Proceedings of the Institute of Acoustics

DESIGN AND EVALUATION OF DIALOGUES FOR AUTOMATED TELEPHONE SERVICES

J C Foster (1), R Dutton (1), M A Jack (1), S Love (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

Small vocabulary, speaker-independent speech recognition over the telephone network has now reached a level of performance which allows it to be considered for data transfer applications such as banking and catalogue shopping. A key consideration in the development of these applications will be the design of the speech interface since speech is the sole communication mode between the user and the service.

The research reported in this paper is currently addressing speech interface design issues through a series of large-scale field experiments using a new real-time Wizard of Oz (WOZ) experimental methodology designed to permit the investigation of users' attitudes towards simulated telephone services and the evaluation of the perceived usability of such services. During the experiments, specific features of the speech interface are manipulated and the dependent variables of user attitude and perceived usability are measured using telephone and postal questionnaires.

Among the features of the interface which can be independently modified in the WOZ experiments are a simulated speech recogniser and a dialogue scheme. One of the major new features of the work carried out to date is that it is based on the parametric simulation of an existing speech recognition technology. This allows experiments to be carried out with both current recognition performance levels and recognition performance level extrapolated beyond those currently achievable. In this way, it is possible to address, among other key issues, the shape of the usability function for automated telephone interfaces for different levels of recognition performance.

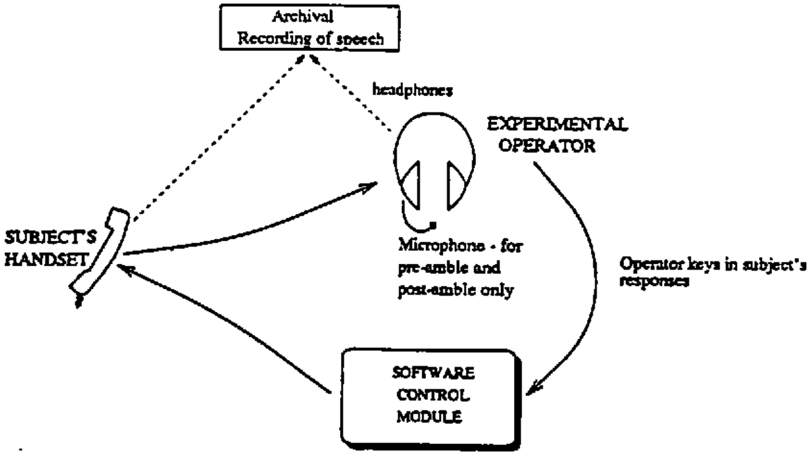
Modifying the speech interface can also involve changing the dialogue scheme used in the automated telephone service. This allows important user interface and human factors issues to be addressed in the experiments such as the impact of voice quality, the degree to which conventional 'beep' prompts (used in addition to spoken prompts) influence the progress of the dialogue, and the impact of dialogue structure and prompting strategies on users' responses to the service.

Results from the project on the effects on usability of changing the accuracy level of the simulated speech recogniser in the telephone interface have been reported elsewhere [1,2]; the present paper is concerned mainly with dialogue issues.

2. A NEW WIZARD OF OZ SCHEME

WOZ experiments involve the use of a hidden operator who, unknown to the experimental subject, simulates some aspect of the performance of a computer [3]. In the case of the experimental work reported in the present paper, contact between subject and computer is over the telephone line. Consequently, it is a relatively easy matter to ensure that subjects are not aware that a human being plays any part in the running of the automated telephone service.

The WOZ experimental configuration which includes the user/subject, the WOZ operator and the speech interface software is shown in Figure 1.



This software
a) simulates recognition errors
b) delivers prompts and questions to subjects
c) controls timings of prompts and other delays
d) controls dialogue sequence
e) records all statistics including operator's keystrokes

Figure 1: The WOZ Experimental Configuration

In experiments, the Software Control Module handles the initiation of contact at the subject's home or workplace, the delivery of the dialogue prompts to the subject, the registering of keystrokes by the experimental operator made in response to the spoken input from the subject, the on-line generation of recognition errors when these are required and the recording of all statistical data on keystrokes and timings.

The WOZ software runs on an IBM-compatible PC with a plug-in circuit card providing connection to the telephone line. In addition to simulating the speech recogniser, the software controls the archival recording of all interactions between the subject and the operator.

This particular WOZ setup has a number of important advantages over previous WOZ investigations into speech interfaces. These include the extensive control it gives over the experimental conditions; the constraints imposed on the operators who key in the subjects' spoken responses; and the fact that all other experimental conditions, such as the introduction of speech recognition errors are entirely under software control.

When entering subjects' spoken responses to the dialogue prompts, the WOZ operators follow certain rules or keystroke protocols. Different protocols cause the experiment to follow very different paths through the space of possible dialogues. In the experiments carried out to date involving the simulation of an isolated word recognition system, the protocol is defined such that if the expected answer is spoken by the subject in response to a dialogue prompt, even in the context of extraneous linguistic or paralinguistic expression, that expected answer is keyed in by the operator. For example, if a subject in response to the prompt "Please say the next digit" says "It's three" or "Um...Er...three" instead of simply "three" as expected, the operator keys in the digit "3". This protocol implies a generous view of the capabilities of an isolated word recognition system in the speech interface, including aspects relevant to wordspotting methods. It does mean, however, that fewer dialogues fail in experiments, an important consideration in the early stages of the research work. A stricter keystroke protocol would require that only distinct input be accepted by the WOZ operators. Later experiments in the programme will compare the effects of these and other protocols and will develop protocols for connected word recognition and more realistic word spotting techniques.

When the WOZ software returns control from the automated telephone service to the experimental operator, it is important to decide exactly what to tell subjects about the success or failure of the interaction and about he purpose of the experiment. Experience has shown that the wording at this point in the experiment has a major impact upon users' responses to the usability of the service. For example, in experiments to date subjects have been told that when such services become generally available, their credit card account would be automatically debited. This gives them a wider appreciation of the implications of using this service, a fact reflected in their responses to the telephone questionnaire which is delivered by the operators immediately following subjects' use of the service and to the postal questionnaire which subjects complete as soon as possible after completion of the telephone call.

3. THE SPEECH INTERFACE

The speech interface used in the telephone experiments reported in this paper consists of a simulated speaker-independent, isolated word recogniser coupled to a dialogue scheme which controls the delivery of all prompts to users. The simulated speech recogniser at the heart of the WOZ software is modelled on a detailed parametric characterisation of a current speaker independent speech recognition system. The performance data for this recogniser have been analysed in terms of a speaker-specific variable g(s), an utterance-specific variable $p_w(u)$, and a word/lexical pair $q_{w,r}(u)$. These parametric distributions are shown schematically in Figure 2.

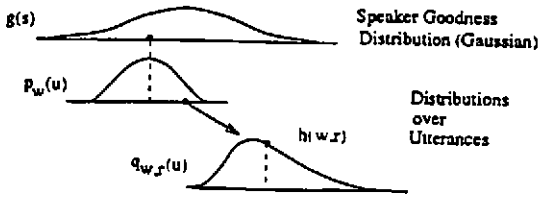

g(s)
Speaker Goodness Distribution (Gaussian)
$p_w(u)$
Distributions over Utterances
h(w,r)
$q_{w,r}(u)$

Figure 2: The Simulated Recogniser

Access to the speaker-specific variable g(s) allows experiments to be designed on a "worst case" basis, using speakers with a low value of "goodness" representing the "goats" in a population who find difficulty in speaking to automated systems. Access to the utterance dependent variable $p_w(u)$, allows experiments with acoustic variations produced over time with a given set of input words. Access to the word/lexical pair parameter terms h(w,r) and $q_{w,r}(u)$ allows investigation of the effects of acoustically and phonetically confusable words.

In operation within the WOZ system, the recogniser simulation acts as an 'error' generator, assuming the input to the simulator is a correct input word (keystrokes) and calculating a probability score y(s,w,u,r) for correct recognition, query or rejection as a function of the three parameters:

$$y(s,w,u,r) = g(s) + p_w(u) + h(w,r) + q_{w,r}(u) \qquad [r \neq w]$$
$$= 0 \qquad [r = w]$$

User interactions with the simulated speech recogniser are mediated via a dialogue consisting of a linked network of prompts and opportunities for input from the user. One such dialogue for which results are reported below involves the user entering a 16-digit credit card number in four blocks, each of four digits. This generic dialogue could form a component in a banking or catalogue ordering automated telephone service application.

A typical session with this dialogue starts with the message from the service welcoming the user and continues with brief instructions on how to enter the credit card number, one digit at a time, following the 'beeps'. Each of the possible end states of the simulated recogniser (following user input) directs the dialogue along a defined dialogue path. For example, if an input is accepted and the user is in the middle of reading a block of digits, a single 'beep' prompts the user for the next digit in the block. If, on the other hand, the last digit of the block is accepted and there are blocks of digits remaining, the service thanks the user for reading the number so far and prompts for the next block of digits. Queries by the recogniser activate subdialogues in which the first and (if necessary) the second hypotheses are offered to the user for confirmation. Rejection of input by the simulated recogniser is followed by a request for repetition.

In the case of no input from the user (and therefore no keystroke from the WOZ operator), the "recognition window" times out after a predetermined period and the dialogue responds with a suitable prompt.

The current dialogue allows a maximum of three attempts by the user to enter each digit in the credit card number before the automated service signals failure and returns the user back to the WOZ experimental operator.

4. EVALUATION OF ATTITUDES AND USABILITY

The two principal instruments used in this work to measure users' attitudes towards the automated telephone service are a telephone questionnaire and a postal questionnaire. The telephone questionnaire consists of a number of questions asked by the human operator immediately after completion of the experiment with the automated service. Among other questions, subjects are asked how many mistakes they thought they made and how many they thought the service made; how long they thought it took them to read their credit card number; and how they evaluate the service overall. For most of these questions, the actual data (such as the number of mistakes and the duration of the call) are already known from the transcription made automatically by the WOZ software during the experiment. Discrepancies between the data and subjects' perceptions are valuable in evaluating subjects' perceptions of the effectiveness and efficiency of the service, both of which contribute to measurement of the usability of the service [4,5].

The postal questionnaire consists of several sections eliciting subjects' attitudes and general comments on the service. The attitudes are measured with 22 responses on 7-point Likert scales each with a mid neutral point. These data are used to derive detailed statistical profiles of users' attitudes and the perceived usability of the automated telephone service. The questionnaire items have been designed on the basis of extensive user interviews to identify the key salient attributes of such automated telephone services and the stability and sensitivity of the questionnaire has been proven over a total of seven separate experiments with different subjects.

5. DIALOGUE DESIGN AND EVALUATION

Effective dialogues are the key to the development of usable automated telephone services. The experiments carried out to date on the generic dialogue described above have highlighted the range of difficulties users can experience with such services and have suggested a variety of possible solutions which are being actively investigated.

5.1 Unexpected Responses

Users do not always respond with a word or words from the active vocabulary. Part of the learning process involved in using an automated service is to know what can be said at each request for input by the service. This is, of course, a general problem in the design of interfaces of all types.

In one large experiment (256 subjects), there were 49 examples of unexpected responses to the service. The variety of responses included, very commonly, answering with a digit as well as 'yes' or 'no' in reply to confirmation prompts. For example, in reply to 'Did you say six?', the user might say 'No, five' or 'Six, yes' rather than simply 'yes' or 'no' as prescribed.

One strategy used in the experiments to avoid this problem is to remind users whenever possible of the active vocabulary by including reference to it in the prompts. For example, in recovery or confirmation prompts, where the active vocabulary is {yes, no}, phrasing such as 'Answering 'yes' or 'no', did you say ...?' was used. Interestingly, when at the end of the dialogue the number was read back to subjects without this reminder, 11% of subjects gave an inappropriate response, suggesting that learning how to use services of this type is something that needs continual reinforcement.

5.2 Inappropriate Timing of Responses

Inappropriately timed responses are those spoken outside the time window in which the speech recogniser is active. One common type of inappropriately timed response is speaking during a pause in a syntactically complex prompt consisting of either a sequence of independent sentences or a single sentence having a main clause - subordinate clause structure. In both these cases, there tends to be a short pause between the relevant parts of the prompt and if the first part is semantically interrogative or is understood, pragmatically, as requiring a response, many users supply the response before completion of the prompt. Again, this is a major issue for automated telephone service interface design, requiring 'intelligent' strategies for either avoiding the situation or for curtailing the delivery of the prompt if the user is recognised to have supplied an appropriate input.

In one experiment, 18% of subjects gave inappropriately timed responses at some point in the dialogue. This is a high figure pointing to a problem that needs to be addressed in dialogue design. However, it is overshadowed by the problem of subjects speaking on or before 'beep' prompts. In experiments carried out to date, most spoken prompts were followed by a fixed-duration, constant frequency, system generated tone, referred to as the 'beep'. In the case of the best dialogues (those where there were few or no queries or rejects because the simulated recognition accuracy was high), users heard a 'beep', input their digit, heard another 'beep' and so forth to the last digit of the number. In the case of queried or rejected input, subjects heard, in addition, a spoken prompt followed by a 'beep'.

In one large experiment, 56% of subjects spoke on or before the 'beep'; 8.6% of subjects spoke on or before the 'beep' four or more times. Evidence from this and other experiments suggests that while many users are able to adjust their response behaviour as their understanding of the service improves during a single dialogue transaction, some users do not learn to respond in the way the service requires.

5.3 Hesitation and Silence

Silence by the user in response to a request for input is generally a sign of serious dialogue failure. In a few cases it may be due simply to the user being distracted but in the great majority of cases it is due to the fact that the subject does not know how to respond. Careful study of the contexts in which silences occurred in experiments has identified a specific structural configuration responsible for causing silence: the unexpected shift following a series of verbal prompts each followed by 'beep' to a simple 'beep' prompt used without the support of a verbal prompt. The latter are often used for rapid-entry dialogues. The confusion arises because subjects are unable to interpret the meaning of the isolated 'beep'. In one large experiment, 14.5% of subjects were silent at least once during the session, most due to this cause. Consistency in prompting strategy is an essential element in maintaining user confidence during interactions with automated services.

6. CONCLUSIONS

A New Wizard of Oz experimental scheme has been described which permits large-scale field experiments into telephone-based speech interfaces for specific applications. Important dialogue features have been described for one application and general lessons about dialogue design discussed.

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] J C Foster, R Dutton, S Love, I A Nairn, N Vergeynst and F W M Stentiford, 'Intelligent Dialogues in Automated Telephone Services', to appear in Proceedings of the workshop on Interactive Speech Technology, 15 May 1992, NEC Birmingham, Ergonomics Society (1992)

[2] M A Jack, J C Foster and 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)

[3] N M Fraser and G N Gilbert, 'Simulating Speech Systems', Computer, Speech and Language, 5, pp81-99 (1991)

[4] ISO, 'Ergonomic Requirements for Office and Visual Display Terminals (VDTs), Part 11: Usability Statements', International Standards organization, ISO CD 9241-11 version 2.5 (1990)

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