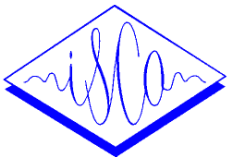

ITRW on Multi-Modal Dialogue
in Mobile Environments
Kloster Irsee, Germany
June 17-19, 2002

REALISM AND NATURALNESS IN A CONVERSATIONAL MULTI-MODAL INTERFACE

G. Power, R. I. Damper, W. Hall and G. B. Wills
Department of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK
{gp98r|rid|wh|gbw}@ecs.soton.ac.uk

Abstract

As computing becomes ever more pervasive in everyday life, new interface metaphors are urgently required for mobile and multi-modal applications. In this paper, we consider the issues of realism and naturalness in virtual 'talking head' characters. Specifically, we address the two questions: (1) What is the most appropriate degree of visual realism for a talking head, and does this vary with the degree of interaction? (2) To what extent should the naturalness of the synthetic speech match the realism of the talking head? Experiments are described that provide partial answers, by asking subjects to rate the interfaces on five attributes, as well as providing informal comments. Indications are that users prefer an intermediate level of visual realism, perhaps because this matches the underlying technology (animation, speech synthesis) best. Question (2) is very difficult to answer because of the difficulty of controlling naturalness in a synthesiser. Using three different TTS engines, we found that ratings across attributes varied with the synthesiser although average overall scores were very similar. Interestingly, subjects were not always aware when different synthesisers were being employed.

1. Introduction

The accelerating convergence of computer and communications technologies [Damper et al., 1994] together with ever-increasing use of these technologies by non-technical members of the public demand new interaction metaphors for mobile environments. One such metaphor gaining popularity is that of the conversational partner or 'agent', in which the device appears to the user as a virtual human, capable of understanding natural language and generating synthetic speech. In this paper, we focus on the 'talking

head' aspect of such interfaces [Marriott et al., 2000]. Although synthetic speech does not yet match the quality of natural speech [Duffy and Pisoni, 1992, Ralston et al., 1995, Olive, 1997], intelligibility scores for the best text-to-speech (TTS) systems approach that of human speech [Kamm et al., 1997] and can accurately manifest personality as well [Nass and Kwan, 2000].

There are many issues surrounding the appropriate use of speech technology in these new interfaces [Furui, 1995]. The goals of the present work are to answer the following questions:

1 What is the most appropriate degree of visual realism ('anthropomorphism') for a talking head, and does this vary with the degree of interaction?

2 To what extent should the naturalness of the synthetic speech match the realism of the talking head?

These are clearly crucial matters yet they do not seem to have received very serious attention to date. In this paper, we describe experiments aimed at providing answers, so as to advance the use of these new interfaces.

2. Character Design

The characters we developed use Microsoft Agent [Microsoft, 1999] and Haptek's Virtual Friend (Haptek, `http://www.haptek.com`). Microsoft Agent services support the presentation of software agents as interactive personalities within the Microsoft Windows environment. It allows developers to easily incorporate anthropomorphic conversational interfaces into software.

We initially wanted to script a Verbot (Verbot Virtual Personalities, `http://www.verbot.com`) to create our on-screen character. This, however, proved unsuccessful, as the scripting language it uses is very limited. The only way to enable communication between the Verbot and any other software is by using the command line, as the Verbot has been given the ability to load other programs. This is unfortunate as the Verbot has exceptional lip synchronisation and a wide range of facial expressions. As such, we decided to capture the frames from the Verbot software and animate them using the Microsoft Agent software.

There are no generally-accepted guidelines for creating conversational, anthropomorphic interface agents. However, Trower has proposed a set of guidelines for designing effective conversational software agents [Trower, 1997], which we have considered in the creation of our characters. For the first two experiments (see Section 3 below) we had three characters of increasing realism (Figure 1), A being a smiley face, B a cartoon character and C a realistic female character. After negative feedback from the first two experiments concerning the female character, which was described as being

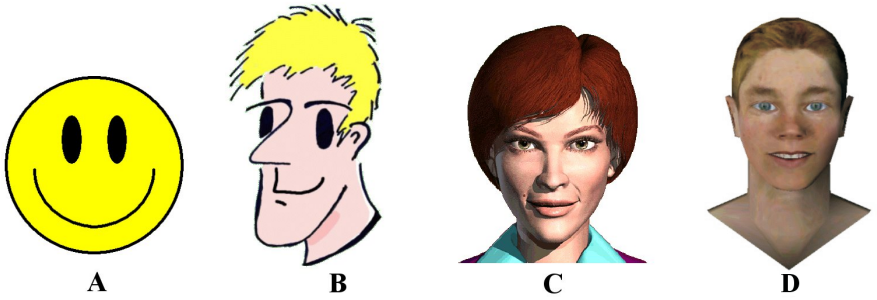

Figure 1. Visual appearance of the three characters used for initial experiments illustrating the range of realism employed: (a) smiley face; (b) cartoon character; (c) realistic female character; (d) realistic male character Erin.

scary by a number of subjects, the character was changed to a Haptek's Virtual Friend character called Erin, see D in Fig. 1. Erin used the Virtual Friend API, and it is a realistic smoothly animated a 3D character.

3. Effect of Degree of Realism

In this section we describe the three experiments we carried out in chronological order. For these experiments, the voices of characters A, B and C were implemented using the Microsoft TTS engine (voices Mike and Mary). However, the version of Virtual Friend available at the outset of this work did not allow the speech engine to be changed from that provided by Haptek.

3.1 First Experiment: Two-Way Interaction

For this experiment, we implemented software that works as an interface for a web-based search engine, using code we developed, that accepts Sherlock plug-ins (Apple Online, `http://www.apple.com/sherlock/`) which are used to parse results produced by various search engines. Two-way communication was allowed between the subject and the character in this experiment, see Table 1 for the subject profile. The ability of this system to target different sources was exploited to give different capabilities to the software. For example, if the subject asks the character to find a photograph, the software only targets archives that contain pictures. We restricted the software to a few simple functions: `find website`, `find picture` and `find links`. The `find website` function also targets Google but selects the most relevant website. For the `find picture` function, the software targets the Altavista image index. However, during a pilot experiment, we found that it didn't produce very good matches so we deactivated the `find`

Table 1. Subject profile for the two-way interaction experiment.

Number of subjects	11
Gender distribution	8 male and 3 female
Age distribution	Age group 18-30

`picture` feature. Having the other two functions enabled us to instruct subjects to carry out open-ended tasks.

The natural language parser can handle three types of inputs:

- Commands to find a set of links.
- Commands to find one page.
- Keywords.

The software parses the first two types of command, extracting search keys and forwards these to Google and waits for the results. A command that cannot be parsed by our software is sent to Netscape Search, which also accepts natural language queries.

The set of tasks required the subjects to search for and browse sites on:

- their academic/work interests;
- hobbies;
- music group(s) of their choice;
- favourite sport(s);
- a company or product of choice;
- favourite TV programme(s).

To reduce the chances that the subjects' opinion was influenced by the specific tasks they decided to carry out, their actions were restrained by the guidelines given to them. In addition, to ensure that the results were not influenced by the order in which tasks were undertaken, the order was randomised. Two of the six tasks were randomly allocated to each of the three virtual characters A, B and C (Fig. 1).

Task sheets were printed out and handed to the subject as required; when the subject finished a set of (two) tasks, the next would be given to them and the next character they had to interact with would be started. The order in which the characters were presented to the subjects was also randomised . In

this experiment, the characters used had a neutral expression throughout (not showing any emotion).

The subjects' opinions were recorded using a questionnaire, completed once they had carried out all the tasks as instructed. They were asked to rate each character using a seven-point scale for each of the attributes, 1 being the lowest (i.e., unhelpful) and 7 the highest (i.e., helpful).

Subjects rated a set of character attributes: friendly, pleasant, interesting, intelligent, helpful. These were decided upon by consideration of the traits thought central to the application—a simple anthropomorphic web-searching assistant. The first three attributes were deemed important to user-friendly interaction and the last two were considered to impact on fulfilling the users' needs and expectations. Subjects were given a brief definition of the five attributes in everyday language. Informal comments were also elicited.

To check whether or not the full complexity of the two-way interaction experiment was necessary to answer our requirements, two simpler experiments were designed. These will now be discussed.

3.2 Second Experiment: No Interaction

The second experiment consisted of completing a web-based form where subjects were asked to assess the characters just by looking at still pictures, i.e., no interaction. The helpfulness attribute was omitted as it was thought that it wasn't appropriate to judge how helpful a person or caricature was from a still picture. This experiment used 31 subjects.

3.3 Third Experiment: One-Way Interaction

The third experiment consisted of a lecture presented to undergraduate students (see Table 2 for the subject profile) by the three different characters A, B and D in Fig. 1. (Recall that Erin, character D, replaced female character C when the first two experiments revealed the imperfections of the latter.) The main part of the presentation was divided up into three parts. Part one took 12 minutes, part two took 7 minutes and part three took 5 minutes. The first part was presented by character A, part two by character B and part three by character D. It was decided that each part should take less time than the previous one to prevent the students from getting bored and irritated with the presentation. The attributes we examined for the characters were as in the first experiment.

For this experiment, characters had two different expressions: neutral and smile. The characters would speak with a neutral expression and smile from time to time. Character D is slightly different from characters A and B, in that Erin is rendered in 3D and has smoother animation compared to the other two.

Table 2. Subject profile for the one-way interaction experiment.

Number of subjects	79
Gender distribution	77 male and 2 female
Age distribution	Age group 18-25

Table 3. Normalised scores for friendliness in the three experiments.

	A	B	C	D
No interaction	0.94	0.71	0.57	–
One-way interaction	0.73	0.51	–	0.53
Two-way interaction	0.73	0.73	0.48	–

4. Results of Realism Experiments

In this section, we present the subjects' mean opinion scores for each character and each attribute of that character, for each of the three experiments. The issue of concern is how character realism affects the scores. The reader is reminded that as we go from character A through B to C or D, the level of realism increases. Where statistical significance is mentioned, this was tested using appropriate non-parametric methods [Siegel, 1956].

Friendliness. As a general finding, the friendliness rating decreases with the degree of realism of the character—the friendliest being the most abstract character, i.e., character A, then the cartoon, character B, and finally the least friendly character being the least abstract character, either C or D (see Table 3). For the no-interaction experiment, the differences observed between the average score for each character were statistically significant with a confidence of more that 95%. In the two-way experiment, characters A and B were rated equal and both were significantly friendlier than character C, with a confidence level greater than 90%.

Character A scored very highly on friendliness in the no-interaction experiment, viz. 0.94. Subjects scored this character less highly, average 0.73, during the two interactive tests (one-way and two-way). In the case of character B, the no-interaction experiment results in a score of 0.71, which decreased in the one-way-interaction experiment to 0.51, and then significantly increased by the two-way-interaction experiment to its highest value of 0.73. This pattern of the one- and two-way interaction test was frequently observed in the other characteristics evaluated.

The general finding that the realistic characters were deemed less friendly than the more abstract characters is in accordance with McCloud's and Thorison's ideas [McCloud, 1993, Thorisson, 1996]. A possible reason might be that the realistic characters might be intimidating subjects. This is supported by comments made by our subjects.

- The following comments were given about the female character C used for the first and second experiments (two-way and no interaction):

  was scary

  scares me

  is very scary!

  is SCARY!!!!!! She's staring at me. She has a pimple. I think she is laughing at me.

  C is evil

- Regarding male character D used for the third experiment (one-way interaction), the following comments were given:

  is far too scary . . .

  looks neurotic

We have considered a few possibilities for this. After the first two experiments (two-way and no interaction), we felt that it might be the particular character C that was 'scaring' the subjects: hence the choice of a different character for the third experiment (one-way interaction), when we switched to the Haptek character, Erin (D), who was much better animated and was three-dimensional. However, subjects still commented they found the realistic character 'scary'. We also considered that subjects might be expecting realistic-looking characters to move in a less cartoon-like fashion—more like humans. If this was the case, it would be improved by using the better animated character. The idea that subjects expect more of the more realistic characters can be seen in a comment about character D during the one-way interaction experiment:

> *doesn't behave like a real human, eye movements, you get confused*

However, even though the other two more abstract characters did not "behave like a real human", no comments were made about this as the subjects did not expect them to behave as such.

Pleasantness. The degree of pleasantness also decreased with the degree of realism. This pattern was observed in all three experimental conditions (Table 4).

Table 4. Normalised scores for pleasantness in the three experiments.

	A	B	C	D
No interaction	0.84	0.64	0.55	–
One-way interaction	0.63	0.51	–	0.49
Two-way interaction	0.74	0.69	0.49	–

Table 5. Normalised scores for interest of the character in the three experiments.

	A	B	C	D
No interaction	0.49	0.61	0.61	–
One-way interaction	0.45	0.52	–	0.50
Two-way interaction	0.47	0.55	0.58	–

Non-parametric statistical tests show that in the no-interaction experiment the differences between A and B and A and C were significant with a confidence higher than 95%, but the difference between B and C was not significant. On the other hand, in the one-way experiment, the differences between A and D and B and D were statistically significant with a confidence level higher than 95%, but the difference between A and B was not.

On reflection, we might expect pleasantness and friendliness to be closely related. This can be clearly seen in Tables 3 and Tables 4 as they follow similar patterns. Another similarity to the friendliness attribute is that in the one-way interaction experiment the subjects gave this attributes its lowest scores. However, the score is higher for the two-way interaction experiment, showing again that it might be better to not have any interaction with the characters if the only interaction possible is partial. Once again, the overall pattern of the scores is preserved across the different conditions.

We also considered that less realistic characters leave more open to the imagination and so are less likely to be disliked. McCloud argues that when viewers see cartoon characters they see a reflection of themselves [McCloud, 1993], in that cartoons are like an empty shell that enables us to not just watch the cartoon, but to "become it".

Interest. We have identified that the highly abstract character was classed as being less interesting to look at than the less abstract ones (Table 5). In this case, the scores for the different characters were very similar. However, as with other attributes, the users gave the interest attribute its lowest score in the one-way interaction experiment.

Table 6. Normalised scores for intelligence in the three experiments.

	A	B	C	D
No interaction	0.45	0.57	0.72	–
One-way interaction	0.44	0.56	–	0.60
Two-way interaction	0.45	0.58	0.65	–

The non-parametric tests showed that in the no and one-way interaction experiments the differences observed between the characters A and B and A and D were marginally significant with a confidence level of 85%. However, the difference between the characters B and D was not statistically significant. None of the results in the two-way experiment were found to be statistically significant.

Intelligence. The results in Table 6 apparently show that perceived intelligence increases with the degree of realism of the characters. It was in fact expected that subjects would see the more realistic characters as more intelligent, since they seem more human. This expectation is in line with the results of [Koda and Maes, 1996].

In spite of these results, and the way that they conform to our expectation, the statistical evidence of a significant effect is a little weak. In the one-way interaction experiment, the difference between character B and D was not statistically significant. However, the differences between A and B and A and D were significant with a confidence level greater than 95%. On the other hand, in the two-way interaction experiment, only the difference between characters A and C was even marginally significant with a confidence greater that 90%.

No significant effects from the degree of interaction were observed on the scores each character having almost the same score in each experiment except character C in the no-interaction experiment, where it scores its highest value.

Helpfulness. As mentioned earlier, this characteristic was not evaluated as part of the no-interaction experiment because of its inconsistency. The scores obtained in the two remaining experimental conditions failed to show any statistically significant differences. The obtained values were mostly just above 0.5. It appears that this characteristic is rather too subjective to give very meaningful results.

Before leaving this section, we mention that during the one-way interaction experiment, subjects were asked to choose a favourite character. The over-

Table 7. Normalised scores for perceived attributes for the cartoon character B with three different TTS voices.

	Microsoft/Mike	Digalo/Gordon	AT&T/adult male 1
friendliness	0.41	0.49	0.44
pleasantness	0.36	0.41	0.40
interest	0.33	0.47	0.37
intelligence	0.53	0.47	0.53
understanding	0.51	0.41	0.50
AVERAGE	0.43	0.45	0.45

whelming majority chose B, the cartoon character. That is, users seem to prefer a moderate level of abstraction in the virtual characters.

5. Effect of the Character's Voice

Having addressed question 1, we now turn to question 2, that is, to what extent should the naturalness of the synthetic speech match the realism of the character? This is a very difficult question to answer, since it is difficult to control the naturalness of the speech. In an attempt at least to vary the naturalness, we have evaluated three different TTS synthesisers with one of our characters—the cartoon character B.

The experimental design was largely that of the one-way interaction experiment in Section 3.3, using the same 79 subjects. However, the helpfulness attribute (which seemed not particularly useful) was replaced by ease of understanding (which seems highly relevant to the issue at hand). Character B read a different 5 min passage for each of three different synthesisers. These were: (i) Microsoft's TTS engine with (American English) voice Mike; (ii) Elan's Digalo with (British English) voice Gordon; (iii) AT&T Bell's TTS with adult male 1 voice (American). Table 7 shows the average scores obtained.

Non-parametric analysis of variance revealed that the score profiles of the three synthesisers were significantly different. On average, however, all three are rated very similarly. AT&T is rated better than Digalo on understanding and worse on interest. From the informal comments, we were surprised to discover that many subjects thought that two of the three voices (Microsoft and AT&T) were the same.

6. Conclusions

In this work, we have attempted to study the effect of visual realism of a 'talking head' character and the naturalness of its (synthetic) voice when the

character acts as an assistant in web-searching tasks. Experiments to explore the effect of visual realism led to the following conclusions:

- *The more realistic faces were seen as less friendly and even 'scary'*: We feel that the 'scary' impression is due to the character's behaviour not matching the subject's expectations of the human-looking face. That is, there is a mismatch with the underlying technology (e.g., animation, text-to-speech synthesis). We suggest that this technological barrier should be overcome with time.
- *The more abstract a character, the more friendly and/or pleasant it seems*: It is possible that less realistic characters leave more open to the imagination and so are less likely to be disliked.
- *The more abstract the character, the less 'interesting' it is to look at*: It is not surprising that subjects rate the simpler characters as less interesting than more complex ones.
- *Subjects favoured a moderately abstract character*: The cartoon character (B) is most subjects' favourite, even though this is not reflected in higher average scores than for the other two characters. It seems that its popularity is due to the fact that it scores moderately well on all attributes, as opposed to the other characters, which score very well on some attributes and very low on others.

Chronologically, the two-way interaction experiment was performed first. Thereafter, the degree of interaction was simplified to determine whether or not the full complexity of the earlier experiment was necessary to answer our questions. Indications are that broadly similar results are obtained for all three versions (no interaction, one-way, two-way). It should be possible to exploit this finding to simplify future experimentation.

Turning to the issue of the impact of the naturalness of the synthetic speech, it is very difficult to study this because of the difficulty of controlling naturalness. We attempted to do so by using three different TTS engines and found that ratings across attributes varied with the synthesiser although average overall scores were very similar. Interestingly, subjects were not always aware when different synthesisers were being employed. This may be because the subjects were focusing more on other aspects of the interaction than on the specifics of each character's voice, which would indicate that the voice may be relatively unimportant to the success of the interface.

7. Acknowledgements

This work was partly funded by the UK Engineering and Physical Sciences Research Council through a PhD studentship to author Guillermo Power. The

authors would like to thank Dr. Samhaa El Beltagy for help with programming issues and for providing the natural language parser used for the two-way interaction experiment.

References

[Damper et al., 1994] Damper, R. I., Hall, W., and Richards, J. W., editors (1994). *Multimedia Technologies and Future Applications*. Pentech Press, London.

[Duffy and Pisoni, 1992] Duffy, S. A. and Pisoni, D. B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*, 35(4):351–389.

[Furui, 1995] Furui, S. (1995). Prospects for spoken dialogue systems in a multimedia environment. In *Proceedings of European Speech Communication Association (ESCA) Tutorial and Research Workshop on Spoken Dialogue Systems: Theories and Applications*, pages 9–16, Vigsø, Denmark.

[Kamm et al., 1997] Kamm, C., Walker, M., and Rabiner, L. (1997). The role of speech processing in human-computer intelligent communication. *Speech Communication*, 23:263–278.

[Koda and Maes, 1996] Koda, T. and Maes, P. (1996). Agents with faces: The effects of personification of agents. In *Proceedings of Fifth IEEE International on Robot and Human Communication (RO-MAN '96)*, pages 189–194, Tsukuba, Japan.

[Marriott et al., 2000] Marriott, A., Beard, S., Haddad, H., Pockaj, R., Stallo, J., Huynh, Q., and Tschirren, B. (2000). The face of the future. *Journal of Research and Practice in Information Technology*, 32(3–4):231–245.

[McCloud, 1993] McCloud, S. (1993). *Understanding Comics: The Invisible Art*. Harper Perennial.

[Microsoft, 1999] Microsoft (1999). Microsoft Agent Software Development Kit.

[Nass and Kwan, 2000] Nass, C. and Kwan, M. L. (2000). Does computer generated speech manifest personality? An experimental test of similarity-attraction. In *Proceedings of ACM Conference on Computer-Human Interaction, CHI 2000*, pages 329–336, The Hague, Netherlands.

[Olive, 1997] Olive, J. (1997). The talking computer: Text-to-speech synthesis. In Stork, D., editor, *HAL's Legacy: 2001's Computer as Dream and Reality*, pages 101–131. MIT Press, Cambridge, MA.

[Ralston et al., 1995] Ralston, J. V., Pisoni, P. B., and Mullennix, J. W. (1995). Perception and comprehension of speech. In Syrdal, A. K., Bennett, R. W., and Greenspan, S. L., editors, *Applied Speech Technology*, pages 233–288. CRC Press, Boca Raton, FL.

[Siegel, 1956] Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Kogakusha, Tokyo, Japan.

[Thorisson, 1996] Thorisson, K. (1996). *Communicative Humanoids: A Computational Model of Psychosocal Dialogue Skills*. PhD thesis, Program in Media Arts and Sciences, MIT, Boston, MA.

[Trower, 1997] Trower, T. W. (1997). Creating conversational interfaces for interactive software agents. Tutorial at ACM Conference on Computer-Human Interaction, CHI'97. Available for download as file `twt.htm` from:

`http://www.acm.org/sigs/sigchi/chi97/proceedings/tutorial/`.