Effectiveness of Mapping Human-Oriented Information to Feedback From a Software Interface

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

Abstract. *This paper describes an experiment which tests anthropomorphic user interface feedback against non-anthropomorphic user interface feedback. The experiment has considered the effectiveness and user approval of the anthropomorphic and non-anthropomorphic user interface feedback. The experiment is worthwhile, as the Computer Science community is not in agreement concerning effectiveness and user approval of anthropomorphic user interface feedback. The specific area or setting for the experiment has been English as a Foreign Language (EFL). The results from the experiment are very promising being statistically significant. They show that the anthropomorphic user interface feedback was more effective and preferred by users.*

Keywords. User interface, feedback, anthropomorphism.

1. Introduction

Improving user interface feedback is a general goal of Human Computer Interaction (HCI). This paper presents the results of an experiment comparing two different types of user interface feedback. The issues investigated were effectiveness and user approval of the feedbacks. The two types of feedback tested were anthropomorphic and non-anthropomorphic in nature.

An anthropomorphic user interface feedback usually involves the attribution of human behaviour or appearance or both to something which is not human. Alternatively it can be the direct representation of a human such as video [4]. A specific example is the Microsoft Office paper clip. A further example is a video of a human giving feedback in a special context [11], [12], e.g. EFL training. A non-anthropomorphic user interface feedback does not manifest any human qualities and is usually neutral. A specific example in a Web context could be a user trying to find a particular location on a location finder Web site. The site could present the user with a diagram or map of the relevant location. A further example could be a message box which appears on the screen stating that some 'illegal operation' has taken place and the user clicks 'OK' to close the box.

There has been a necessity to look at the effectiveness and user approval issues of anthropomorphic feedback at the user interface. This is because the issue has caused a division in the Computer Science community, where certain individuals are for anthropomorphism at the user interface (e.g. Agarwal [1], Guttag [6], Koda and Maes [9], Maes [10] and Zue [15]) and others are against anthropomorphism (e.g. Shneiderman in [5] and [13]). However regardless of each researcher's stance, each has not provided concrete enough evidence to support their conclusions.

Hence the experiment described below, which is part of a suite of experiments [11] [12], begins to resolve this issue. The experiment is in the domain of software for in-depth understanding. EFL pronunciation for Italians was chosen as the general experimental area and specifically correcting wrong sound reproductions. A prototype (described below) covering vowel and consonant phrases was developed with the 2 types of feedback available. The phrases were designed to cover typical errors Italian non-native English speakers could make.

2. Hypotheses

It was of interest and importance to find out if using video as the anthropomorphic user interface feedback would be effective and liked by users. This would be the direct mapping of human-oriented information to software interface feedback. Alternatively it was of equal importance to find out if using two-dimensional images with guiding text as the non-anthropomorphic user interface feedback would

be effective and liked by users. This would be the indirect mapping of human-oriented information to software interface feedback. Furthermore, it was hoped that the results could give software interface designers some useful guidelines for their development activities. As mentioned in the previous section, the experiment has been conducted in the EFL pronunciation area. This was specifically to correct sound reproduction problems when pronouncing words/phrases.

(Effectiveness in this case was defined as 'successful' self-correction of pronunciation. If a student took two attempts before a 'successful' self-correction for the video feedback on an exercise, compared to one attempt for the two dimensional images with guiding text feedback on another exercise, the latter would be classified as being more effective.)

Two null hypotheses (H_{0A}) and (H_{0B}) were tested:

- (H_{0A}) - There will be no difference between the 2 conditions (video and two-dimensional images with guiding text) - for effectiveness.
- (H_{0B}) - There will be no difference between the 2 conditions (video and two-dimensional images with guiding text) - for user preference.

Two alternative hypotheses (H_{1A}) and (H_{1B}) were also considered:

- (H_{1A}) - Video feedback will be more effective than two-dimensional images with guiding text.
- (H_{1B}) - Users will prefer the video feedback to the two-dimensional images with guiding text feedback.

Furthermore, this experiment, where EFL pronunciation was the specific context, rested within the broader area or domain of software for in-depth understanding, learning or educational software. This would hopefully allow one to make a generalisation based on the results of this experiment, to cover the broader area of similar software.

2.1. Users

The prototype used was designed and built with exercises for Italian users. Therefore, 18 Italian subjects from various regions of Italy were used. All the users were adult males and females with different backgrounds who spoke English with an Italian accent. Subjects had marginally differing standards of English. The subjects were found from the student population and by the subjects themselves suggesting others they knew.

2.2. Experimental Design

A within users design was used. This means that all subjects tried the same set of vowel and consonant phrase exercises developed. The exercises were designed to exercise typical language speaking areas where Italians tend to make errors when speaking English [8]. All subjects tried two methods of user interface feedback which was randomly allocated. Therefore one particular exercise phrase would not always be used in conjunction with one feedback method. The Anthropomorphic user interface feedback tested was video of an EFL tutor giving feedback on user pronunciation when an error was made by a subject. This was tested against a non-anthropomorphic user interface feedback. This consisted of two-dimensional images with guiding text giving feedback on user pronunciation when an error was made.

2.3. Variables

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

- Two-dimensional images with guiding text. These were professionally drawn images of facial cross-sections giving feedback on user pronunciation, where mouth shape and tongue position etc. were represented. The images were of the kind found in [2],[3],[14] and educationally approved of as shown in [2],[3],[14].
- Video. The video was of a real EFL tutor giving specific feedback on user pronunciation, similar in style to a class room setting.

Both types of feedback had in common the factor of giving very specific feedback to a subject concerning exactly where in a word they had erred.

To measure feedback effectiveness the dependent variables used were on a designed observation protocol, where this consisted of keeping a score of the amount of mistakes a subject made when attempting a particular exercise. For statistical purposes the scores given to each user's successful self-correction

consisted of three points for a successful self-correction at the first attempt. Two points for a successful second attempt, one point for a successful third attempt and no points if the subject failed on the third attempt. Therefore the user was given a maximum of three attempts per exercise.

To measure user preference to the feedback the dependent variables used were in a designed questionnaire, where users were asked to rate on a Likert scale their opinions (e.g. *Terrible* $_{1,2,3,4,5,6,7,8,9}$ *Excellent,* where 1 shows negative opinion and 9 shows the most positive opinion) of the helpfulness of the feedback modes they tried.

The dependent measures used were by observation, e.g. how many times (up to a maximum, of three times) did a subject need to view the feedback (video or diagrams with guiding text) in attempting an exercise and what were the subjects' opinions on the feedbacks and general interface (Likert scales).

2.4. Apparatus and Materials

The equipment used for the experiment was a PC with external speakers and microphone running Windows 95, 400 MHz and 128 Mb RAM. The ASR engine used was IBM ViaVoice [7] Executive (including text-to-speech), fully trained with a male Italian accented profile. A full training was the reading of 496 English phrases, predefined in ViaVoice. A female profile was also obtained for use with female subjects. The prototype was engineered with C++ Builder 3 and the ViaVoice Software Development Kit (SDK).

2.5. Description of User Interaction

This section will describe a typical user interaction during the running of the experiment. Therefore upon executing the program, the system presented a welcome screen immediately followed by a screen with a short textual introduction to the exercises.

The user was then presented with an options screen (composed of two buttons) where they chose from consonant or vowel exercises. When an option was chosen, an exercises screen would appear composed of buttons each containing an exercise. Furthermore, the feedback buttons would be visible, but 'greyed out' until an error was detected. From this point on one could 'toggle' to the other category of exercises, e.g. if consonant exercises had been chosen, then one could traverse to the vowel exercises without having to run the program from the beginning. Also there was at all times a button for ending the program.

The user would then start the exercises, which would activate the ASR engine. A dialogue box instructing the user to select an exercise would be displayed. Each exercise would then be activated by the user clicking the relevant buttons.

Having clicked on a button, the exercise phrase would appear in a text box and be read out by the text-to speech engine.

Having uttered the phrase the user would be acknowledged by the system text-to-speech as being correct, incorrect or not having been understood by the system. If the system did not 'understand' the user, a message to that effect would also be displayed graphically.

A correct result would mean the user could move to the next exercise. An incorrect result would make active the specific feedback methods. These could be selected by clicking on the appropriate buttons (during the experiment this was directed by the author who conducted the experiment).

Having acknowledged the feedback by clicking 'OK', the user could then have a further attempt at self-correction - up to a maximum of three attempts.

When the exercises were activated and being used, the user could ask for 'help', by uttering the word 'help', or click on the relevant help button. This would result in text-to-speech encouragement and a further reading of the phrase being exercised, e.g. assuming the user has said 'help', the system would reply by saying, 'Try it slowly or one word at a time. I'll repeat it for you - '. This would be followed by the relevant exercise phrase.

2.6. Procedure

This procedure was carried out in the same way for all subjects using the same equipment and questionnaires. Each subject was treated in the same manner. This was all in an effort to control any confounding variables.

The experiment took about one hour to complete per volunteer. Each subject was booked an appointment during the day. Upon meeting the subject they were given a brief

overview of the purpose of the research. Then they were asked to read a small paragraph (same for all subjects) of text in English, to informally ascertain the quality of their English (no subject had 'perfect' English).

A verbal introduction to the system itself was given, to help the subject overcome any false notions about the system. This was then followed by giving the subject a quick demonstration of how the system behaved.

When the subject felt they were ready to start the experiment, they were given the head mounted microphone to put on. The subject then proceeded to try out the exercises proceeding with the interaction as described in the previous section. If the system flagged the subject as having made a 'pronunciation error', the author directed the subject to the type of feedback they should try out. Once feedback was selected, the same feedback would be used for the question where the error happened, up to a maximum of three times.

At the end of the experiment the subjects were given a questionnaire to complete (see next section for the main categories covered by the questionnaire).

2.7. Results

This section presents the statistical results of the experiment and these are interpreted in section 2.8 below. The data collected was firstly concerned with the effectiveness of the interface feedback, and secondly with the user approval of the feedbacks. For effectiveness, the scores obtained for all subjects (see section 2.3 for scoring formula used) were plotted on a Normal Probability Plot which revealed the data to be approximately normally distributed. This gave confidence in the raw data's validity. The data were then used in a t-test for the determination of feedback effectiveness where the video had been tested against the diagrams and text, in relation to the stated hypotheses. For 18 subjects, the t observed was 2.14, and the t critical (5 %) was 1.74. This is shown in tabular format in table 1 below:

Table 1 - Comparison of video vs. diagrams and text (EFL experiment)

	Comparison of Video Vs. Diagrams and Text
t-Observed	2.14
t-Critical (5%)	1.74

There is a statistically significant difference between the effectiveness of the anthropomorphic and non-anthropomorphic Graphical User Interface (GUI) at the significance level of 5%. Thus the hypothesis (H_{0A}) is rejected.

User approval issues were determined by the scores allocated by the subjects in a post-experiment questionnaire, where the scale available was from 1 to 9 – 9 being the most positive score one could allocate. The means of the user scores were calculated along with the relevant standard deviation. This is shown in tabular format in tables 2 – 5, where each table heading was a separate category in the questionnaire. Furthermore, each specific question in each category is reflected on the left most column of each table.

Table 2 - Overall user preferences (EFL experiment)

	Overall User Preferences	
	Mean	Standard Deviation
Video	8.17	1.10
Diagrams and Text	7.11	2.17

Table 3 - General interface issues (EFL experiment)

	General Interface Issues	
	Mean	Standard Deviation
Colours Used, Comfort	8.39	0.92
Text, Readability	8.22	1.48
Text, Understandability	7.83	3.94

Table 4 - Buttons (EFL experiment)

	Buttons	
	Mean	Standard Deviation
Labelling Quality	8.22	1.00
Layout Clarity	8.44	0.92
Consistency	8.44	0.85

Table 5 - System usability (EFL experiment)

	System Usability	
	Mean	Standard Deviation
Ease of Use	8.06	0.99
Logic in Sequencing of Screens	8.44	0.78
Relevance of Information Given During Interaction	8.28	1.23
Clarity of Language Used by Pronunciation Assistant	8.28	1.28
Effort Required on User's Memory	8.00	1.71
Clarity of Exits from the Program	8.00	1.24
Helpfulness in Improving Pronunciation	8.17	1.04

Each mean score was in the high positive 7 and 8 regions, where 9 was the best score a particular interface feature could receive. Most

of the standard deviations further reveal consistency in the user allocated scores.

Finally, in response to a general preference question on the questionnaire, 95% of subjects felt they would like to use the system on their own and felt that using the system would improve their confidence in English.

2.8. Conclusions

The result of the t-test regarding the comparison of the video interface feedback (anthropomorphic) with the diagrams and guiding text interface feedback (non-anthropomorphic) shows that there is statistical significance in the results. The level of successful self-corrections was higher with the video interface feedback than with the diagrams with guiding text.

Users rated highly both types of interface feedback (Table 2). However the video conditions received higher more consistent average scores. The standard deviations show the greater consistency concerning user responses to their preference of the video feedback. Not reflected in the scores allocated by subjects, was the verbal feedback received at the end of the experiment. Users spoke more positively of the video feedback than for the other condition. During the experiment some subjects while being directed to the two-dimensional image and guiding text condition, specifically asked to play the video feedback. Further it was noticed that one of the subjects would pay attention only to the video feedback. If presented with the diagram and guiding text condition they appeared to make no effort at reading and applying the advice given via the diagram.

The results for the general interface features which appeared in all experimental conditions (Tables 3-5) show that the prototype was in a highly usable state, meeting with user approval. Subjects rated highly all aspects of the interface as can be seen from the mean scores. Furthermore, the standard deviation results show in most cases (except for the Text - Understandability ratings) a good degree of consistency in the mean scores, showing that subjects were generally in agreement with each other. These results give confidence that good interface design practice was employed throughout the development stages. Further, the highly usable interface will not have tainted the results concerning the effectiveness and user approval issues of the feedbacks. The reason for this is that, if the general interface was badly designed, one of the feedbacks could have been seen as providing more effective results or having more user approval. This could come about if e.g. incorrect use of colour was used, where one of the feedbacks was difficult to read or view compared to the other. This would probably result in that feedback being less effective or disliked by the subjects. Although this is a simple example, it clearly shows the many possible problems that one could have with a badly designed interface in such an experiment.

Therefore, from the results obtained, the two null hypotheses (H_{0A} and H_{0B}) raised at the beginning of the experiment can be rejected. There was a difference in the results with respect to effectiveness and user approval. The video was more effective than the diagrams with guiding text. Furthermore, users on the whole preferred the video interface feedback. Thus the second hypotheses (H_{1A} and H_{1B}) can be confidently accepted. It is also interesting and perhaps significant that there is a relationship between the more effective interface feedback and user preference. This raises the suggestion that users probably quite quickly ‘sense’ what helps them most or makes things easier for them and therefore are more positive towards feedback that helps them to achieve effectively their tasks. This idea is supported by the fact that software packages with bad on-line help systems tend not to have their help systems used very often.

The generalisation concerning the results is that in the context of software for in-depth understanding, learning or educational software, an anthropomorphic user interface feedback is desirable. This is with particular reference to video and the suggestion that it would be more effective. Software packages which might benefit from these findings could cover other aspects of language learning, e.g. reading, grammar, improving one’s accent and other language groups. The packages would not need to be confined to language learning, but could cover other areas of learning, e.g. PC building and mechanics. Extending the example of PC building and using the results of this experiment, it is likely that a course composed mainly of video modules (involving a human tutor) would be more effective and preferred by users, than a textual/diagrammatic construction

manual. This is due to the fact that observing a human building a PC and listening to that person's advice, e.g. in avoiding some pitfall, will permit the learner to observe actual human hands and their co-ordination. A course could be devised in such a way as to allow one to build a PC on-line and receive anthropomorphic feedback (e.g. video of tutor) on the process.

The author would recommend to software interface designers of software for in-depth understanding, learning or educational software to include anthropomorphic user interface feedback, especially video, as these are likely to give users better results. Other methods of feedback can also be included such as diagrams with guiding text (depending on the context involved). However, it would not be recommended to develop such systems with solely non-anthropomorphic feedback.

One of the limitations concerning this experiment is that it would have been desirable to conduct the same experiment on a much more long-term basis with the same subjects. This however was not possible due to time constraints, but more so due to the fact that the majority of the subjects used were in the United Kingdom (U.K.) for a short period of time.

3. References

[1] Agarwal, A. Raw Computation. Scientific American. 1999, 281: 44-47.

[2] Baker, A. (1981). Ship or Sheep? An Intermediate Pronunciation Course, Cambridge University Press.

[3] Baker, A. (1998). Tree or Three? An Elementary Pronunciation Course, Cambridge University Press.

[4] Bengtsson, B., Burgoon, J.K. et al. The Impact of Anthropomorphic Interfaces on Influence, Understanding and Credibility. Proceedings of the 32nd Hawaii International Conference on System Sciences, 1999. IEEE.

[5] Bradshaw, J. M. Software Agents, AAAI Press, MIT Press. 1997.

[6] Guttag, J. V. Communications Chameleons. Scientific American. 1999, 281: 42,43.

[7] IBM, IBM ViaVoice 98 User Guide, IBM, 1998.

[8] Kenworthy, J. Teaching English Pronunciation, Longman. 1992

[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] Shneiderman, B. Designing the User Interface Strategies for Effective Human Computer Interaction, Addison-Wesley, 1992.

[14] Ur, P. 1996 A Course in Language Teaching - Practice and Theory, Cambridge University Press

[15] Zue, V. Talking With Your Computer. Scientific American. 1999, 281: 40,41