



Rapid prototyping as a design tool for dialogues employing voice recognition.

Stephen M Furner¹

ABSTRACT

Flexible voice recognition units are currently available as an add-on for popular micro-computers. These systems provide an opportunity to prototype speech dialogues for proposed products or services. This type of prototype can be useful in two ways. Firstly, to investigate the design of speech-based interfaces employing behavioural experimentation to quantify the effects on user attitude and performance of variations in interface design. Secondly, as a focus for discussion and informal evaluation within a design project producing a specific product or service.

INTRODUCTION

The design of a user interface employing voice recognition presents a significant problem for the Human Factors of software engineering. The dialogue with a user is transitory and subject to errors in the recognition process. Also, the characteristics of operation of the recognition technology are not the same as those of a human listener. Thus, the structure of a user dialogue should provide a framework within which the recognition system can successfully be exploited to allow the user to effectively communicate with an applications system.

Currently, there are no simple software tools, or rules of thumb, which can be employed to produce "easy to use", robust, dialogues, within which a recognition system can be embedded. However, speaker dependent recognition systems are now available for use with popular micro-computers such as the IBM PC, BT Merlin M5150 or M5200, or similar machines. These systems can be employed as a design tool to aid in the production of speech-based dialogues. By rapidly producing simulations of a proposed speech product, the characteristics of the dialogue become available for practical evaluation and refinement.

RAPID PROTOTYPING

To produce robust, reliable, user dialogues, an iterative design process can be employed. Here, a simulation of the dialogue is produced. This prototype is built so that it is easily modifiable. The prototype may only offer part of the functionality of the full dialogue, where a specific area has previously been identified as being problematic.

Once produced, a prototype is available for a development cycle of evaluation and modification. A simulation can be evaluated, modified to take account of the results of the evaluation, and then evaluated again. This cycle can be repeated until the dialogue reaches an acceptable level of performance in the evaluation procedure.

DIALOGUE DEVELOPMENT FOR VOICE PRODUCTS

Typically engineering design is not simply a question of a single individual designing and building an item of equipment to meet his or her own specifications. There may be many groups dealing with specialist areas that have an interest in the design of a finished product. Here, a prototype may act as a focus for decisions made within a project.

¹Human Factors Division, British Telecom Research Labs., Martlesham Heath, Ipswich, Suffolk, IP5 7RE

Evaluation and modification can be carried out as a result of demonstration, consultation and discussion amongst the design team and other interested groups.

The portability of a personal micro-computer, allows a prototype to be taken to meetings and displayed for comment. If the prototype is sufficiently robust in operation it can be left with a specific group so that they can use it when they go on to consult with other interested parties. However, the type of information required about the performance of a potential voice product will determine the way in which the information is obtained. If a precise quantification of performance is required, then rigorous behavioural experimentation should be employed.

A FORMAL EXPERIMENTAL APPROACH

Since formal behavioural experimentation is an academically defined procedure, it is, in some ways, easier to discuss than procedures taking a qualitative view of a product. A practical example of the formal behavioural approach was that taken within a project applying speaker-independent recognition and sophisticated signal detection to answering machine dialogues.

It was hypothesised that a telephone answering machine providing a dialogue based around a conversational question and answer exchange, would produce a significant improvement over a conventional answering machine dialogue structure. Also, if there was an improvement, it was not clear how large it would be. It was obvious that precise information was required which compared user behavior and attitude for the two types of dialogue, thus the formal experimental approach was needed.

The two answering machines were fabricated using two micro-computers fitted with a voice recognition card each. 120 subjects of the BT volunteer subject panel in Ipswich were sent letters asking them to phone a telephone number about an experiment which was being organised. 60 subjects were allocated to the ordinary answering machine, the other 60 the conversational question and answer dialogue. The calls to the machines were monitored and tape recorded for analysis.

The subjects were identified from the messages they left and were invited to visit the BT Human Factors experimental centre in Ipswich. Here, they filled in an attitude questionnaire and took part in a group discussion about answering machine usage. Subjects who could not attend this session were sent the questionnaire to fill in at home.

The results indicated that there was a significant difference in user performance with the two dialogues, but that the attitude of the subjects towards the two was not significantly different when statistically tested. Using the categories for caller behavior developed by H Maskery (1981) the results for caller behavior were:-

	Conventional	Conversational
slam down	14%	7.7%
no message	20%	2.6%
successful message	66%	89.7%

A chi-square test was significant at 0.05 with $\chi^2=8.06$ and 2 degrees of freedom.

A simple content analysis of the messages left by the callers for the information requested in both dialogues revealed:-

	Conventional	Conversational
name, address, availability for an experiment	30.8%	3%
name, telephone number, availability	30.8%	3%
name, telephone number, address, availability	25.6%	80%
all other combinations of items	12.8%	14%

A chi-square test was statistically significant at 0.001 with $\chi=27.7$ and 3 degrees of freedom. This indicated a clear advantage for the conversational dialogue since 80% of the successful messages contained all of the items of information requested by the machine; this was only 25% for the conventional dialogue.

The average time for a successful call to the conventional machine was 46.5 seconds whereas it was 97.9 seconds for the conversational machine, a t-test indicated that this was statistically significant at 0.0001 with $t=5.6$ and 37.6 degrees of freedom. This time difference could be reduced by improving the spoken prompts in the dialogue, and improving the structure of the dialogue to make it more efficient and reduce caller errors.

The relative frequency probability (rounded to 2 decimal places) of a successful message being obtained from a caller to the conventional machine was 0.66, for the conversational machine it was 0.90. Of the successful messages, the relative frequency probability of obtaining all of the items of information requested by the machine was 0.26 for the conventional, and 0.80 for the conversational machine. The probability of two independent events occurring, that is event A and event B occurring, is given by the probability of event A multiplied by the probability of event B. Thus, the probability of obtaining a successful message and all the items of information requested in the dialogue was, for the conventional machine $0.66 \times 0.26 = 0.17$, and for the conversational machine $0.90 \times 0.80 = 0.72$.

The results of the evaluation indicated that the conversational dialogue had clear advantages over the conventional dialogue in terms of user performance. For any random group of calls to a conventional machine only 17% of the callers could be expected to provide all the information requested in the dialogue, but for the conversational machine the owner could expect to find that 72% of the callers would have provided a message with all of the requested information in it. However, this performance advantage was not reflected in a more favorable attitude towards the conversational dialogue by the callers, it also took longer to use than the conventional dialogue.

The answering machine evaluation demonstrates the way in which speech-based interfaces can be subjected to measurement and comparison of the performance of design variations. However, this degree of numerical precision is only important if it is of some practical value, since in a design situation the information will have a limited shelf life (Furner 1987).

CONSULTATION AND QUALITATIVE EVALUATION

This is a less formalised procedure than the experimental approach, in which a prototype is used to act as a convenient means of demonstrating the potential product for comment. It also allows obvious difficulties to be dealt with as a result of the observation of subjects attempting to use it. The use of this procedure is demonstrated by the development of a proposed dialogue for a facility to be offered to the general public over the public telephone network. The recognition system was to be speaker independent. It was intended that a typical caller would telephone the recognition system, provide a numeric identification code, then provide a short message of numeric digits and some special characters such as a full-stop or a pounds-sign.

The recognition system was itself being developed in parallel with the dialogue. The intention was that the dialogue would be ready for use directly the recognition technology became available. By this method the overall time scale for the development of the technology to a point where it could be introduced as a service, could be kept to a minimum. A speaker dependent simulation using a micro-computer based recognition system was produced. This provided the basic functionality of the proposed service.

Over a series of meetings between the Human Factors Division, the BT research division developing the recognition technology, and the BT operating division interested in

providing the service, the prototype dialogue was iteratively refined. At each meeting the current version of the dialogue was demonstrated for comment. Between meetings the dialogue was modified to take account of the comments. The meetings tended to focus on different aspects of the performance of the dialogue, for example correction procedures for transpositions or recovery from a non-recognition.

When the dialogue had reached a stage where it was considered that it had been developed sufficiently to provide a viable service, the consultation procedure was expanded. The prototype system was taken from the research laboratories in Suffolk to offices of the BT operating division in London, so that it could be demonstrated to allow for comment by a wider range of interested groups within BT. It was demonstrated to a selected audience and then left to be used by those interested in the proposed service who wished to try it for themselves.

After feedback from the expanded consultation had been included in the prototype dialogue, a short user trial was carried out. Here, a small group of naive subjects were observed attempting to use the dialogue. This was followed by a group discussion about their experiences with the dialogue.

As a result of the information gained from the informal consultation and evaluation with the prototype dialogue system a series of flow-charts, and associated notes, were prepared for the design staff producing the recognition technology. This described a version of the dialogue which could be used to exploit the technology to expand the services BT can provide to its customers. This project illustrates the way in which a prototype can be exploited to improve dialogue performance by exposing it to a wide range of informal evaluations.

CONCLUSIONS

Two examples have been given in which two different approaches were taken to dealing with a proposed product utilising voice recognition. Both procedures proved useful for obtaining information about the design of a speech product. From these it is clear that micro-computer based recognition systems provide a useful dialogue design tool. Also, that the way in which this tool is used is dependent upon the type of information required within the design process for the product being prototyped.

REFERENCES

Maskery H. S. (1981) Telephone answering machines - an investigation into users behavior, University of Loughborough, Memo. 244

Furner S. M. (1987) Practical information about interface operating characteristics for engineering design, IEE Colloquium Digest 1987/38 "Evaluation Techniques For Interactive Systems Design"

Acknowledgement is made to the Director of Research British Telecom for permission to publish this paper. Acknowledgement is also made to John Reah, Bryan Mensforth and John Miles for their valuable contributions to the project work reported here.