures condition (n = 11)

1. No emotion shown in voices.
2. Strange double negatives in dialogue.
3. I have met and worked with this salesman.

Subjects in +Gestures condition (n = 17)

1. It was boring because the vocal intonation was all wrong. It would have be [sic] rating 1 except for humorous content.
2. Good animations, but voices could have been more human.
3. Voices were very monotone. Not interesting to listen to.
4. Computer generated voices sound really bad. It is hard to understand and boring.
5. Dodgey [sic] voices.
6. If I was the customer I'd have walked out much sooner.
7. Q9. Depends on what you mean by engaging? From the point of view of the ability of the computer to generate dialogue, then yes. Ordinarily – no.
8. The conversation was too computer sound. Should be more natural.
9. The speech wasn't very natural, and some of the answers did not really relate to the questions.
10. The video was quite comical which meant it was engaging.

References

E. André, T. Rist, S. van Mulken, M. Klesen, and S. Baldes (2000). The Automated Design of Believable Dialogues for Animated Presentation Teams. In J. Cassell, S. Prevost, J. Sullivan, and E. Churchill (Eds.), *Embodied Conversational Agents*. Cambridge, Massachusetts: The MIT Press.

Brigitte Krenn, Hannes Pirker, Martine Grice, Paul Piwek, Kees van Deemter, Marc Schröder and Martin Klesen (2002). Generation of multimodal dialogue for net environments. In *Proc. der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS'02)*, Saarbrücken.

Piwek, P. (2003). A Flexible Pragmatics-driven Language Generator for Animated Agents. In *Proceedings of the10th Conference of the European Chapter of the Association for Computational Linguistics (EACL03; Research Notes)*, Budapest.

M. Schröder & J. Trouvain (2001). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *In Proc. of 4th ISCA Workshop on Speech Synthesis*, Blair Atholl, Scotland.

M. Schröder & J. Trouvain (to appear). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. To appear in *International Journal on Speech Technology*, Special Issue following the 4th ISCA Workshop on Speech Synthesis.