Proceedings of the Institute of Acoustics

# TOWARDS A USABILITY MEASURE FOR AUTOMATED TELEPHONE SERVICES

S Love (1), R Dutton (1), J C Foster (1), M A Jack (1), I A Nairn (1), N Vergeynst (1) and F W M Stentiford (2)

(1) Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh
(2) BT Laboratories, Martlesham Heath, Ipswich

## 1. INTRODUCTION

Usability can be defined as 'a concept comprising the effectiveness, efficiency and satisfaction with which specified users can achieve specified goals in a particular environment' [1].

Increasingly in the development of new technology, the attitudes and perceptions of potential users are regarded as important considerations which have to be taken into account in the design and alteration of systems and services. Users can provide useful information, indicating where the strengths and weaknesses of a system lie. The main problem is that these attitudes and perceptions are difficult to measure. Evidence suggests that there is a diversified approach to developing methodologies for measuring usability. However, in [2] Poulson argues for a general purpose measuring tool (a questionnaire) which can be used to assess the perceived usability of different systems in various settings. The methodological approach reported here for the measurement of usability of automated telephone services is similar to Poulson's in one respect - namely the use of a questionnaire. However, the work in addition, employs a field experimental approach.

The present paper describes the experimental approach in detail, paying particular attention to its value for usability research. It then goes on to describe a usability metric developed specifically for the evaluation of telephone-based speech interfaces and based on a detailed postal questionnaire. Issues of validity, reliability and sensitivity with respect to the questionnaire are discussed.

## 2. EXPERIMENTAL DESIGN

All experimental designs are concerned with the ways in which one or more dependent variables may be affected by one or more independent variables. In the case of automated telephone services, the dependent variable is the perceived usability of the automated telephone service; the independent variable is the performance level of the speech recogniser. In the experiments reported here, the accuracy level of the recogniser is under the control of the experimenter and can be varied to determine how recogniser performance affects the perceived usability of the system.

Laboratory experimentation offers an opportunity to define and measure both the independent and dependent variables. It also reduces the possibility of uncontrolled or extraneous variables having an effect on the dependent variable at the expense of the independent variable. The major drawback of using a laboratory experiment approach in usability research is that there is a danger that effects might not generalise because of the artificiality of the conditions. The laboratory can be an intimidating place and can never fully simulate the environmental conditions in which individuals would use an automated telephone service.

Consequently, the usability research paradigm described in this paper uses an approach which may be characterised as field experimentation. By conducting research in the user's natural environment, with independent variables still under the control of the experimenter, it is possible to obtain results which can be generalised to real situations. One effect is that, because of the familiarity of their surroundings, subjects may experience less anxiety when taking part in the experiment.

One drawback of this approach is that the experimenter has limited control over extraneous variables. The work reported here takes into account the possibility of environmental noise affecting the subject's performance when considering usability scores.

The results reported later in this paper are from matched-pair experiments. These involve matching subjects in one experimental condition (for example, one level of simulated speech recognition accuracy) with another experimental group (for example, a different level of speech recognition accuracy) in terms of relevant variables which are determined by the nature of the research being undertaken (for example age, sex, and geographical region). The main advantages of adopting this design include the fact that there are no ordering effects (performing better on a second task due to the practice effect obtained by having carried out the first task) and that subject variables can be partly controlled through matching the same stimulus (i.e. levels of recogniser) on both groups. The main problems with adopting this design are that it is difficult to find perfect matches and the loss of one member of the pair means the loss of the whole pair when it comes to analysing the data.

A decision also has to be made concerning the experimental setup to be used within the exper-

imental design. For the purposes of the research reported here, the evaluation of an automated telephone service, a Wizard of Oz simulation methodology was used.

Most of the Woz studies of human-machine interaction reported in the literature [3] have focussed on the problem of characterising dialogue features in application domains such as rail and timetable enquiry services. The Wizard of Oz system reported here is distinguished from most previous studies by the choice of highly constained application domains; by the degree of control the software provides over the experimental variables; by the care being taken to quantitatively measure users' attitudes and perceived usability; and by the large subject samples used in the experiments. One of the major new features of the work is that it is based on a realistic simulation of an available speech recognition technology allowing experimentation with recognition performance levels extrapolated beyond those currently available. This allows the experiments to address, among other key issues, the shape of the usability function for automated telephone interfaces for different levels of recognition performance. The details of the Wizard of Oz experimental setup and the simulation of the speech recognition system used in the present research have been reported elsewhere [4,5].

## 3. USABILITY METRIC

In designing an attitude measurement tool such as a questionnaire there are several important issues which must be addressed. In an attitude scale there should be an equal number of positive and negative statements relating to the system and these items should cover a wide range of attitudes. Such balance overcomes the danger that the overall score could reflect the users' tendency to agree rather than diagree with the questionnaire statements (an effect known as 'response acquiescence set') instead of providing valid information on usability. Furthermore, the attitude scale should provide an overall score for usability. However, in addition to estimating the significance of the overall score, it is important to examine the scores for individual items, since these may highlight specific problems relating to the usability of the service. There should also be a series of open-ended questions and a general comments section included in the questionnaire to allow users to express opinions and perceptions not covered by the attitude scale. Including these provides a rich source of qualitative data which can augment the quantitative data provided by the attitude scale. For the particular measurement tool reported here a seven-point Likert scale with a mid-neutral point has been used.

Three important aspects to be considered in the development of a measurement tool are its reliability, validity and sensitivity. Reliability can be defined as the extent to which a test accurately produces the same results on different occasions and validity could be defined as the extent to which a test actually measures what it claims to measure [6].

The most straightforward way of assessing reliability is the Test-Retest method. This involves administering a particular test to sets of subjects on more than one occasion. Using the Pearson-Product Moment Correlation Coefficient calculated on the data collected provides a 'reliabilty

coefficient' [7], and the closer this is to unity the greater the reliability of the test.

There are also several ways of investigating the validity of a usability metric. One of the categorisation systems used is known as content validity. This involves evaluating the content of the test to ensure that it is representative of the area it is supposed to cover. The measurement tool presented here is representative of a broad concept such as usability since the underlying uses comply with the work described in [1] and [2]. Also related to the concept of validity is the issue of external validity. This implies that the subjects used in any experiment should be representative of the general user population since this will allow generalisation of any significant research findings to the the population at large. This procedure highlights the importance of sampling in experimental work. In general it is considered that the ideal sample size when investigating an independent variable which is regarded as having a medium sized effect is 30 [8].

## 4. QUESTIONNAIRE VALIDATION

The first experiment reported here aimed to examine whether the user evaluation metric (the questionniare) predicted increased user satisfaction and usability when the performance of the simulated recogniser in an automated telephone service speech interface was set at two different levels. By adopting a matched-pairs design it was hypothesised that this would help to give an indication of the reliability and validity of the measurement tool.

Two subject groups each of 20 subjects were drawn from a subject panel and matched for sex and age group. Each matched pair of subjects was given the same task of reading a 16 digit credit card number one digit at a time to the automated telephone service. In each matched pair, one subject experienced the speech recogniser set to 90% accuracy, the other experienced 100% speech recognition accuracy.

Subjects were posted a cover letter explaining the experiment, a replica credit card, a questionnaire and a reply-paid envelope for returning the questionniare. They were telephoned by a WoZ operator at times which had previously been arranged with each of them. The operator read out a preamble which greeted the subjects and prepared them for their task in the experiment. The subjects were then told that they were being put through to an automated service and the WoZ system was activated. The operator keyed in the subject's responses which were categorised as digit, yes/no or a reject (i.e. an unrecognised utterance). If there were no response when one was expected, this was recorded automatically as a silence after a period of 5 seconds had elapsed.

After subjects had reached the end of the dialogue they were put back to the operator who told them that in a real situation the number they had just given would be read back to them and confirmed. They were asked to fill in the questionnaire and return it as soon as possible.

Since the data returned by a Likert scale is regarded as being ordinal, it was appropriate to use non-parametric statistics. A Wilcoxon Signed-Ranks test [9] was performed on the attitude scores obtained from the two experimental groups and a significant difference was found($p < 0.05$). A profile showing the separation of responses to the Likert attitude measures was also constructed, part of which is illustrated in Figure 1.

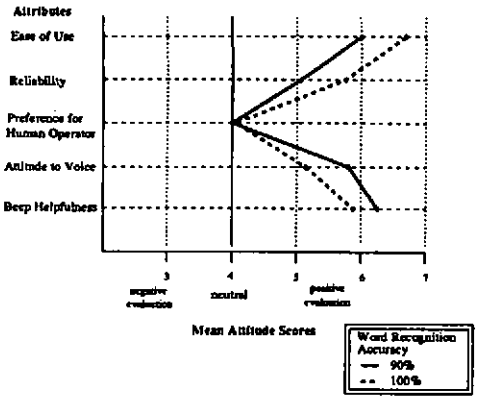


Figure 1: Sample Usability Profile

This provided a valuable source of information regarding which aspects of usability were affected by the manipulation of the simulated recogniser settings. In addition, the profile gave a visual representation of the shift in attitude perception which was confirmed by the Wilcoxon analysis. This was an important finding in many ways. It suggested that not only could the Likert attitude questionnaire discriminate significantly between two levels of the simulated recogniser, it also provided evidence that the metric had reached a certain degree of reliability.

The question of the validity of the content of the questionnaire was addressed by conducting a Pearson-Product Moment Correlation on the 22 variables in the questionnaire. The results of this analysis indicate that the variables were well correlated, suggesting that the content of the questionnaire was well disposed to assessing users' perceptions and attitudes to usability.

## 5. QUESTIONNAIRE SENSITIVITY

A second experiment was conducted to evaluate the sensitivity of the questionnaire for measuring user attitude to the automated telephone service at four levels of the recogniser. By using four different recogniser settings and having a large sample size in each experimental group, it was expected that further evidence of the reliability and validity of the usability metric would

be obtained.

The experiment used 256 subjects divided into four experimental groups (subjects matched for sex and age group) with simulated recogniser set at 85% (group 1), 90% (group 2), 95% (group 3), and 100% (group 4). Each subject grouping had the same credit card number across the four experimental settings of the recogniser.

A Friedman Two-Way Analysis of Variance (a non-parametric test) [9] was performed upon the sum scores of the Likert attitude measures for each subject across the four experimental groupings. A significant effect was found for differences in perceived usability due to the recogniser level ($p < 0.001$). This effect was explored by conducting a series of Wilcoxon Signed Ranks tests on each pairing of the four experimental groups. No significant differences were found between group 1 and group 2, group 1 and group 3, and group 2 and group 3. However, significant differences were found between group 1 and group 4 ($p < 0.001$), group 2 and group 4 ($p < 0.001$) and group 3 and group 4 ($p < 0.001$). This effect is also apparent from a comparison of the profiles of the mean responses for each of the 22 variables across the four experimental settings.

A Pearson-Product Moment Correlation test was performed to measure the significance of the correlations between the subjects scores on the 22 variables within each experimental condition. For each condition, their was evidence to suggest a highly significant amount of correlation between the variables. A series of Wilcoxon Signed Ranks tests and Related T-tests [7] were performed on pairings of variables across the four experimental groups. Both these parametric and non-parametric sets of tests returned exactly the same set of significant pairings.

## 6. CONCLUSIONS

The results of the two experiments reported above show that it is possible to construct an experimental setup and measurment tool which can reliably and validly assess users' attitudes and perceptions towards a simulated automated telephone service. There must however be an ongoing assessment and evaluation of this usability metric. As it stands it can provide us with an overall figure for usability, but of equal importance is the ability to breakdown usability into a series of subcategories and assess how each of these contribute to the users' overall perceptions and evaluations. An example of a process of this kind is widely available in the literature concerning the development of psychometric tests to measure intelligence. If intelligence can be defined and measured, the evidence suggests that it has to be assessed by a test which involves a series of diversified subtests with the individual's performance in each of these combined to give an overall score for intelligence - the Wechsler Adult Intelligence Scale [10] being one of the most well-known.

The development of equivalent diversified subtests for usability assesssment requires the preliminary investigation of subconstructs within the general domain of usability. One approach

to this could be through the application of factor analysis to the data made available in the experiments described earlier.

Factor Analysis [11] is a statistical method which looks for "clusters" of variables which are statistically related. The method attempts to explain the relationship between test results. Factor analysis does, however, require the intuitive naming and interpretation of the factors which have been statistically identified. It does not prove the existence of factors; rather, it provides supporting evidence which may allow us to claim, in this particular instance, that usability can be organised in a particular way.

## 7. ACKNOWLEDGEMENTS

The authors wish to acknowledge the support for this research from BT's Strategic University Initiative and the contributions made by other members of the Dialogues for Systems team at BT laboratories (Martlesham Heath, Ipswich).

## 8. REFERENCES

[1] INTERNATIONAL STANDARDS ORGANISATION, 'Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)', ISO CD 9241-11, (1990)

[2] D POULSON, 'Towards Simple Indices of the Perceived Quality of Software Interfaces', in IEE Colloquium - Evaluation Techniques for Interactive System Design. IEE, Savoy Place, London, (1987)

[3] N M FRASER & G N GILBERT, 'Simulating Speech Systems', *Computer, Speech & Language*, 5, pp81-99, (1991)

[4] J C FOSTER, R DUTTON, S LOVE, I A NAIRN, N VERGEYNST & F W M STENTFORD, 'Intelligent Dialogues in Automated Telephone Services', to appear in Proceedings of the workshop on Interactive Speech Technology, NEC, Birmingham, Ergonomics Society (1992)

[5] M A JACK, J C FOSTER & F W STENTIFORD, 'Intelligent Dialogues in Automated Telephone Services', to appear in Proceedings of the International Conference on Spoken Language Processing (ICSLP 92), Banff, Alberta, Canada (1992)

[6] J RUST & S GOLOMBOK, 'Modern Psychometrics: The Science of Psychological Assessment', Routledge, London, (1989)

[7] D C HOWELL, 'Statistical Methods for Psychology', PWS Publishers, (1987)

[8] H COOLICAN, 'Research Methods and Statistical Methods in Psychology', Hodder & Stoughton, London, (1990)

[9] S SIEGAL & N J CASTELLAN, 'Nonparametric Tests for the Behavioural Sciences', McGraw-Hill International Editions, (1988)

[10] D W WECHSLER, 'The Measurement and Appraisal of Adult Intelligence, 4th Edition', Williams & Wilkins, Baltimore, (1958)

[11] D CHILD, 'The Essentials of Factor Analysis', Cassell Education Ltd, (1990)