AUTOMATIC SPEECH RECOGNITION: THE USER'S VIEW

A.F. STARR*, S.M. HUDSON+, D.M. JONES+ & C.R. FRANKISH"

ABSTRACT

As part of an Alvey-sponsored project on the human factors of speech system interfaces, users' views of human factor aspects of automatic speech recognition (ASR) were elicited from over thirty sites in the UK and USA. Comments from users and researchers suggested that: a) Many successful applications allow flexibility for the user in vocabulary selection and training; b) The acceptability of systems is aided by offering economy and naturalness of input; c) Recognition rates are currently too low to be acceptable in some critical applications, such as avionics. It was observed that successful applications were generally those where the original task offered a rigid structure, allowing ready application of syntax. The implications of these findings for research on human factors are outlined.

INTRODUCTION

The findings in this report are based on a series of visits to sites in the UK and USA where speech recognition is being used or researched. Users and researchers were asked about their ASR device, its mode of use and their preferences with regard to a range of human factors questions. The aim of this part of the project was to obtain users' and researchers' opinions which would allow identification of:

a) Problems encountered in applying ASR;
b) Areas requiring human factor guidelines for systems designers;
c) Problems that require experimentation to establish the guidelines.

HUMAN FACTORS ISSUES

Comments from users and researchers have been broken down into a number of topics which have implications for the guidelines.

Dialogue Design

The ability of users to remember the relations between spoken commands and required system responses has an important bearing on vocabulary size. Generally, vocabularies of less than 40 items appeared to cause no memory problems for users trained on the original task (before the introduction of speech input). The task itself often provided intrinsic prompts; e.g. faults noticed in inspection, and symbols for digitisation in cartography, were present as part of the task. In avionics applications this type of prompting was not always available and memory problems for recalling vocabularies of more than 30 items were reported. The value of intrinsic and other forms of prompting need to be established.

* Smiths Industries, Bishops Cleeve, Cheltenham. GL52 4SF.
+ Dept. Applied Psychology, UWIST, Llwyn-y-Grant, Cardiff. CF3 7UX.
" Dept. Psychology, Bristol University, Bristol. BS8 1HH.
With the exception of avionics, many applications allowed users to select their own vocabulary items. They replaced individual items that gave poor recognition performance, or were difficult to recall. In some cases this involved the whole vocabulary being changed.

One difficulty highlighted was recognizer confusions with natural pairs of words, such as 'on' and 'off'. When one of the items was changed, users were often confused about which word had been replaced. Changing both members of a pair seldom solved recognizer confusions and often gave rise to human ones, e.g. 'activate/deactivate' for 'on/off'. This problem needs further study.

ASR allows words or phrases to be used in the place of complex alphanumeric codes and number sequences, e.g. names instead of telephone numbers. This enhances the dialogue, making it more economical and natural, therefore easier to remember and use. Wherever this feature was offered, users praised it.

In most of the applications, the original version of the system had employed a strict task structure. The introduction of a similar syntax on the speech system, to improve recognition, was therefore readily adopted. It was not clear from the comments whether ASR would benefit from a more flexible syntax or from natural language inputs. This topic needs to be investigated, so that recommendations can be made for the design of ASR devices.

Feedback

All users reported a need for some indication that their message had been received. Visual feedback did not necessarily have to duplicate the verbal input. In many cases, activity in response to a command produced visible effects that were thought to be adequate (for example, by one display being substituted for another). In some situations, particularly those which were 'eyes-busy', auditory feedback was also thought to be effective, e.g. simple tones indicating recognition success, or a voice output message of some kind. However, voice output was sometimes annoying because of the length of time it took, the accent used (in the UK a North American accent was disliked), and the sex of the speaker. In telephone applications, the voice should match users' expectations, whereas in an auditorily-cluttered avionics environment, the novelty of a female voice was thought by some pilots to be an advantage. Users complained when no information or auditory information only was provided to indicate changes in the activation state of the device. Recommendations for feedback may have to be strongly qualified by the task being performed.

Error Correction

Various methods of error correction were used. These included clearing an entire data field and re-entering all items (this was favoured in avionics applications), and deleting one item at a time by moving a cursor backwards. Overwriting of the last item was often possible and considered to be helpful. Subdividing long data fields into groups of items, individually accessible for correction, was found to be beneficial for complex information, such as navigation co-ordinates in avionics. However, no systematic approach to the selection of error correction procedures was found, nor was there any clear indication of when particular types of procedure should be used.
Training

Three sorts of training were considered: voice, task and template preparation. Voice training was not used in any application, but some users and system builders thought it could be a viable possibility. Pilots disliked the idea, but tended to have good control of their voices anyway as a result of using radio links.

In all applications surveyed, task training had been carried out in advance of ASR training, as the ASR system was always introduced to users skilled in the task. ASR familiarization and training were generally informal and often carried out by an operator who was involved in the systems design or who showed interest in the technology. This situation was acceptable to other users, but the effectiveness of such a method was difficult to assess.

Updating of a template(s) at the time when recognition problems occurred was not considered by pilots to be feasible in avionics, but was performed in applications where the task could be halted temporarily. Initial template elicitation, and frequency of template updating, varied both between and within applications. Time between updates ranged from a few days to two years. This may be due to a combination of factors: the consistency of an individual's voice, the quality of the ASR system, the template-elicitation procedure, or users' expectations of the system.

Reversion

A reversion system can have two functions: it can be used as an alternative to ASR in the event of a complete failure, or to enter single items when ASR performance is poor. A reversion system is essential for critical applications such as avionics. It must be available at all times and be up-to-date with the current status of the ASR system. When recognition was poor in critical tasks, users preferred to employ the manual reversion system most of the time.

Immediate reversion was not possible in some applications, e.g. for inspection tasks, the only alternative involved paper and pencil and remote data entry. However, backup from a keyboard was available in most cases, and was used when an item was being poorly recognized, or a spoken vocabulary item was not recalled, or for error correction. Users did not anticipate that problems could be caused by differences between the vocabulary and syntax of the two methods. Research is required to establish criteria for reversion modes of ASR systems.

Workload, Stress and Fatigue

The ASR tasks were undertaken in conjunction with various other tasks, e.g. flying, visual inspection, data entry. The advantage of ASR was that it did not require diversion of hands and eyes from the primary task. It was not clear how these tasks interrelate with speech. More information is required on the consequences of combining speech with other tasks.

Workload and stress (physical, cognitive and emotional) were highlighted as problems in the avionics application. However their full effects on the use of ASR, and consequences on the perceived benefits of its applications, were not clear.
Vocal fatigue was thought to be a problem by some individuals, particularly when the voice was used all day. Some operators alternated use of the voice with other tasks such as manual input, or stored a separate set of templates to use in the morning and the afternoon.

Studies are required to establish the effects of stress, and to produce schedules for template training and updating and for the allocation of time spent on voice input and other tasks.

Recognition Performance

Users' expectations of recognition performance requirements for their application varied. Pilots felt recognition rates of 97% or above were necessary. Disabled users tolerated 70% but admitted that this caused frustration.

Microphones

The importance of microphone selection was apparent, but reasons why opinions differed were unclear. Inconsistent microphone position was often blamed for poor recognition performance. Headsets were sometimes seen as a status symbol but elsewhere provoked teasing, or were considered to be uncomfortable, or were disliked for "tethering" the user to the device. While telephone-style handsets were preferred by some users, others felt they were heavy and difficult to use.

CONCLUSIONS

Experimentation is required in the following areas in order to establish human factors guidelines for the implementation of ASR systems. To produce a system that is acceptable to the user and that is also efficient, the guidelines should encompass the following topics:

a) Selection of vocabulary items, syntax, and prompting.
b) Suitable methods of error correction and feedback.
c) Procedures for training the user and eliciting templates.
d) Interference from concurrent tasks and various types of feedback.
e) Effects of stress and fatigue.
f) Performance requirements of different applications.
g) Microphone selection.