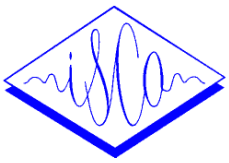

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

IDENTIFYING USABILITY ATTRIBUTES OF AUTOMATED TELEPHONE SERVICES

R. T. Dutton [1], J. Foster [1], M. A. Jack [1] and F. W. Stentiford [2]

[1] *Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh, EH1 1HN, Scotland, UK*
[2] *BT Laboratories, Martlesham Heath, Ipswich, IP5 7RE*

ABSTRACT

This paper reports on research involving a series of large-scale field experiments using a new Wizard of Oz (WOZ) system for the investigation of users' attitudes towards automated telephone services. The paper focuses on the identification of a service's usability attributes, those features and characteristics of a system which influence the effectiveness, efficiency and satisfaction with which specified users can achieve specified goals in a particular environment. The attributes which are found to be most salient are used to form the content of a questionnaire designed to measure attitudes towards usability of automated telephone services. Experimental results derived by using such a questionnaire are discussed.

KEYWORDS : usability attributes, automated telephone services, speech interfaces, Wizard of Oz, questionnaires.

10.21437/Eurospeech.1993-312

1 INTRODUCTION

This paper reports research into one key aspect of the usability evaluation of automated telephone services: the identification of user-perceived salient attributes of such services. These attributes represent the features and characteristics which will ultimately determine the effectiveness, efficiency and satisfaction with which users can achieve their goals when using automated telephone services.

Results are presented which relate to a series of large-scale field studies of simulated automated telephone services. These studies involve the use of the Wizard of Oz technique. Particularly important features of these studies include the use of large subject samples, realistic tasks carried out in natural user environments, and the use of telephone and written questionnaires to evaluate users' attitudes towards the automated services. These questionnaires are also used to elicit salient attributes.

Some twenty attributes are identified and evaluated in the experiments. Analysis of the results shows that a small number of these twenty attributes are highly salient to users. It is recommended that these highly salient attributes should be given priority consideration in the engineering design of new automated telephone services. The paper concludes with a description of how the identified attributes can be used to develop valid and reliable evaluation tools and appropriate usability metrics.

2 THE WOZ EXPERIMENTAL WORKBENCH

The main feature of the Wizard of Oz (WOZ) method is that a human (known as the 'wizard' or 'accomplice') simulates all, or part of, the system output. The essential requirement of any WOZ experiment is the deception involved in making subjects believe that they are interacting with a real computer-operated system.

The WOZ system used in this research uses a parametric simulation of a speech recogniser. The simulator enables the investigation of telephone service usability issues which incorporate speech recognisers with accuracy better than current state-of-the-art systems. The role of the experimental operator (the Wizard) is reduced to keying in users' responses. The Wizard is not required to make recognition decisions.

The WOZ experiments use large groups of subjects who have been quota-sampled from the UK population by a professional market research company. Experiments adopt matched-set design methodologies, matching subjects for age group and sex. A particularly important feature of the work is the use of realistic tasks to be carried out by subjects in their home or work place.

A typical task involves subjects entering their personal identification number, ordering an item from a catalogue, and then making payment through a credit card account. To complete such a task successfully, subjects are required to speak a range of different numbers and reply to system

requests for confirmation.

3 IDENTIFICATION OF ATTRIBUTES

An initial set of usability attributes was identified from a pilot study involving observation studies, interviews with naive users, and a review of the literature. Usability attributes are described by ISO as "the features and characteristics of a product which influence the effectiveness, efficiency and satisfaction with which specified users can achieve specified goals in a particular environment" [1].

Debons et al. refer to usability attributes as "items affecting satisfaction" [2] and offers a list of 10 items which apply to visual user interfaces. Poulson [3] created a general purpose measurement tool, in the form of a questionnaire, which can be used to assess the perceived quality of software interfaces by users, identifying 'dimensions' which are required to produce an effective scale. Other research has concentrated on producing guidelines to help the design of usable interfaces. Molich et al. present a checklist of "usability considerations in a good dialogue" [4] ('dialogue' here refers to that of a visual system interface) to be used during evaluation procedures. Attributes which appear repeatedly in the literature relate to reliability, concentration required, ease of use, speed, enjoyment, satisfaction, ease of learning and quality of instruction.

3.1 DISCUSSION OF SELECTED ATTRIBUTES

From the pilot studies and the literature review, an initial list of twenty-two attributes was drawn up relating to the usability of automated telephone services. These included the attributes mentioned above together with a number of further attributes, some specific to speech interfaces, such as, the clarity of the voice and other prompts, the likability of the voice, and comparability to a human-operated service.

3.2 USER-DERIVED RANK ORDERING OF ATTRIBUTES

In order to determine the salience of individual attributes, an experiment was carried out in which a group of subjects was presented with the list of twenty-two attributes identified as being appropriate to the simulated telephone service they had used. Subjects were asked to write down the six attributes they considered to be the most important and then to rank the six in order of importance. Subjects were also asked to write down any additional attributes they considered important but which had not been included in the list provided.

A second matched group of subjects was asked to identify six important attributes and then indicate their order of importance in the same way. These subjects were not provided with a list of candidate attributes from which to choose. In this way, subjects' responses were expected to be made without any influence from previous exposure to the findings devised from the literature and preliminary evaluation studies.

Tables 1 and 2 below show the top ten attributes which appeared most frequently in the lists given by subjects from each of the two groups. The frequency (F) represents sum of the number of 'votes' for each attribute (the ranking of importance given by subjects is not taken into account in these tables).

ATTRIBUTES	Frequency (F)
Clear instructions	16
Clarity of voice	15
Option to start again	13
Easy to follow	12
Ease of use	9
Reliability	9
Good understanding of speech	9
Not too long	9
Convenience	7
Efficiency	7

Table 1: Responses from 20 subjects given a list from which to select attributes

ATTRIBUTES	Frequency (F)
Clear voice	8
Clear instructions	8
Adequate answer time	6
Simple to use	6
Convenience	5
Rapid service	5
Speed	4
Efficient	4
Checks number is correct	4
Clarity	4

Table 2: Responses from 20 subjects without a proposed list of attributes

The results of Tables 1 and 2 show that there is a strong similarity between the type of attributes chosen by subjects from the two groups. The attributes identified from the literature and pilot studies were found to be fully included in the list given spontaneously by subjects who had no proposed list. The subjects who had no exposure to the proposed attributes gave a much greater diversity of attributes in their lists. Seventeen of the attributes mentioned by this group were mentioned by one subject only. This accounts for the much lower frequency counts for the most popular attributes given by the unprompted group.

None of the subjects provided with the attribute list wrote down any additional attributes that they had thought of

themselves. This gives a strong indication that the proposed list included most, if not all, of the automated services salient usability attributes. The attributes given the highest ranking were, in both groups, clarity of voice and clear instructions.

4 FURTHER IDENTIFICATION OF ATTRIBUTES

The initial set of twenty-two attributes was augmented by studying comments made by subjects after experiments. New features which were widely mentioned were considered as candidates to be added to the original attribute list. These included:

Start again facility. The facility to go back and read out the numbers again if the user had made a mistake.

Help and repeat facilities. The facility to repeat instructions and prompts, and to ask for help if difficulties arise.

Security and confidentiality. Users seek assurance that information they give to the service will stay confidential and that there would be no security problems in relation to their credit card account.

Pacing of service. The pacing of the human-computer interaction.

5 WEIGHTING ATTRIBUTE SALIENCE

Using the augmented pool of attributes, a further experiment was conducted to determine subjects' ratings of the relative importance of each attribute. In the experiment, subjects used a simulated home shopping telephone service and were then given sets of attributes from which they were asked to indicate the importance of each on a four-point scale.

Subjects' responses were analysed by assigning values to the four categories of the importance scale as follows: 1 = "not important", 2 = "fairly important", 3 = "very important" and 4 = "extremely important". Table 3 shows the attributes which were generally given the highest and the lowest responses. The results shown are taken from 154 subjects.

ATTRIBUTES	Average score
Security of credit card number	3.82
Reliability	3.69
Confidentiality of information	3.66
Efficiency	3.58
Clear instructions	3.58
Understood well	3.48

Table 3: Attribute scores (highest ratings)

ATTRIBUTES	Average score
Liking the beep	1.73
Liking the voice	1.95
Don't need much concentration	2.11

Table 4: Attribute scores (lowest ratings)

Security, confidentiality, reliability and efficiency were consistently given very high ratings of importance suggesting that users need to feel the service incorporates these attributes since it is dealing with their credit card accounts. Also it is paramount that clear instructions are given using a clear voice.

Aesthetic considerations, such as how likeable users find the beep tone and the voice, are relatively unimportant compared to other attributes.

6 DEVELOPING A USABILITY QUESTIONNAIRE

Using the results described above, a questionnaire has been devised specifically for the evaluation of users' perception of the usability of automated telephone services. The questionnaire uses the Likert rating scale technique [5], where the strength of agreement with a clear statement is measured. A seven-point Likert scale was used throughout, with labels ranging from "strongly disagree" to "strongly agree". Statements reflecting the identified attribute concepts form the basis of the questionnaire content. For example, 'speed of service' became "I thought the automated catalogue service took too long". Wordings were carefully chosen so that subjects were exposed to an equal number of positive and negative attitude statements.

One of the experiments using this questionnaire evaluated the sensitivity of the questionnaire for measuring perceived usability of a simulated telephone service at four different levels of recogniser accuracy. A total of 256 subjects were used, divided into four matched groups, matched for sex and age group.

The positions on the Likert scales were given weights from 1 (a very negative response) to 7 (a very positive response) for scoring purposes. The sample means (see Figure 1) across all attributes indicate a rise in positive evaluation of the service from 85% to 100% recognition accuracy for the automatic speech recognition component. A Friedman Two-Way Analysis of Variance [6] was performed on the summed scores for each subject across the four experimental groupings. A significant effect was found in perceived usability due to recogniser level ($p < 0.001$). Significant differences ($p < 0.001$) were found between group 1 (85% accuracy) and group 4 (100% accuracy), between group 2 (90% accuracy) and group 4, and between group 3 (95% accuracy) and group 4 . These results suggest that the questionnaire is reliably detecting changes in users' attitudes as a function of speech recognition accuracy.

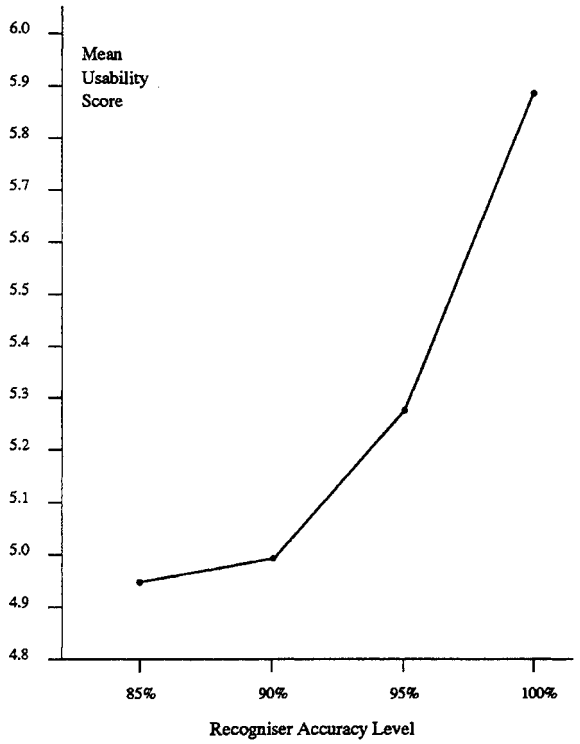

Figure 1: Mean Scores on Likert Attitude Scales for each Four Levels of Simulated Recogniser Accuracy

To further investigate the questionnaire validity, a usability profile was constructed (see Figure 2) to give a visual representation of the shift in attitude perception for each of the usability attributes as a function of recogniser accuracy. This was done by taking the means of responses to each of the statements across all subjects in each group for each of the speech recogniser accuracy settings.

The results demonstrate that the questionnaire is sensitive enough to measure differences in perception of service usability when the recognition accuracy level is varied. The validity of the content of the questionnaire was addressed by conducting a Pearson-Product Moment Correlation [7] on the responses to the attribute statements. The results of this analysis indicate that the variables are well correlated, suggesting that the questionnaire is a valid method of assessing user's perceptions and attitudes towards the usability of automated telephone services.

7 CONCLUSIONS

This paper has reported on the identification of salient usability attributes for automated telephone services. These salient attributes have been used as the content of a questionnaire which has proved to be a reliable and valid measurement tool of user-perceived usability.

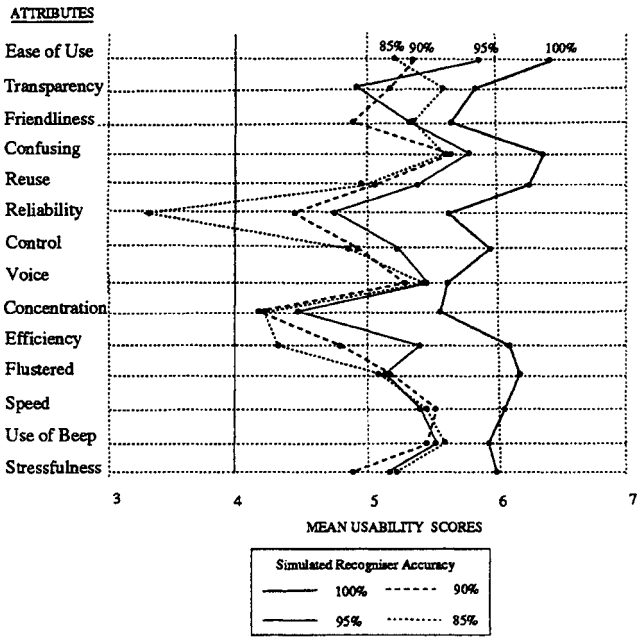

Figure 2: Section of the Usability Profile

8 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 and the Intelligent Dialogues team at CSTR.

REFERENCES

[1] ISO. *Ergonomic Requirements for Office Work with Visual Display Terminals (VDT's)*, 1990. ISO CD 9241-11.

[2] A. Debons, W. Ramage, and J. Orien. Effectiveness model of productivity. In Hanes L. F. and Kriebel C. H., editors, *Research on Productivity Measurement Systems for Administrative Services: Computing and Information Services*, volume 2. NFS Grant APR-20546, July 1978.

[3] 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.

[4] R. Molich and J. Nielson. Improving a human-computer dialogue. *Communications of the ACM*, 33(3), March 1990.

[5] A. B. Anderson, A. Basilevsky, and D. P. J.Hum. Measurement: Theory and techniques. In Rossi P. H., Wright J. D., and Anderson A. B., editors, *Handbook of Survey Research.* Academic Press, 1983.

[6] S. Siegal and N. J. Castellan. *Nonparametric Tests for the Behavioural Sciences.* McGraw-Hill International Editions, 1988.

[7] D. C. Howell. *Statistical Methods for Psychology.* PWS Publishers, 1987.