THE ROLE OF HUMAN FACTORS TESTING IN SPEECH TECHNOLOGY

Dr Elizabeth Bednall
Josephine Chessari
Human Factors Team
Telecom Research Laboratories

Three studies are reported which focus on human factors issues in three speech-based telecommunications products. The purpose of the paper is to describe how human factors methodology can be applied in different ways. The first product shall be referred to as System A, the second as System B and the third as System C. The methods used provided valuable information regarding the human-computer interface for each system. These included:

1. user needs analysis
2. heuristic evaluation
3. observation of users interacting with the system while completing typical tasks
4. measurement of performance of the system
5. interpretation of questionnaire data
6. testing of product managers on the tasks

The results of these studies clearly illustrate the need for human factors testing to be incorporated into the design process. Such testing ensures a cost-effective way of optimizing usability and customer satisfaction when a product is finally released into the market place.

1. INTRODUCTION

It would be easy to assume that because speech recognition systems offer a "natural" interface between humans and computer systems, the design of such interfaces is just "common sense" and requires little effort. Paradoxically, it may be more difficult to get such an interface right, given well learned conventions in human speech behaviour which may or may not translate easily to the human-machine interface.

Human factors specialists ideally have a background in cognitive psychology and are well placed to study the behaviour of humans whenever they are required to interact with computer-based systems. The objective of any human factors study is to discover problems the human operator is likely to experience when using a given system. Solutions can then be recommended which take into account the cognitive capacity of the typical user.

There are a number of tools which human factors specialists use to shed light on problems with a particular interface. Usually, a combination of such tools provides the best view of design problems from the human factors perspective. Human factors studies may be carried out quite early in the design process, ideally in a test, modify, re-test (iterative) fashion, to ensure that the best possible system from the user's point of view is created.

It will become clear that an important component in deciding how to test a particular product is to establish, in advance, who the intended market is likely to be. In particular, despite the fact that the Systems A and B are nearly identical in many aspects, the target market is completely different. It is essential, then, to place the product in its correct context, before trying to establish how well it works.

2. "SYSTEM A" STUDY

System A is a telecommunications product which requires the user to dial into the service using a telephone. The user then enters a total of 20 digits to connect to a set of recorded voice announcements which instruct the user on how to proceed. Digit entry can be achieved in two ways. On all phones, the user can enter the number by saying the digits into the phone. The speech recognizer accepts isolated digit entry and prompts the user if it fails to recognize a digit. On phones connected to exchanges which enable Touchtone dialling, that is, tone rather than decadic (pulse), the number can be entered by using the keypad to key in the digits. In Australia, there is at best 50% availability of Touchtone. It is seen as essential then, to ensure that the voice recognition component of the system is very easy to use, as well as being attractive to customers. Target users for this product are international travellers and business people within Australia who travel interstate.

2.1 Method

- User needs analysis: This involved talking with the design team to establish exactly what the product was, and for whom it was intended.

- Heuristic evaluation: Two human factors specialists interacted extensively with the system to learn exactly how it worked and to pinpoint potential problem areas in the user interface.

- Tasks: Based on the heuristic evaluation, a set of typical tasks was designed to investigate these aspects, as well as to test a wide range of "real life" situations which users might find themselves in. The tasks provided a stringent test of the product, forcing people to make errors to see whether they could recover easily. When such tasks are undertaken in a structured way, and on representative users, it is likely that problems will be identified which might otherwise have gone unnoticed.

- Subjects: A group of 14 people was selected to take part in the study. These included males and females, people with and without accents, and people of a variety of ages, including two children.

- Procedure: Each user was tested individually on the set of tasks. The two experimenters listened in to the interaction of the subject with the system through headphones. Each session was tape-recorded and the behaviour of the user was carefully observed by the experimenters while completing the tasks.

- Data: Careful written recordings of what the user was doing were made during his/her interaction with the system. Performance of the speech recognizer was also monitored. Information was collected via questionnaire on demographics, subject opinions of the service and preference data.

- Testing of product manager: At the end of the human factors tests on subjects, the managers of System A were invited to a meeting where the results of the study were to be presented. One of the managers was asked to be a subject and to submit to the same testing procedures as had the "real" subjects. That person was given the tasks to do while the other people in the meeting looked on. This proves to be an extremely worthwhile way of demonstrating problems with the interface to those who have good product knowledge and who are responsible for the future product launch. Such people are often too close to their own product to see potential problems.

  In addition, tape-recordings of particular subject interactions were played to the product managers to illustrate where users got stuck, how successfully they recovered from such problem spots and to demonstrate any other aspects of the system which were felt to cause frustration and error on the part of the user.

2.2 Results and Discussion

A number of problems with the interface were discovered. There was a strong preference by users for the Touchtone component of the system. The Voice component was generally felt to be a slow and unwieldy method for digit entry. The Help message which is optionally given on connection to the service took just under 2 minutes. Once in the Help loop, there was no method of exiting, even if the user had Touchtone. Information regarding both Touchtone and voice entry was included, despite the fact that the user may only want to listen to one part of Help.

There was a problem of people not understanding the concept of Touchtone, and not knowing whether their phone was connected to an exchange enabling Touchtone dialling. This finding illustrates the need to educate the potential users about Touchtone, until such time when all phones have the Touchtone capability.

It was recommended that Help be made more context-sensitive, and that it should be available at all times. A strategy was suggested whereby a person could determine whether Touchtone dialling was available from the phone he/she was using by pressing a single key . The system would then prompt the user with the appropriate entry information. This was expected to greatly reduce user frustration.

It was felt that the performance of the speech recognizer, while not specifically measured, fell below a desirable level. The system at times failed to recognize both digits and command words. While for some subjects, performance was quite acceptable, for others, this proved to be a stumbling block for easy use of the system. This was especially true when subjects needed to correct errors, which they might have made themselves, or when the system misrecognized a digit. Sometimes the system did not understand the word "*Cancel*" which was the command used to correct errors. This led to a high level of user frustration.

There was found to be inconsistency across different components of the interface. For example, when correcting errors in Touchtone, the star (*) key command required that all digits be re-entered from the start, whereas in voice, the word "Cancel" meant that the user could simply re-say the last digit he/she had entered. This type of inconsistency is confusing for a typical user, who will be required to be able to use both Touchtone and voice modes, depending on the type of phone he/she is calling from.

Changes were made to improve the interface and it was recommended that the performance of the speech recognizer be systematically checked by the designers and if necessary, trained on more representative Australian users.

Finally, it was recommended that serious consideration be given to providing a speech recognizer which allows continuous rather than isolated digit entry. This was due to the observation that people automatically tried to enter digits continuously, possibly this being the more "natural" strategy in speech behaviour.

The results of testing the product manager were very encouraging. Clearly, an important part of human factors work is to demonstrate the findings of any given study to the designers and managers of the product, so that they are convinced of the need to make changes to the interface. It was found to be especially effective to run one of the product managers as a subject so she could see where the interface fell down. This method proved far more powerful than any written document. Even a product manager who was familiar with the system ran into many of the same stumbling blocks as had the naive subjects. After the product manager had performed the tasks, the report on the overall findings was presented, together with examples of specific problems highlighted on audio-tape. Many of the points raised in the report were accepted readily, as the audience and the subject had all experienced the problems either first or second hand. The authors would strongly recommend this type of approach to those undertaking human factors work.

3. "SYSTEM B" STUDY

System B is another telecommunications service, similar in type to System A in that the same speech database was used for both. In System B, the user is required to dial into the service using a telephone, then enter a 4 or 5 digit number which results in connection to recorded voice announcements. Digit entry can again be achieved by either Touchtone or voice.

The intended market for this product is people with children, teenagers, elderly relatives, travelling students or pensioners in their families. It is, therefore, quite a different range of users which needs to be catered for than in System A, which had an international and national focus. System B must cater for users from a broader age range, but a national market only.

3.1 Method

This was the same basic method as used in the previous study with the following differences.

- Subjects: A group of 19 people were selected to take part in the study. These included males and females, of a variety of ages, including six children, aged 9-14.

- Data: Performance of the speech recognizer was logged more systematically than in the previous study.

- Quick Reference Card: In this study, a short User Guide which had been provided with the product, was shown to users. They could refer to it while completing their tasks. It was possible to gain some insight into how well the Guide was working by watching how subjects interacted with both it and the system.

3.2 Results and Discussion

Bearing in mind that the main purpose of the study was to evaluate the interface, the performance of the speech recognizer was measured as far as possible given the limited number of trials and subjects. Caution was used, however, in placing too much weight on the precise estimates, due to sampling considerations. The speech recognizer was found to perform better for adults than for children. As this product is intended for use by young teenagers, among others, it was felt that the speech recognizer would need to be improved, at least for this age group. In some respects, with only a 4-digit code to enter, rather than a 20-digit string, the accuracy level may not need to be as high for this product as for System A.

Many of the interface problems were the same as for System A. The Help message was very long-winded, being over two minutes in length with the additional information about destination selection included.

There was still the problem of people not understanding the concept of Touchtone, and not knowing whether their phone was connected to an exchange enabling Touchtone dialling.

As before, recommendations regarding improving both the interface and the speech recognizer were made. Suggestions were also made regarding the User guide with respect to issues of layout, consistency and content.

4. "SYSTEM C" STUDY

System C is a speaker verification product. This system allows enrolment of a person's voice into a database by having the speaker say digits, in a continuous string, into a telephone handset. System feedback and prompting is provided via a computer screen. When a user attempts to access the system, he/she logs in as a particular identity and the system prompts the user to say sets of digits into the phone. The characteristics of that person's voice are analysed and compared with the voice print of the claimed identity. If there is a closer match with a "world model" (based on a sample of English speakers, from the UK) than with the claimed identity, access is denied. If, on the other hand, the match with the claimed identity is better, then the user is accepted as that identity.

Clearly, in applications such as in banking and in telecommunications, it is vital that the false acceptance rate be very low indeed. In fact, Moody (1991) claimed that for this system "no one has yet been able to gain access as an impostor". With this claim in mind, the performance of the system was tested in the current study using an Australian sample.

4.1 Method

- Heuristic evaluation: Three human factors specialists interacted with the system to learn exactly how it worked.

- The System: The speaker verification system was set up so the subject was required to speak digits into a telephone handset. The software ran on an IBM compatible 386. Digit sets were displayed on the computer screen and the subject was required to say the digits continuously into the phone.

- Tasks: Based on the heuristic evaluation, an experiment was designed to test the performance of the product. Specifically, the false recognition (false acceptance) rate was measured.

- Subjects: A group of 30 people from the laboratories volunteered to take part in the study. Half were males and half were females, and some had "non-Australian" accents.

- Procedure: Each subject was tested individually. They were required to enrol themselves into the database, then, attempt to be falsely accepted when they tried to access 12 "model templates" which had been enrolled previously into the database. These "model templates" were based on 12 people selected from the laboratories. Six were males and 6 females. Two had "foreign" accents, one being a Danish woman, and the other a French man. There was also one man and one woman with an English accent. The rest had "Australian" accents.

- Rationale: The templates were randomly presented to subjects so they were unaware of the identity of each template. This is quite a strong test of the system, as we made it as difficult as possible for a would-be impostor. The rationale was that if people could "break into" the system with no knowledge of whom they were meant to be imitating, then it should be easier for a person with prior knowledge (and at worst, a trained imitator) to fraudulently access the system.

4.2 Results and Discussion

- The false recognition rate was found to be 2.3% for this sample.

- The range of false recognition rates ranged from 0% to 8.8%.

- 25 out of 30 (83%) subjects were able to illegitimately access at least one of the 12 model templates.

- Only 5 out of 30 (17%) subjects were unable to achieve improper access. Of these, 3 had English accents, one had an Australian accent with a very distinctive rhythmic pattern, and the remaining subject was an Australian with no special characteristics apparent.

- The number of templates (out of 12) that a single subject could illegitimately access ranged from a maximum of 4 down to zero. Five subjects accessed 4 templates out of a possible 12 (33%), 6 subjects accessed 3 out of 12 (25%), 6 subjects accessed 2 out of 12 (17%), 8 subjects accessed 1 out of 12 (8%) and finally, 5 subjects were unable to access any (0%).

- Although the majority of false recognitions occurred within the same sex, (50 out of a possible 180 templates - 28%) surprisingly, a few did occur across sex, (8 out of a possible 180 templates - 4%).

- One of the templates, that belonging to a woman with an English accent, was found to have the voice print which was most easily "broken into". This occurred for 16 of the 30 subjects (53%).

Clearly, the performance of the speaker verification system was not nearly as good as the designers claimed for the product. The reasons for this can only be speculated upon. It seems likely that the fact that an Australian sample was used to test the product had serious implications for the way the system performed. Why should this be the case?

Suppose the "world model" is based on a cross section of speakers from the UK. It could be that many of the Australian subjects' voices are closer to an individual model template (one of the 12 in this study) than to the UK world model. This would explain why the three English subjects could not access any of the templates as their voice prints were perhaps closer to the world model and more distant (distinct) from the template models. But what about the model template belonging to the woman with an English accent whose template was most frequently accessed by "impostors"? In this case, the subjects' voice prints must have been closer to that model template than to the world model. Possibly, this occurred because this particular model template was, in fact, so close to the world model that such confusions were possible. However, this did not occur in the case of the model template of the man with the English accent, so this explanation seems fairly weak. This is especially so given that one of the model templates, a man with a strong French accent was illegitimately accessed by 8 out of the 30 subjects. His voice print ought to have been very distinct from the world model, hence ought not to have been confused with other people's.

Whatever the true explanation of the false recognition rate found is, clearly, the results have demonstrated the need to tailor such systems to the market for which they are intended. It is dangerous when assessing speech systems, to try to extrapolate results from one country to another, simply because they share a common language.

5. CONCLUSION

The three studies, taken together, illustrate the importance of undertaking proper human factors evaluation and testing of products before they are launched into the market place. Such work ensures that systems are usable and do, in fact, cater for the precise needs of the people for whom they are intended. In the first system described here, the interface was thoroughly tested on representative users performing typical tasks. Changes to the interface were suggested based on these tests. In the second system, while the interface was very similar to that of the first system, the focus of the study was somewhat different, given the differences in the intended market for the product. In each case, it was recommended that the database for the speech recognizer be constructed using voice samples which are representative of the target market.

In the third system, the interface was much less important as it was very straightforward, so the emphasis of the work was on how well the system performed, but once again, focussing on the population for which it was intended, that is the Australian market, rather than the one for which it was originally designed.

Why is human factors important in the design cycle of these types of products? In the first two cases described, it is assumed that successful first-time use of both services is essential. Difficulties experienced in initial interactions with a system may deter people from using the service at all, even though a user might eventually learn how to use a "tricky" system if he/she persevered. There is clearly a trade-off between finding mechanisms which ensure that new users will be given all the support that is necessary, and keeping annoyance levels for the experienced user to a minimum. The type of testing carried out for these products uncovered many problems, some of which could be easily fixed by minor modifications to the interface. Others involved more comprehensive changes which the designers would need to evaluate as being worthwhile or not, given cost considerations. In the third system, the human factors evaluation was shown to be essential in testing the designers' claims for the product. The results had serious implications for the intended application.

Finally, the point must be made that ideally, any suggested changes for an existing interface should be subjected to further human factors testing to ensure that the modified system works well, and in fact is an improvement over the original one. This is currently happening for System A. Then, once released, field trials should be carried out to evaluate the usability and acceptance of the product in the market place.

ACKNOWLEDGEMENT

The permission of the Director of Research, Telecom Research Laboratories, to publish this paper is hereby acknowledged.

REFERENCES

Moody, A. (1991) "Speaker Verification", (Ensigma Ltd, Archway House, Welsh Street, Chepstow, UK).