EVALUATING AUTOMATIC SPEECH RECOGNITION AS A COMPONENT OF A MULTI-INPUT DEVICE HUMAN-COMPUTER INTERFACE

B.A. Mellor†, C. Baber‡ and C. Tunley‡

† Speech Research Unit, DRA Malvern, England.

‡ Industrial Ergonomics Department, University of Birmingham, England.

ABSTRACT

This paper reports on an investigation into the basic properties and requirements of automatic speech recognition as an input device to a trial human computer interface. The experiments required subjects to carry out a simulated target acquisition and report filling task, with the available input devices being automatic speech recognition, trackball, function keys or a simultaneous combination of all three. Experiments were carried out under varying workload to examine the degradation of overall interface and individual input device performance under user stress.

An approach at modelling interface performance using a critical path analysis approach is introduced. Modelling of the interface developed here has shown a good match to the experimental results.

Although use of the prototype speech recogniser was found to be both slower and less accurate than the manual mode inputs it was possible to estimate a required word accuracy of around 94% which would allow speech entry to provide an equivalent performance.

1. INTRODUCTION

It is clear that future human machine interactions (HMI) will be carried out via a number of interaction modes, rather than the manual entry and visual output modes currently used. Automatic speech recognition (ASR) will enable one additional mode of interaction. The choice of interaction devices and the combination of these devices in a true multi-modal interface will depend on many factors; individual component performance, user experience and ability, the computer application and fundamental properties of the various interface modalities can all be expected to influence the overall interface design.

This paper reports on experiments carried out at the Speech Research Unit, DRA Malvern, into the assessment of various interface designs for a simple target acquisition and report entry task. This task was chosen as being representative of many generic 'real world' data capture applications whilst being sufficiently simple to allow control over experimental design. Evaluation of the interface was based on completion time and task completion accuracy.

In addition to the main task, varying degrees of workload stress were provided to the experimental subjects by a way of a mental workload task. The effects of stress on the subjects' performances with the various interface devices was examined.

Development of novel interfaces would be aided by a suitable modelling tool. Such a tool would allow performance predictions of an interface design given atomic performance measures for individual interface devices, taking into account any conflicting requirements. This paper examines the use of an interface model based on a critical path analysis (CPA) approach, with the experimental results being used to validate the model.

In addition to modelling a prototype interface, it would be desirable, for the procurement of ASR devices, to have a specification for recognition performance which will provide the required system performance. The results of the trials are used to specify a recogniser word accuracy for the application developed.

2. DETAILS OF THE EXPERIMENT

2.1. Experimental Task

The experimental task chosen for the experiments involved target acquisition and report entry. The 'real world' application could comprise an operator controlling a remote pan and zoom camera to locate and report on objects of interest. Data entry and control in the experiments was chosen to be by keypad (which could be implemented as a wrist mounted unit), mini-trackball (hand held), speech input or a simultaneous combination of all three.

To simplify the display and remove problems related to legibility, the display shown at Figure 1 was developed. The lower part of the display presents the subject with a choice of simple geometric shapes, one of which is highlighted. The target acquisition task required the placement of the display cursor over the designated target before the issuing of a 'Select' command.

After selection of the correct shape, the subject issued a 'Zoom in' command to display an enlarged colour view with a direction arrow. The detailed view replaced the symbols at the lower part of Figure 1. The form at the upper left of the display is then filled out. All possible options for shape, colour and direction are displayed in the data entry form by way of labelled 'radio buttons'.

Report entry and cursor control for target acquisition were carried out via the three input devices. It was intended that each input device would be used in an appropriate fashion. The trackball was used as standard, with selection by device buttons. Data was

entered by selecting the individual radio buttons on the display. The ‘Zoom in’ command was selected via the Window display’s button.

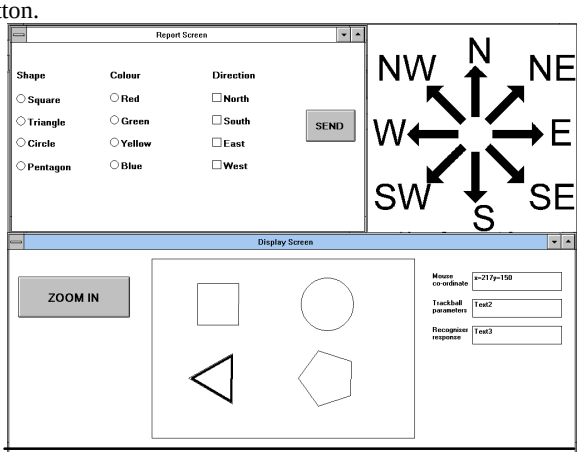

Figure 1. Screen display for target acquisition task. Data entry form is shown at the top left. The compass rose is for reference in the report filling section.

For target acquisition, the function keys provided speed increments in the selected direction, diagonal movement being allowed. The target was selected by pressing a central key. As with many key based interfaces, the focus of the interface jumps to relevant options. After target selection, the focus jumps to the ‘Zoom in’ button requiring only a single ‘entry’ key to proceed. Selection of entries in the data form was by ‘tabbing’ the focus of the display between fields and selecting the desired option using the ‘select’ key.

ASR control of the cursor in the acquisition task was by the spoken vocabulary “Up’’, ‘‘Down’’, ‘‘Left’’, ‘‘Right’’; providing directional speed increments. Target selection was by the vocabulary word ‘‘Select’’. After selection, only the spoken word pair ‘‘Zoom in’’ was available. Data entry was carried out by speaking the relevant data field names. A language model allowed two modes of data entry, either direct access to any field, or selection of a data group followed by a field in that group.

Workload to the subjects was provided by a mental arithmetic task played aurally during the trials. Mathematical problems were provided every five seconds, comprising simple addition and subtraction from a given starting point. At various points in the task, the subjects were required to provide the current total to ensure that the workload task was being carried out. Three levels of workload were trialled; no workload, medium workload which involved mathematical problems without carrying of digits (i.e. total between 0 and 9) and high workload which required mathematical problems with a carry (total between 0 and 19).

2.2. Experimental Implementation

The experimental environment described above was implemented on an IBM compatible personal computer in the GUIDE environment [B.A. Mellor, M.J. Tomlinson and N.J. Coleman, 1995]; a tool-kit of Microsoft Visual Basic routines intended to allow rapid prototyping of multi-modal interfaces. The window based working environment provided an event driven dialogue management system to control the overall interface, lending itself to a multi-modal interface with simultaneous inputs.

The target shapes and their properties were decided by a random number generator. Data on target types and operation timings were stored directly to computer hard disk. In addition, for ASR input, the recognised words were stored and the input speech signal recorded to allow a measure of the ASR device performance itself.

The speech recognition device used was based on the Speech Research Unit’s ‘AURIX’ rapid prototyping tool [R.W. Series, 1994]. The central processor unit runs a sub-word hidden Markov model based, fully continuous, connected word recognition algorithm. The model set used was composed by extracting suitable context sensitive sub-word units, using a decision tree method, from a large data-set trained on a large speaker, phonetically balanced corpus. This speaker independent and task independent model set, intended for rapid prototyping, limited the maximum performance of the ASR device. The vocabulary was constrained using a finite state syntax.

The hand-held mini-trackball used comprised a ball of approximately 8mm set in a square box of approximately 3cm to a side. Two buttons, placed to the left and right of the ball, were supplied for selection. The device was procured from a commercial source.

The function keys used were from the main PC keyboard, comprising an offset Maltese cross, with up as ‘T’, down as either ‘V’ or ‘B’ (to facilitate left or right hand operation) ‘F’ as left, ‘H’ as right and ‘G’ as enter/select. A mask was provided to cover all the remaining keyboard keys. Key strokes were captured by the controlling GUIDE application and processed by the dialogue management system to provide the relevant functionality.

2.3. Experimental Conditions

Six subjects took part in the experiments with two others used for initial trials. One input device was trialled per day, with the order randomised. All subjects were classed as expert, rather than naive users. Five trials were carried out per input, with the order of workloads being varied. Within each trial, the subjects were required to carry out six acquisition and report entry tasks.

The experimental trials were carried out in a soundproofed booth. Subjects were placed sufficiently close to the screen that legibility was not in question and with easy access to the input devices used.

The mathematical problems used for the workload task were played from a pre-recorded DAT tape. After validation of the maths task, the subject was prompted with the correct running total and the trial continued. There were no pauses during the trials without the workload task. The mathematical problems were presented via far-field speakers rather than headphones.

The use of a SURE SM10 microphone as input to the ASR device minimised pickup from the prompts.

Before and after the experiments, the subjects were questioned on their subjective evaluation of the input devices, allowing comparison tests of overall preferences [C.J. Tunley, 1995].

3. RESULTS

3.1. Objective Performance Measures

Several objective measures of performance were available from the trials. These included time to completion of the whole task, time to completion of each task sub-component (e.g. acquisition, zoom-in and reporting), accuracy of the data entry, accuracy in completion of the mathematical workload task and ASR accuracy.

Tables 1, 2 and 3 below summarise the key performance criteria of task completion time and task accuracy for the trials for all subjects and workloads.

Input Device	All Workload	No Workload	Medium Workload	High Workload
ASR	7.1	6.4	11.5	8.7
Keypad	0.84	0.7	0	2.0
Trackball	1.4	1.3	2.0	1.4
Mixture	7.1	4.8	4.6	4.2

Table 1. Summary of report entry error rates. Values are average percentage error rate over all trials for all subjects.

Input Device	All Workload	No Workload	Medium Workload	High Workload
ASR	27±15 sec	24±15 s	35±11 s	35±16 s
Keypad	14±6 s	12±3 s	16±7 s	20±7 s
Trackball	12±3 s	11±2 s	13±3 s	15±5 s
Mixture	13±6 s	11±4 s	15±5 s	20±8 s

Table 2. Summary of task completion times for the entire task, times averaged over all speakers over all trials.

Input Device	All Workload	No Workload	Medium Workload	High Workload
ASR	14±10 s	12±12 s	15±6 s	16±9 s
Keypad	9±3 s	8±2 s	10±2 s	11±4 s
Trackball	9±4 s	8±2 s	11±5 s	13±5 s
Mixture	9±5 s	8±3 s	11±5 s	13±9 s

Table 3. Summary of task completion times for the report filling task, averaged over all speakers over all trials.

Based on completion time and accuracy, it can be seen that the trackball provided the highest overall performance. This superiority was dominated by the target acquisition phase, as the report filling task indicates that the keypad was both quicker and more accurate than the trackball at all workloads. It is clear that for a complex task, the choice of optimal interface devices will depend heavily on all the styles of interaction required.

ASR provided the lowest performance interface, a result which was dominated by very variable recognition performance. The degradation of performance with workload appears less than for the other inputs. This is probably due both to the initial low performance and a failure to manage the workload task. It was generally found that the rehearsal mechanisms used to maintain the workload task running total interfered with the speech input mode more than the other inputs.

Most subjects used a combination of speech and trackball inputs in the mixed input experiment. It is clear that using the trackball to aid in the target acquisition provides a significant improvement in performance over the use of ASR in isolation, but does not match the performance of trackball input for report entry.

Figure 2 shows the variation of task completion time against speech recogniser word accuracy, for all three workload conditions. The plot shows that, predictably, the task completion times decrease with increasing ASR performance. The variation of performance is high, both between and within subjects.

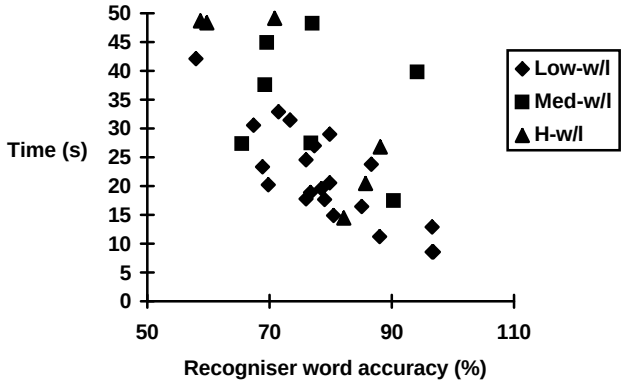

Figure 2. Scatter plot of whole task completion times against ASR percentage word accuracy for each workload level, low medium and high.

3.3. Subjective Results

Subjectively, the trackball proved the preferred input device for target acquisition, which was expected for a spatial positioning task. The keypad was generally preferred over ASR control for target acquisition especially when ASR performance was poor.

For report filling, most subjects rated ASR as the preferred input device followed by the trackball. Keypad control was preferred over trackball entry, possibly due to the highly structured menu design. The preference for ASR held even if the performance of the ASR device was poor.

All subjects reported that the workload task did not provide a significant problem unless an error occurred, at which point, the extra requirement for error spotting and recovery led to a rapid breakdown in overall performance. This affected the ASR more seriously than the other input modes.

It was noticed that the increase in workload affected the report filling task more than the acquisition task for all input modes.

This could be due to the more textual demands of the form filling task compared to the spatial manipulation task of target acquisition.

4. CPA MODELLING

As more complex and demanding multi-modal computer interfaces are developed, the requirement for a suitable interface modelling paradigm becomes increasingly obvious. Ideally, such a model would take atomic performance figures from the human factors literature and predict relevant performance metrics for given interface designs for a computer application. This should be extensible to allow prediction of variations in performance due to environmental factors, such as vibration (which will degrade use of keyboard entry and visual displays), noise (which will degrade speech input and output) and hands-busy operation (which would mitigate against use of a mouse or keyboard).

An approach to such interface modelling has been attempted using a critical path analysis approach. For the task described in this paper, a sequence of interface events was identified, along with competing events from the workload task. The time to completion for each atomic event was taken from either open literature or from previous research e.g. [B.A. Mellor and C. O'Connor, 1995].

Use of this model to predict the performance of the overall task using the trackball, without workload stress, predicted a time to completion of 13.2 seconds, compared to the actual value of 13.2 ±3.1 seconds. As ASR accuracy will affect overall completion times, Table 4 shows predicted times for various accuracies. In the trials, the mean overall time for task completion with ASR was measured at 34.7±11 seconds with an average word accuracy of 74 ± 20%.

ASR Accuracy (%w/a)	Predicted completion time (secs)
60	31.8
70	31.1
80	30.4
90	29.0

Table 4. CPA prediction of task completion time for ASR input for differing ASR performance.

By including an assumption as to how workload will affect the completion of the task, an attempt can be made to model the effect of stress on the interface. Table 5. shows the predicted completion times of the task given the extra workload tasks.

Workload	Predicted ASR Timing	Actual Time	Predicted Trackball Timing	Actual Time
No word	12.23 s	12±12 s	8.36 s	8±2 s
Medium Work	15.06 s	15±9 s	9.64 s	11±5 s
High work	16.28 s	16±9 s	10.87 s	13±5 s

Table 5. Predicted completion times for the report filling task for ASR and trackball inputs under workload.

From the results of the CPA modelling, we can suggest that this approach to interface modelling shows promise and will be developed further to investigate more complex interfaces.

5. ASR PERFORMANCE REQUIREMENT

Observing Figure 2, it is obvious that the ASR word accuracy strongly affects the overall performance of the interface. Given the time to completion for the other inputs, the ASR accuracy required to provide equivalent performance can be estimated. For these trials, this correspond to an ASR word accuracy, over the whole trial, of 94% without workload, and 84% with high workload. This latter figure is dominated by user error and hence a realistic figure would be higher than this. These figures have not yet been implemented in the CPA model of the interface to confirm the performance gain.

6. SUMMARY

An investigation of various interfaces to a generic computer application have been carried out. In general, an interface comprising a hand held trackball was found to provide the highest overall performance. A function keypad provided higher performance in a structured report filling task, but was less efficient at a spatial positioning task. ASR providing a speech mode of input was found to be subjectively popular amongst the trial subjects but provided the lowest actual interface performance. Using a trackball in conjunction with ASR provided higher performance than for speech input alone. It is expected that if the ASR recognition word accuracy could be improved from an average 74% to around 94%, on the task vocabulary, the ASR could provide an equivalent performance to the manual mode inputs.

A critical path analysis approach to interface modelling has been shown to provide a good approximation to the actual results suggesting that this approach may hold benefits for modelling of future, novel man machine interfaces.

7. REFERENCES

1. B.A. Mellor and C.O'Connor, *User Adaptation to Voice Input Interfaces*. Proc. ESCA. Workshop Spoken Dialogue Systems. Vigso, May 1995

2. B.A. Mellor M.J. Tomlinson and N.J. Coleman, *The Generic User Interface Design Environment, GUIDE - Overview and Features*. Proc. ESCA. Workshop Spoken Dialogue Systems. Vigso, May 1995.

3. C.R. Tunley. *A Comparison of Input Devices for the Proposed Future Fighting Soldier System.* Msc. Thesis, University of Birmingham, September 1995.

4. R.W Series. *The AURIX Speech Recognition Development Kit.* Proc IOA Conf on Speech and Hearing, Windermere, November 1994.