An Investigation into the use of Error Recovery Dialogues in a User Interface Management System for Speech Recognition

Presented at Interact '90, Cambridge

Technical Report No. 108

Mary Zajicek
Jill Hewitt

August 1990

AN INVESTIGATION INTO THE USE OF ERROR RECOVERY DIALOGUES IN A USER INTERFACE MANAGEMENT SYSTEM FOR SPEECH RECOGNITION

Mary ZAJICEK & Jill HEWITT*

Department of Computing and Mathematical Sciences, Oxford Polytechnic, Gipsy Lane, Headington, Oxford OX3 0BP. Tel: 0865 741111 Fax: 0865 819666

*School of Information Sciences, Hatfield Polytechnic, College Lane, Hatfield, Herts. AL10 9AB. Tel: 0707 279327 e-mail comqjah@hatfield.uk.ac

Experiments were carried out to assess new users' attitudes to different versions of a speech input word processing system providing different error recovery strategies. Whilst they preferred a simple error message to none at all, a more complex recovery dialogue lead to decreased satisfaction with the system. This paper describes the experiments carried out and explores possible reasons for the results.

1. INTRODUCTION

The purpose of our investigation was to assess the effects of different error recovery strategies on first time users of a speech input word processing system. The system has been developed as part of our ongoing work into speech interfaces in a project entitled the "Intelligent Speech Driven Interface Project" (ISDIP)#. The investigation forms a part of a continuous programme to improve the usability of the system following an iterative design approach.

Earlier investigations (Hewitt & Furner 1988) had highlighted the importance of error recovery in a system where the recognition rate was not 100%, and a study of user satisfaction with the system (Zajicek 1989) lead us to believe that a more informative error recovery dialogue would be well received.

Our results were not what we were expecting, and have lead us to reconsider the interface design and the form of error recovery we might offer. In particular we propose that a more close adherance to human-to-human conversational strategies should be investigated.

2. THE EVALUATION

Our experiments were designed to ascertain "Whether first-time users prefer an informative error recovery dialogue to a minimal one or none at all". The 20 participants were selected from the staff in the Computer Science department at Hatfield Polytechnic, all of them had keyboard skills, but none had used a speech recognition system before. The task they were asked to complete was to input and save to disk a short letter, using speech alone; following this they completed an attitude questionnaire and made any comments about the system. Diagnostics were recorded automatically for each session, giving us recognition rates and the types of error that occurred.

2.1 The Recognition System

The voice recognition system used was a VOTAN VPC 2000 isolated word recogniser with limited vocabularies (of up to 64 words). It was therefore necessary for the users to train all the words they would need to create edit and save the letter; these were contained in two vocabularies, the main one containing the words and editing functions and the other devoted to file handling commands. Movement between the vocabularies was to be made explicitly using a "switch" command.

2.2 An Experimental Session

Subjects were assessed individually and were not allowed to view sessions preceding their own so that the system would be new to each one of them. They were given a short demonstration of the training process followed by a part of the letter being dictated and then saved to disk. The training program was started for them and they trained each word just once before embarking on two separate attempts to create the letter.

They were told to watch the screen for possible error messages whilst inputting the letter - no error tone was used for this experiment. They had access to a limited number of editing functions and an UNDO command and were asked to try to ensure that the letter was correct, but to abandon attempts to correct a word if after 3 tries they could not get it right.

Prompts from the experimenter were given to enable them to save the file they had created and quit from the word processor.

2.3 The Word Processing System

The word processor used has been specially created for voice input, it has a limited range of editing functions

Funded by NAB under the NAB3 initiative in Artificial Intelligence.

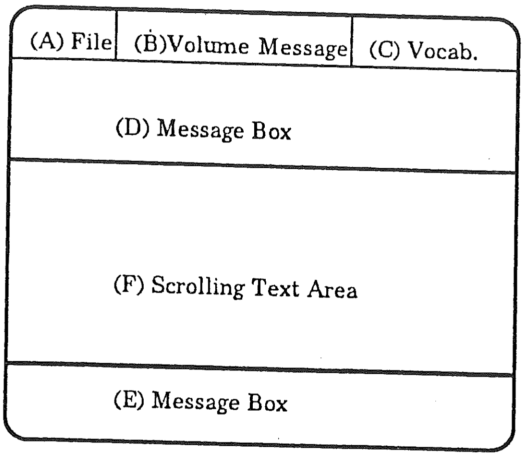
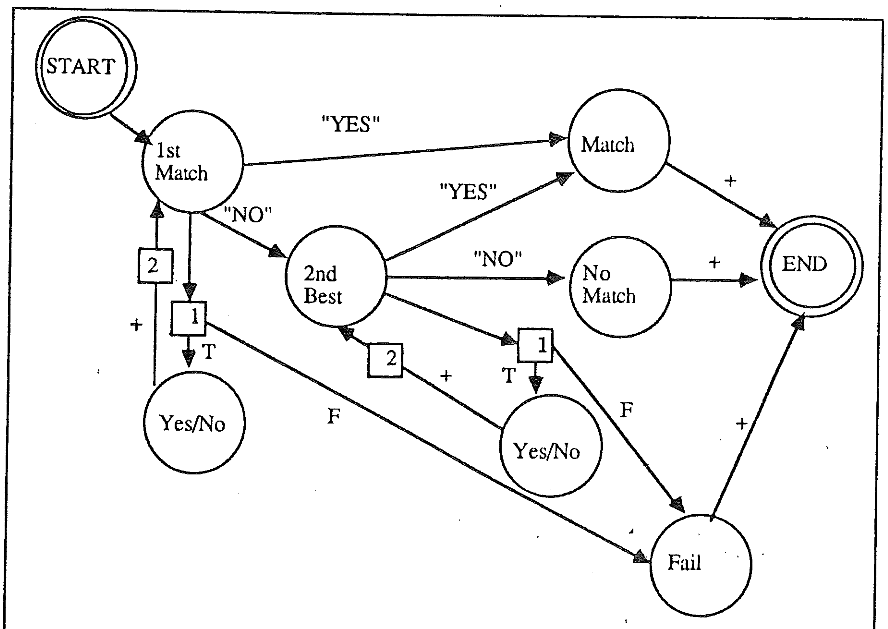


Figure 1

which are sufficient for simple text creation tasks. The screen design includes special fields related to the voice recognition system, an outline is given in Figure 1. The labelled regions are used as follows:

A. The name of the file being edited
B. Reserved for the messages "Speak Louder Please" or "Speak Quieter Please"
C. The name of the vocabulary currently in use by the Recogniser
D. The main area for error messages.
E. Reserved for sub-dialogues between user and the interface management system and for auxiliary error messages.
F. A scrolling text area.

Pop-up windows appear in the text area for the file menu and the edit menu. The participants did not use the edit menu but were provided with a limited set of editing commands in the form of cursor movement, delete (previous character) and undo (last action).



1st Match
Write to message area D
"Close Match, Do You Mean" 1ST_MATCH

2nd Best
Write to message area D
"Second Best, Do You Mean" 2ND_MATCH

Match
Clear Message Areas

No Match
Write to message area D
"No Match"

Yes/No
Write to message area E
"Please Answer Yes or No"

Fail
Write to message area D
"Giving up after" X "attempts"

1. X = LIMIT
(4 in experiment)

2. set X to 0

Figure 2

2.4 Recovery from Recognition Errors

Three strategies were used in the evaluation:

i. No error recovery:- the system will always find a match for an utterance even if it is not a good one. In this case, users will get the wrong word coming up on the screen. If they notice it immediately they can use the 'undo' command to remove it, otherwise they have to delete it character by character. It is possible that users may not be able to get the correct word if it has been badly trained and is too 'close' to another.

ii. Minimal information:- if the system does not find a good match for an utterance it will print the message "Don't recognise that word" in the error message area (D) and wait for the user to say something else. Wrong words will appear less frequently on the screen, but, depending on the setting of the recognition threshold, users may get stuck on a particular word which was perhaps poorly trained.

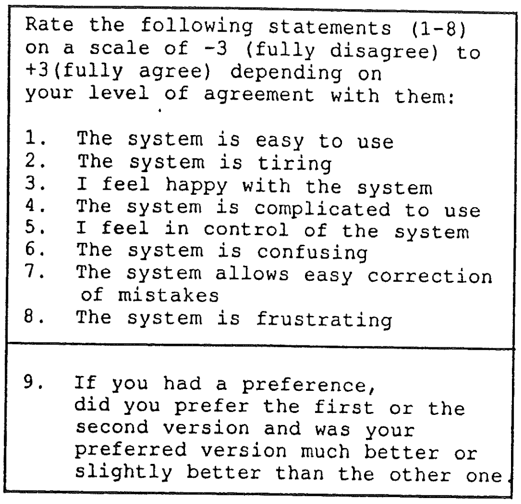
iii. Close Match dialogue:- in the case of an uncertain recognition, the system will offer a recovery dialogue as shown by the USE (Wasserman et al.1985) state transition diagram in Figure 2. If the first word offered is not the one they want they are offered a second choice. If that too is incorrect they will have to say the word again. The recovery dialogue uses both the message areas on the screen, the lower one being reserved for the message "Please answer Yes or No" if one of these words was not detected in answer to a question.

The system does in fact have a fourth recovery strategy - the "failsafe" which was not used for this experiment. It offers a selection of all possible words and highlights each in turn, asking the user to whistle when the required one is reached. This is essential when it is being used for 'real' work, particularly if the user is handicapped and cannot resort to using the keyboard.

Subjects were divided into two groups, the first each used two versions of the system, one with no error recovery beyond the undo command (strategy 1) and one with the minimal information (strategy 2). The second group used the version supporting strategy 3 for both tasks, but they were not told that there was no difference between the versions and were asked the same questions as the first group.

2.5 The Questionnaire

On completion of the task, subjects were asked to consider 8 statements related to the system and to score them on a scale of -3 to +3 depending on their level of agreement. These are reproduced in Figure 3. They were also asked which version of the system they preferred, if any, and whether they had a strong or marginal preference for it (question 9). They were given the opportunity to make any other comments regarding the system, including any suggestions for improvements they might have.



Rate the following statements (1-8) on a scale of -3 (fully disagree) to +3(fully agree) depending on your level of agreement with them:

1. The system is easy to use
2. The system is tiring
3. I feel happy with the system
4. The system is complicated to use
5. I feel in control of the system
6. The system is confusing
7. The system allows easy correction of mistakes
8. The system is frustrating

9. If you had a preference, did you prefer the first or the second version and was your preferred version much better or slightly better than the other one

Figure 3.

2.6 The Diagnostics

Whilst the subjects were using the recognition system a diagnostics program running in the background was collecting information. The statistics collected for all three strategies included:

- Number of words spoken
- Number rejected by the recognition unit
- Number of undo commands given by the subject

In addition, for the third strategy, the following were collected:

- Number of accepted 1st matches
- Number of accepted 2nd matches
- Number of rejected 2nd matches
- Number of yes/no errors

Two recognition rates were calculated automatically :-

- The basic recognition rate:
words recognised/words spoken

-Rate plus Undo's
(words recognised + undo commands)/
words spoken

We felt it was necessary to also calculate an adjusted rate:-

(words recognised + undo commands + errors in finished letter) / words spoken

This was not a very exact measurement, but allowed some account to be taken of mistakes that the subject could not correct or had not noticed when creating the letter.

These diagnostics are summarised in Figure 4.

Strategy:	1	2	3
Avg. no. utterances	95	76	89
Basic Recog. Rate	100	87	93
Rec. Rate incl. undo	93	83	89
Adjusted Rate	90	82	88
Undo Commands	7	3	3
Yes/No errors	-	-	5

Figure 4.

3. RESULTS OF EXPERIMENTS

User scores on questions 1-8 of the questionnaire were totalled, providing a numeric measure of their satisfaction with the system generally. Since dialogue structure was the only variable component, changes in measure were taken to reflect different scores for dialogue structures. Diagnostic files were analysed for changes in recognition rate, the number of yes/no errors, and the number of undo commands. The letters created by subjects were examined for errors.

The results were analysed to find the answers to two questions:

3.1 Are users happier with a description of no match rather than wrong machine action?

The primary indicator was the result of question 9, which asked subjects to indicate their preference. It was found that eight out of nine preferred the description of no match. This preference was shown to be unrelated to recognition rate which in some cases was slightly better for the 'wrong machine action' version. The number of undo commands was predictably higher for this version. The number of errors in the finished letters were similar.

3.2 The hypothesis that "Users find no difference between more explanatory dialogue with the offer of a second best match and basic dialogue which either offers a no match message or wrong machine action"

The results of questionnaire answers were grouped as follows:

Group 1 - those who experienced no error recovery (wrong machine action) or minimal error recovery dialogue in the form of the error message "I don't recognise that word" (strategies 1 and 2 as described in section 2.4).

Group 2 - those who experienced error messages describing first and second best choices (strategy 3 described in section 2.4)

The score for each subject on questions 1-8 was computed, taking into account the sign change for negative questions. The scores were then subjected to a one-tailed Mann-Whitney U Test of the null hypothesis above. It was found that this hypothesis could be rejected at .02% level and a significant shift in favour of basic dialogue was detected. The results indicated that users preferred the basic dialogue with either a 'no match' message or wrong machine action.

4. DISCUSSION

The aim of experimentation was to analyse users' response to different dialogue structures, and to establish usability guidelines for speech driven dialogue design. The results have clearly shown users' preference for different dialogue structures and indicated the importance of fundamental dialogue design issues. They have also provided insight into the user's relationship, and perception of, a speech driven system, in particular the problems that are encountered when the user subconsciously expects natural human dialogue structures rather than restricted, conventional computer interaction dialogue.

4.1 Levels of sub-dialogue activity

Previous work (Zajicek 1990) had shown that users were, in principle, in favour of a high level of explanatory sub-dialogue. They expressed the opinion that it contributed to feelings of being in control of the system and gave them confidence in achieving error recovery. They made the assumption that the more explanation there was the more information they had to work with in handling the speech driven word processor, i.e. 'Knowledge is power!'

Researchers however, were aware of problems associated with information overload (Nusbaum 1986) and the need for simple dialogue on a multi-functional screen (Dye and Cruickshank 1988). Although the concept of more information more power may be sound, the increased activity involved in reading sub-dialogue messages contributed to cognitive overload, and detracted from the fluency of operation of the word processor.

The aim of experimentation was to observe subjects using the word processor and gain a rating for different forms of feedback dialogue, enabling a comparison between a users' expectations and the actual experience.

4.2 Comparison between the strategies of no error recovery and a description of no match.

The answer to question 9 of the questionnaire showed that users are happier with a description of no match rather than wrong machine action. This indicated that the dialogue displayed in the top sub-dialogue area, if in the brief form of a description of no match, was considered to be more effective than no dialogue at all. This was consistent with previous users' positive view of the value of increased sub-dialogues.

The result was also consistent with rules of human conversation in that if a word is misheard, the listener will give the equivalent of a no match message such as "Pardon" or "Can you repeat that please" rather than act on an unlikely guess. The strength of the description of no match was that it was succinct and required no further action.

4.3 User Preference for a basic dialogue or no error recovery over a more explanatory one

This result shows that although a 'no match' message was preferred to wrong machine action, users were less well satisfied with a system giving more explanatory error messages. This result was not consistent with previous experiments where users had expressed a preference for increased explanation, the experience was then in some way different from the expectation.

It is well known that help facilities which enable a user to 'get started' with a system soon become irritating when the user becomes competent, however in these experiments all users were inexperienced and persisted in overlooking useful information.

Observations showed that subjects frequently misused the yes/no answer system provided for selecting first and second best matches. They failed to answer yes or no even when prompted and often continued to scan the upper dialogue area for clues to their apparent deadlock. This is substantiated by the large number of yes/no errors in their diagnostic files (a yes/no error occurs when the system is expecting one of those words and the user says something else). It must be noted that even the most experienced researchers sometimes forgot to consult the lower sub-dialogue area for information when error situations arose. The yes/no dialogue had been flagged as a problem in previous prototypes and these experiments have helped to clarify aspects of the problem.

4.4 Mapping Dialogue to the User's Conceptual Model

Although a yes/no response is an effective method for clarifying the user's intention in keyboard interfaces where typing 'y' or 'n' is quicker than re-typing a command, subjects' behaviour has shown its use to be at variance with their conceptual model of a speech driven interface.

The user's conceptual model of a speech driven interface appears to be more closely related to human conversation than conventional keyboard dialogue, although other researchers (Newell, 84) maintain that people use unnatural speech when addressing machines. The usual conversational response to a misheard word is to repeat it. The preferred basic dialogue mode offered the no match message and the chance to repeat the utterance of the word. Although it is less sophisticated it provides a close match to human conversational strategies.

However if we pursue the analogy with human conversation, there are situations when if a word is misheard the listener will say 'did you say ....' and the speaker will naturally say yes or no rather than repeat the misheard word. This dialogue strategy is usually employed after several repetitions have been tried.

4.5 Dealing with information overload

As described above, feedback dialogue is presented in several areas above and below the main text of the word processor. It was hoped that users would become familiar with the function of individual sub-dialogue areas, automatically scanning them when particular information was needed. The results of experimentation have shown that the sub-dialogue area below the word processor text was not easily recognised by the user as performing a yes/no dialogue role.

Reasons for the apparent neglect of the lower sub-dialogue area have not yet been established, but two explanations are offered. Firstly that the position of the lower sub-dialogue area does not correspond to the user's conceptual model of the interface and their expectation that all information will be displayed *above* the word processing text; secondly, that the sub-dialogue itself is not comfortable for the user.

It is possible that a sub-dialogue utilising speech output would be more acceptable, satisfying the user's expectations of a more natural conversational mode. If the dialogue (whether spoken or screen output) was invoked only after several attempts to recognise a word it would conform more closely to a human to human conversational strategy.

5. CONCLUSIONS

The experiments have provided insight into natural assumptions made by users of a speech driven interface. Recognition rates are good enough to instill feelings of confidence in word recognition leading users to assume that dialogue structure emulates human conversation. They overlooked confirmation messages in sub-dialogue areas and in fact behaved as though they assumed 100% recognition.

The two particular dialogue strategies offered for assessment in these trials differed in that the basic dialogue strategy emulated natural conversation when a word is misheard. The more explanatory dialogue strategy, while conforming to normal computer interaction confirmatory rules, did not conform to the model of natural conversation. It was found to introduce confusion, and was less popular with subjects.

These results indicate that there exists a point at which expectation and conceptual models of a speech driven interface cease to be those of standard computer interaction, and take on the characteristics of natural conversation. The point may be determined by the level of confidence in the recognition of utterances. The ISDIP system appears to have reached this point indicating that consideration should be given to a dialogue modelled more closely on natural conversation.

Further trials will need to be completed to investigate the usefulness of speech output in a dialogue and the degree to which a 'natural' conversation can be emulated. It is important also to assess more experienced users, particularly those who *have* to use the system in order to complete a task (i.e. disabled users), since their perceptions of the usefulness of the various strategies may be different from those of new users described in this paper.

6. REFERENCES

Dye, R. and Cruickshank. (1988) A System for Composing and editing text using natural spoken language. Proc. Speech '88, Institute of Acoustics.

Hewitt J.A. & Furner S.(1988) Text Processing by Speech: Dialogue Design and Usability Issues in the provision of a System for Disabled Users, in People & Computers IV, Jones D.M. & Winder R. (eds.)

Newell, A.F. (1984) Speech - The Natural Modality for Man-Machine Interaction. Proc. Interact '84.

Nusbaum, H.C. (1986) Human Factors Considerations in the Design Large Vocabulary Speech Recognition Devices in Proc. Speech Tech 1986.

Wasserman A.I., Pircher P.A. Shewmake D.T. and Kersten M.L. (1987) Developing Interactive Information Systems with the User Software Engineering Methodology in Readings in Human Computer Interaction, Baecker R.M., and Buxton W.A.S. .

Zajicek, M. (1990) Evaluation of a Speech Driven Interface, Proc. IEE UK IT 1990.