Anthropomorphic Vs Non-Anthropomorphic Software Interface Feedback for Online Systems Usage

Pietro Murano

University of Salford, Computer Science, School of Sciences, Gt. Manchester, M5 4WT, UK
p.murano@salford.ac.uk

Abstract This paper answers an important question concerning the effectiveness of anthropomorphic user interface feedback. The issue of effectiveness has been unresolved for some time, despite the efforts of various prominent computer scientists. By means of a carefully controlled tractable experiment, significant statistical evidence has been found to suggest that anthropomorphism at the user interface in the context of online systems usage is more effective than a non-anthropomorphic method of feedback. Furthermore, the results can be generalised to most software systems for online systems usage, thus potentially changing the way user interface feedback is designed, developed and thought about. This will lead to the improvement of user interfaces making them more usable, more effective and more accessible to all. Computer systems are being used more and more, where potentially every household and work environment will have a computer in the near future. Hence making systems accessible to all, including 'non-traditional' users is becoming increasingly more important. This research is making a contribution to this general aim.

1 Introduction

User interfaces using some element of anthropomorphism have attracted the attention of various computer scientists, where some have attempted to make conclusions on the effectiveness and user approval of such interfaces. Currently there is a dichotomy between computer scientists. Some are in favour of interfaces using some anthropomorphic element (e.g. Agarwal [1], Cole et al. [4], Dertouzos [5], Guttag [6], Koda and Maes [9], Maes [10] and Zue [17]) while others are against such interfaces (e.g. chapter by Shneiderman in [2] and [16]). However both sides do not provide concrete enough evidence to suggest that they are correct in their stance.

Certain individuals in favour of anthropomorphic interfaces (e.g. Koda and Maes in [8] and [9]) appear to begin with the premise that these are beneficial in nature, despite leaving the issue open concerning the appropriateness of using facial expressions at the user interface as expressed by Maes [10]. Others such as

Microsoft [2] are favourable to anthropomorphic interface feedback as can be seen by the Persona Project which uses a parrot to help a user find and play music tracks. They indirectly justify their use of such an interface by using the work of Nass et al [14]. However the work by Nass does not prove that anthropomorphism is more effective.

The main sceptic concerning anthropomorphism is Ben Shneiderman [16]. He argues that anthropomorphism is ineffective and mainly disliked by users. He uses evidence from various sources, e.g. Brennan and Ohaeri [3] and Resnik and Lammers [15], to support his stance. However the quoted works did not particularly look at the issue of effectiveness and user approval of such interfaces. These were mainly concerned with user behaviour issues [15] and user interaction in a social sense [3]. Furthermore, the feedback tested by these individuals was solely textual in nature thus not necessarily being representative of other anthropomorphic types of feedback such as 'faces'. Also the experiments by Brennan/Ohaeri and Resnik/Lammers used between users elements in the experimental design, which involved users not trying all types of feedback. It would have been more reliable to have a within users design where subjects tried all types of feedback with appropriate tasks. This is because despite randomisation, a group of subjects assigned to a particular interface feedback could have a confounding disparity compared with the other group(s) of subjects.

Recently the dichotomy mentioned at the outset of this paper has begun to be resolved particularly in the interface feedback setting. The first set of results from the experiment by Murano [12], concerning user approval issues of anthropomorphic interface feedback, show with clear statistical results that in the context of online systems usage, specifically UNIX commands, subjects preferred the anthropomorphic feedback. This was compared against a non-anthropomorphic feedback. Given a choice of the anthropomorphic or non-anthropomorphic feedback, a high proportion of subjects (63.64%) said they would have chosen the anthropomorphic feedback. Furthermore, certain subjects felt more secure with the anthropomorphic feedback. Some subjects expressed that they felt they could trust the information being given to them in an anthropomorphic manner and that they liked it.

Hence the results presented below concern effectiveness issues of such feedbacks and combining the user approval results from [12], conclude this particular experiment. Based on significant statistical evidence (presented below), in the context of online systems usage, specifically using UNIX commands, the anthropomorphic interface feedback was more effective than the standard non-anthropomorphic feedback. This is very useful and interesting for the overall aim of improving access to computer systems and for helping novices use computer systems they are unfamiliar with.

2 Online Systems Usage Experiment – UNIX Commands

2.1 Hypotheses

Answers to the following questions were the aim of this part of the experiment.

- Is a direct mapping (using video of a human as the direct mapping, i.e. anthropomorphism) of human-oriented information to software interface feedback effective? (effectiveness in this case was defined as the user achieving the tasks, the user achieving the tasks in as few incorrect attempts as possible and the user faltering as few times as possible.)
- Is an indirect mapping (using guiding text as the indirect mapping, i.e. non-anthropomorphism) of human-oriented information to software interface feedback effective?
- What guideline(s) can software interface feedback designers receive?

Furthermore, a null hypothesis (H_0) linked to the first two questions above was tested:

- There will be no difference between the 2 conditions (video and guiding text) - for effectiveness.

An alternative hypothesis (H_1) considered was:

- Video (anthropomorphic) feedback would be more effective than the text feedback (non-anthropomorphic).

Further, this experiment, where UNIX commands was the 'foundation', rested within the broader area or domain of software for online systems usage knowledge. The reason for this was to try if possible, to make a generalisation based on the results of this experiment, to cover the broader area of similar software.

2.2 Users

- All the users taking part in the study were adults.
- Males and females took part.
- All the subjects had differing personal backgrounds. This was taken to be individuals studying different courses, having varied birthplaces and having varied hobbies/work experience. This information was elicited by pre-experiment questions.
- Subjects were found through the university population.
- 55 users new to UNIX commands took part in the study.

2.3 Experimental Design

A within users design was used for this experiment. For the experiment the subjects tried a set of designed tasks. These were:

- Task 1 - The displaying of files in a current directory in long format.
- Task 2 - The displaying of the full path name of the current directory.
- Task 3 - Compressing a file showing the resulting percentage of file reduction.
- Task 4 - Deleting a given file.

Subjects used both types of feedback, i.e. the video and the guiding text condition. Two tasks would be given video as the attached feedback and the remaining two would receive the guiding text. The first two tasks were more simple (i.e. less key strokes and easier syntax) compared to the last two tasks which were deliberately made a little more complex (i.e. more key strokes and more difficult syntax). Furthermore these were typical UNIX tasks that a beginner could wish to

do. This was realistic because these were the types of commands the author began with, while learning UNIX with a community of students. For randomisation purposes the feedback conditions were rotated, so that one type of feedback was not always linked to a particular task.

2.4 Variables

The independent variables were the types of feedback, i.e.:

- Guiding text.
- Video.

To measure feedback effectiveness the dependent variables concerned observing for each task if the subjects completed a task and if so how many attempts it took to complete. User hesitation if present was also recorded. If user hesitation was recorded, the number of times it occurred was noted. The dependent measures used were by observation.

2.5 Apparatus and Materials

The equipment used for the experiment was:

- A PC running Windows 95, 400 MHz and 128 Mb RAM.
- External speakers.
- IBM ViaVoice Executive Automatic Speech Recognition (ASR) engine [7] (including text-to-speech), trained with a male English accented profile. A full training was the reading of 496 English phrases, predefined in ViaVoice. An English female profile was also obtained for use with female subjects (in practice the author obtained several profiles for having a better chance with voice matching issues).
- Head mounted microphone supplied with the ViaVoice kit.

The prototype was engineered with C++ Builder 3 and the ViaVoice Software Development Kit (SDK), and it was made to 'look like' a UNIX environment.
Running the prototype would present to the user a screen with a single X-Window containing a 'classic' UNIX type shell prompt. At all times an 'end program' button would be available at the top left part of the screen.

Allowing the system's learning algorithm [11] to take effect, would mean the software agent [11] would infer that the current user was a beginner and therefore present a smaller window to the left of the main X-Window asking the user what they wanted to do. The smaller window always contained a 'close' button in case the user wished to not have any feedback. The user would then input via the microphone the appropriate request for obtaining the required information for accomplishing the given tasks (described in Experimental Design section).

Upon a successful ASR, the prototype would display in the smaller window, relevant interface feedback (this was randomised so that one command was not tied to one type of feedback).

If the guiding text (non-anthropomorphic) was presented, the appropriate command would be displayed in the smaller window. If the video (anthropomorphic) feedback was presented, a video clip of a person verbally uttering the appropriate command would be played in the smaller window.

2.6 Procedure

The procedure described below was carried out in the same way for all subjects using the same equipment, questions asked and methods of observation. Each subject was treated in the same manner. This was all in an effort to control any confounding variables.

The experiment took about 30 minutes to complete per volunteer. Subjects were given £3 in cash as a reward, which they signed for, for their participation.

Each subject was booked an appointment during the day. Upon arrival the subjects were given a brief overview of the purpose of the research and then were asked a set of pre-experiment questions concerning the subject's personal background (see section 2.2).

A verbal introduction to the system itself was given, to help the subject overcome any false notions about the system. This included aspects of how to use the ASR module, e.g. to speak clearly and 'normally' (ASR was used to increase the level of usability). Each person was given an indication of the type of feedback that was being tested. Subjects were also briefed on the system behaviour, e.g. the sequencing of the screens involved in the interaction. Furthermore the subjects were assured that the aims of the experiment were to test the software and hypotheses concerning the software, and not to test the person. Each participant was also given a brief explanation as to what UNIX is, and the concepts surrounding UNIX commands. Furthermore in this explanation subjects were told where they would be typing the commands (i.e. which window) and how a command was to be concluded (i.e. by pressing the return key). A description of what they would see was also given, for the reason that a UNIX interface does not look like a regular PC interface. Each subject was also asked if they understood what each task meant and if they indicated they did not know, these would be explained clearly, usually in terms of the equivalent MS Windows operations (as all subjects had used MS Windows but not UNIX commands).

When the subject felt they were ready to start the experiment, they were given the head mounted microphone to put on. Upon running the program the subject would input (via the microphone) the appropriate feedback request for the first task. Upon a successful ASR, the appropriate UNIX command would be issued to the user, where the user would read the text command or view the video (depending on what feedback was issued). The user would then try to input, via the keyboard, the UNIX command the system advised the subject to use. Assuming the subject entered the command in the correct manner, they could then proceed to try the remaining tasks. If they failed/made mistakes in entering the command, they had the option of calling the appropriate feedback help to view the advised command again. This would go on until they achieved the correct command.

The number of attempts a user had in achieving the appropriate command was recorded. Issues causing wrong attempts were to do with syntax. If the user faltered/hesitated in an obvious manner, this was recorded based on observation of the user. A typical example of a user faltering was the user developing a strong puzzled expression at the moment of seeing the feedback and then trying to use the information to complete a task. A further example was a user verbally saying something to the effect of puzzlement. If a user seemed to get 'stuck', they were encouraged to review carefully the feedback they had already seen. If appropriate,

issues explained at the outset of the experiment (e.g. basic UNIX concepts) would be briefly reiterated. A few users had a lot of difficulty in achieving the tasks correctly. This was due to them not seeing the importance of reproducing the command exactly as they were given it even though they had been briefed to follow precisely the feedback given. It was the aim of the experiment for subjects to achieve the tasks on their own by using the feedback. Hence if too much prompting was given for achieving a task, it was deemed that the person had not succeeded in that particular task, while still recording the number of errors made.

Final results, e.g. the subject entered the correct commands upon having viewed the feedback etc. would be carefully recorded on an appropriate observation protocol sheet.

At all times during the experiment, the author observed the system behaviour and subject behaviour. For each task completed/not completed a score was assigned for use in the statistical analyses. The score for each task was based on a points system. For each task each subject (unknown to them) was started on 10 points. Each incorrect attempt resulted in the deduction of 1 point. Each hesitation observed, resulted in 0.5 points being deducted. If the task was completed, the score would remain as described. However if the task was not completed a further 1.5 points were deducted from the score. This scoring method was devised to represent fairly a subject's results, e.g. it was felt that each hesitation observed should carry less weight than an actual incorrect attempt, due to hesitation observation being potentially more subjective. Also each 'factor' described, was used to reach a single score because it was felt that they were closely linked together, e.g. an incomplete task could have been due to many hesitations. Confidence in the scoring system was also seen by plotting the scores on a normal distribution plot diagram, which showed the data to be approximately normally distributed.

2.7 Results

The data collected for the UNIX commands experiment was concerned with the effectiveness of the interface feedback (see [12]). The scores of all subjects were approximately normally distributed and were used in an F test and a t-test for the determination of feedback effectiveness.

For 55 subjects, combining the four tasks and comparing video over text, the t observed was 9.45, and the t critical (5 %) was 1.68. This is illustrated in Table 1 below:

Table 1. Comparison of Video Vs Text (t-test Over 4 Tasks)

	Comparison of Video Vs. Text (Over 4 Tasks)
t-Observed	9.45
t-Critical (5%)	1.68

For 55 subjects, combining the four tasks and comparing video over text, the F observed was 2.42, and the F critical (5 %) was 1.60. This is illustrated in Table 2 below:

Table 2. Comparison of Video Vs Text (F-test Over 4 Tasks)

	Comparison of Video Vs. Text (Over 4 Tasks)
F-Observed	2.42
F-Critical (5%)	1.60

For the 55 subjects in the combination of the 4 tasks comparing video over text, one subject's data showed an outlier in the normal distribution plot diagram for the video feedback. Hence the outlier was removed and the t-test and F test were conducted again for the remaining 54 subjects over the 4 tasks. The results were a new t observed of 10.21 with a t critical (5%) of 1.67. The new F observed was 5.68, while the F critical (5%) was 1.60. This is illustrated in Table 3 below:

Table 3. Comparison of Video Vs Text (t-test & F-test Over 4 Tasks With Outlier removed)

	Comparison of Video Vs. Text (Over 4 Tasks) Outlier Removed
t-Observed	10.21
t-Critical (5%)	1.67
F-Observed	5.68
F-Critical (5%)	1.60

For 55 subjects, combining tasks 1 and 2 and comparing video over text, the t observed was 4.14, and the t critical (5 %) was 1.68. This is illustrated in Table 4 below:

Table 4. Comparison of Video Vs Text (t-test Over Tasks 1 & 2)

	Comparison of Video Vs. Text (Tasks 1 & 2)
t-Observed	4.14
t-Critical (5%)	1.68

For 55 subjects, combining tasks 1 and 2 and comparing video over text, the F observed was 16.03, and the F critical (5 %) was 1.60. This is illustrated in Table 5 below:

Table 5. Comparison of Video Vs Text (F-test Over Tasks 1 & 2)

	Comparison of Video Vs. Text (Tasks 1 & 2)
F-Observed	16.03
F-Critical (5%)	1.60

For 55 subjects, combining tasks 3 and 4 and comparing video over text, the t observed was 5.67, and the t critical (5 %) was 1.68. This is illustrated in Table 6 below:

Table 6. Comparison of Video Vs Text (t-test Over Tasks 3 & 4)

	Comparison of Video Vs. Text (Tasks 3 & 4)
t-Observed	5.67
t-Critical (5%)	1.68

For 55 subjects, combining tasks 3 and 4 and comparing video over text, the F observed was 1.59, and the F critical (5 %) was 1.60. This is illustrated in Table 7 below:

Table 7. Comparison of Video Vs Text (F-test Over Tasks 3 & 4)

	Comparison of Video Vs. Text (Tasks 3 & 4)
F-Observed	1.59
F-Critical (5%)	1.60

Concerning the 55 subjects, for tasks 3 and 4 comparing video over text, one subject's data (same subject as for the combination of the 4 tasks above) showed an outlier in the normal distribution plot diagram for the video feedback. Hence the outlier was removed and the t-test and F test were conducted again for the remaining 54 subjects over tasks 3 and 4. The results were a new t observed of 7.27 while the t critical (5%) was 1.67. The new F observed was 4.21, while the F critical (5%) was 1.60. This is illustrated in Table 8 below:

Table 8. Comparison of Video Vs Text (t-test & F-test Tasks 3 & 4, With Outlier removed)

	Comparison of Video Vs. Text (Tasks 3 & 4) Outlier Removed
t-Observed	7.27
t-Critical (5%)	1.67
F-Observed	4.21
F-Critical (5%)	1.60

2.8 Conclusions

The results were studied firstly by combining the four tasks together in order to get an overall reliable conclusion. However it was also of interest and importance to look at the combination of tasks 1 and 2 and tasks 3 and 4 as two separate groups, as tasks 1 and 2 were 'simpler' tasks compared to tasks 3 and 4 (simpler is defined by the author as being fewer keystrokes and easier syntax). It was important to try and find out if one type of feedback was more effective for 'simpler' tasks while perhaps not being as effective for more difficult tasks (or vice versa).

Concerning the results for the combination of the 4 tasks, the t-test shows a large significance in favour of the video feedback (anthropomorphic). The F test conducted confirms the large significance in favour of the video feedback. Hence in this context, the video (anthropomorphic) feedback is more effective than the text feedback (non-anthropomorphic), over a span of various tasks of various complexity. Furthermore the significance in this conclusion is strengthened if one removes the outlier described in the previous section. This is because the t test and F test becomes even more reliable when outliers are removed and because the significance figures are much larger.

Significance can also clearly be seen for the combination of tasks 1 and 2. The t-test and F test both show a large significance in favour of the video feedback (anthropomorphic). Clearly the video is specifically more effective over 'simpler' tasks.

There is also significance in favour of the effectiveness of the video feedback for the combination of the 'more difficult' tasks 3 and 4. The t-test shows a large significance in favour of video for effectiveness. The F test is just reaching the 95% confidence margin, but is so close that between this score and the t-test score significance can be taken from the F test as well. The reason the F test score does not greatly exceed the 95% confidence interval is because the outlier of before has been left in the figures. If the outlier is removed, then both the t-test and F test dramatically exceed the 95% confidence interval, allowing one to conclude that the video was more effective for the 'more difficult' tasks 3 and 4.

These results show that the video feedback (anthropomorphic) was significantly more effective compared to the text feedback (non-anthropomorphic). This conclusion applies to 'simpler' and 'more difficult' tasks either combined or separated.

Therefore from the results, the null hypothesis (H_0) raised at the beginning of the experiment, can be confidently rejected. There was clearly a difference for effectiveness. Hence, the second hypothesis (H_1) can be confidently accepted as it hypothesised that the video would be more effective. Confidently one can suggest that video feedback is more effective in a UNIX commands situation, particularly where beginners to the environment are concerned. This result for this particular software domain and particular context within this domain has been missing in the current 'world knowledge' and will therefore add to and hopefully modify the current thought about interface feedback design.

From the discussion, one can generalise concerning these results. In the software domain of online systems usage knowledge, a direct mapping of human-oriented information to software interface feedback, such as video, is highly desirable. The overall suggestion is that this type of feedback would be more effective and

preferred by users [12]. Since all of the subjects for this experiment were complete beginners to UNIX commands, one should take this aspect into consideration. Due to this aspect being of importance it would be prudent to have such software with an alternative means of feedback. This is because once the basic syntax/format is learned in an environment such as UNIX commands, more experienced users are likely to not require the full explicit video feedback. A further aspect that could be considered is to have a more explicit video feedback for beginners and a less explicit video feedback for more experienced users, always with the option of perhaps having the text feedback available if wanted. Various other types of software within this domain (online systems usage) can be used as examples. Software such as word processors, painting packages, spreadsheet packages and interface navigation fall within this domain. The better systems of this type currently use 'tutorials' to help a user achieve systems usage knowledge by sometimes running a closed session showing the various buttons and menu options to be chosen to achieve a particular task. The idea is that then the user can remember or perhaps write down the steps to be followed. These though would be better served by the type of feedback (e.g. video) of this experiment, where a 'tutor' could 'walk' the person through the appropriate steps. This is used sometimes in teaching environments, where in the author's teaching experience, one or more beginners are 'walked' through the various steps for achieving a task, e.g. a mail merge in MS Word or creating combo boxes in MS Access. Clearly in this situation the beginner has the responsibility of practising the steps on their own in order to master them. With such a system online though, the beginner could call the appropriate feedback when required.

It is therefore recommended for designers of this type of software and interface feedback in this particular domain, to use a direct mapping of human-oriented information (anthropomorphism), such as video. It would also be prudent to include other types of feedback, perhaps similar to the current methods of using a closed session for illustration purposes. This would be with the idea of catering for the more experienced user who may not need to have a 'hand holding' feedback.

3 Concluding Remarks

As has been seen from the results and conclusions, the dichotomy concerning the effectiveness of anthropomorphism at the user interface has been resolved specifically for the domain of online systems usage. Hence the overall issue is on the way to being fully resolved by using the results presented in this paper and by using the results addressed by the overall research conducted by the author experimenting in other domains (e.g. [13] addresses the above issues in the context of English as a Foreign Language pronunciation). This has been made possible by designing carefully controlled tractable experiments which deal directly with the questions of effectiveness and user approval of such interface feedback.

Furthermore the results and conclusions add to the world knowledge of user interface feedback design/development, which had been deficient. This is always with the aim of enabling users to use computer systems more effectively with a higher level of satisfaction, where it has been seen that in the context discussed above the anthropomorphic user interface feedback was more effective than the standard non-anthropomorphic feedback.

4 Acknowledgements

The author would like to thank Computer Science at the University of Salford and all the willing volunteers who took part in the experiment.

References

1. Agarwal, A. Raw Computation. Scientific American. (1999), 281: 44-47.
2. Bradshaw, J. M. Software Agents, AAAI Press, MIT Press. 1997.
3. Brennan, S.E and Ohaeri, J.O. Effects of Message Style on Users' Attributions Toward Agents. CHI '94 Human Factors in Computing System,. (1994).
4. Cole, R., D. W. Massaro, et al. New Tools for Interactive Speech and Language Training : Using Animated Conversational Agents in the Classrooms of Profoundly Deaf Children. Method and Tool Innovations for Speech Science Education, London, (1999), Dept. of Phonetics and Linguistics, University College London.
5. Dertouzos, M. L. The Future of Computing. Scientific American. (1999), 281: 36-39.
6. Guttag, J. V. Communications Chameleons. Scientific American. (1999), 281: 42,43.
7. IBM, IBM ViaVoice 98 User Guide, IBM, 1998.
8. Koda, T. and Maes, P. Agents With Faces: The Effect of Personification. Proceedings of the 5th IEEE International Workshop on Robot and Human Communication, (1996), *IEEE*.
9. Koda, T. and Maes, P. Agents With Faces: The Effects of Personification of Agents. Proceedings of HCI '96, London, (1996), British HCI Group.
10. Maes, P. Agents That Reduce Work and Information Overload. Communications of the ACM. (1994), 37(7): 31-40, 146.
11. Murano, P. A New Software Agent 'Learning' Algorithm. People in Control An International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres, UMIST, UK, (2001), IEE.
12. Murano, P. Mapping Human-Oriented Information to Software Agents For Online Systems Usage. People in Control An International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres, UMIST, UK, (2001), IEE.
13. Murano, P. Effectiveness of Mapping Human-Oriented Information to Feedback From a Software Interface. 24th International Conference Information Technology Interfaces, Cavtat, Croatia, (2002).
14. Nass, C., Steuer, J. et al. Computers are Social Actors. CHI '94 Human Factors in Computing Systems – 'Celebrating Interdependence', Boston, Massachusetts, USA, (1994), ACM.
15. Resnik, P.V. and Lammers, H.B. The Influence of Self-Esteem on Cognitive Responses to Machine-Like Versus Human-Like Computer Feedback. The Journal of Social Psychology. (1986), 6 : 761-769
16. Shneiderman, B. Designing the User Interface - Strategies for Effective Human Computer Interaction, Addison-Wesley, (1992).
17. Zue, V. Talking With Your Computer. Scientific American. (1999), 281: 40,41.