Manuscript submitted to *International Journal of Human-Computer Studies,* Version 1, September 6, 1995.

Speech versus Keying in Command and Control: Effect of Concurrent Tasking†

R.I. Damper*, M.A. Tranchant* and S.M. Lewis**

*Image, Speech and Intelligent Systems (ISIS) Research Group,
Department of Electronics and Computer Science,
University of Southampton,
Southampton SO17 1BJ, UK.

**Faculty of Mathematical Studies,
University of Southampton,
Southampton SO17 1BJ, UK.

† Based on a presentation at European Speech Communication Association (ESCA) Workshop on "Spoken Dialogue Systems: Theories and Applications", Vigsø, Denmark, May/June 1995.

Abstract

Speech input is frequently claimed to offer great benefits in human-computer interaction, yet experimental comparisons (conducted over many years) of speech and keying as input media have produced ambiguous results. The sources of this ambiguity must include task-related factors – some tasks lend themselves better to one medium than to the other – but also details of the experimental methodology employed. In an attempt to assess the importance of methodological factors and as a contribution to the continued development of techniques for the assessment of speech technology, we have previously reported [R.I. Damper and S.D. Wood (1995) "Speech versus keying in command and control applications", *International Journal of Human-Computer Studies*, **42**, 289–305] experiments modelled on the influential work of Poock in the early 1980s. Poock's work purportedly showed a very significant superiority for speech over keying in a command and control application. It was our contention that these findings were an artifact arising from the methodological flaw of selecting the command vocabulary to suit the requirements of speech input while making little or no concession to the requirements of keying. We removed this putative flaw by using abbreviated rather than full command in the keying condition and the claimed superiority disappeared. The implication drawn was that a fair comparison of input media requires that each is used along with an interface design tailored to the particular capabilities and requirements of that medium.

There were, however, other differences between our experiments and those of Poock so that other interpretations of these earlier findings are possible. Most notably, his subjects carried out a concurrent, secondary task while ours did not. Since speech input is generally considered to be advantageous in such situations, this difference is potentially important. Also, we did not replicate Poock's original experimental condition of full command keying, so that no very direct comparison of results for this condition was possible.

In this paper, we describe new experiments – again modelled on those of Poock – in which speech input, full command keying and abbreviated command keying are compared under conditions of concurrent tasking. We have tried to eliminate most of the avoidable differences between our studies: however, some remain. Additionally, we have employed an improved statistical design which allows more efficient testing of the statistical significance of observed differences than previously.

We find that speech input is no faster (a non-significant 1.23% difference) and enormously more error-prone (1038%, highly significant) than abbreviated keying, but allows somewhat more (11.32%, not significant) of the secondary task to be completed. Full keying has no advantages whatsoever, confirming the methodological flaw in Poock's work. Our subjects perform less well on speech input than did his under broadly equivalent conditions: we attribute this mainly to unavoidable differences in the two command vocabularies.

If recogniser errors (as opposed to speaker errors) are discounted, however, speech shows a clear superiority over keying. This indicates that speech input has potential for the future – especially for high workload situations involving concurrent tasks – if the technology can be developed to the point where most errors are attributable to the speaker rather than to the recogniser.

List of Figures

Figure 1: Schematic of experimental conditions for (a) original experiments without concurrent tasking and (b) new experiments featuring concurrent tasking. See text for details 22

Figure 2: Time for script entry versus run number for the three experimental conditions averaged across all 12 subjects 23

Figure 3: Number of errors for script entry versus run number for the three experimental conditions averaged across all 12 subjects 24

Figure 4: Number of characters transcribed in secondary task for the three experimental conditions averaged across all 12 subjects 25

List of Tables

Table 1: Experimental design employed in this study. KEY: S – speech; KA – keyed abbreviated; KF – keyed full 26

Table 2: Summary of results – figures given are averages per run across all runs and all subjects 27

Table 3: Entry time (seconds) averaged per run across all subjects for each combination of preceding and current condition 28

Introduction

It is widely assumed that automatic speech recognition (ASR) and speech synthesis offer the key to dramatic improvements in the effectiveness of the human-computer interface since: "For the majority of mankind, speech production and understanding are quite natural and unconsciously acquired processes performed quickly and effectively throughout our daily lives" (Furui, 1995, p. 9). Such assumptions about the usability of ASR (at least) have been criticised (e.g. Newell, 1985; Damper, 1993) as unscientific and unsupported either by the take-up of available technology or by human factors experimentation. Indeed, the argument that speech is somehow the 'natural' or 'universal' input medium is increasingly seen as flawed. Recent attention has moved to consideration of the optimal deployment of speech in multi-modal interfaces featuring a range of input devices and media (Martin, 1989; Damper, 1993; Sharman, 1993; Furui, 1995) based on rigorous, human-factors investigations.

Unfortunately, however, experimental comparisons of speech and competitor input media such as keying have – taken overall – produced equivocal results. For instance, Karl, Pettey and Shneiderman (1993) write: "Despite advances in speech technology, human factors research since the late 1970s has provided only weak evidence that ASR devices are superior to conventional input devices ... early studies taken as a whole are not conclusive." Summarising a comprehensive review of the literature, Martin (1989) states: "As these examples show, the results of formal comparisons between keyboard and speech are often contradictory and ambiguous." Earlier still, Simpson, McCauley, Roland, Ruth and Williges (1985; p. 121) write that experimental comparisons of the relative merits of speech input and conventional means of data entry to computers have often "produced conflicting results, depending upon the unit of input (alphanumerics or functions) and other task-specific variables." In other words, the particular application

and specific requirements of the interface design play an important part. While this must undoubtedly be so, another possible reason for such conflict could be the difficulty of deciding what is an appropriate methodology to effect a fair comparison between different input media.

In this paper, we report our latest work comparing speech and keying in the context of a (simulated) command and control application. Earlier work (Damper & Wood, 1995) was modelled on the influential study of Poock (1980; 1982) which is notable for the degree of superiority which it purports to show for speech over keying. It was our contention, however, that Poock's experiments were unfairly biased in favour of speech relative to keying – by unnecessarily requiring subjects to key verbose commands in full rather than in abbreviated form. When we removed this putative flaw by using abbreviated rather than full command in the keying condition, the claimed superiority disappeared. The implication drawn was that a fair comparison of input media requires an interface design that explicitly attempts to minimize the so-called transaction cycle – the number of user actions necessary to elicit a system response – for each medium.

There were, however, other differences between our experiments so that other interpretations of these findings are possible. Perhaps the most notable such difference is that Poock's subjects carried out a concurrent, secondary task while ours did not. Since speech input is generally considered to be advantageous in such situations (North, 1977; Mountford and North, 1980; Wickens, Sandry & Vidulich, 1983; Berman, 1984, Damper, Lambourne & Guy, 1985; Martin, 1989) by allowing a classical separation of modalities, this is potentially important. Also, we did not replicate Poock's original experimental condition of full command keying, so that no very direct comparison of his results with ours for this condition was possible. The reader is referred to the earlier publication of Damper and Wood (1995) for very full details and discussion of the issues involved.

Our latest experiments are again modelled on those of Poock: speech input, full

command keying and abbreviated command keying are compared here under conditions of concurrent tasking. We have tried, in fact, to minimise all avoidable differences but since Poock used a military, classified system, some details of his work (e.g. the precise command vocabulary) are not reported, which limits the degree of repeatability. In particular, his subjects were familiar with an existing command and control system, whereas ours were not. For this reason, we employed a pre-determined on-screen script to prompt subjects, while Poock's subjects decided for themselves which commands to enter according to a "fixed scenario". In the current work, we have also used a more efficient statistical design, and a more rigorous method of data analysis. Because of the limited information they yielded in our earlier study, no questionnaires were administered. Again, the hypothesis under test is that the observed superiority of speech over keying in Poock's experiments arises from a methodological flaw. (In other words, the omission of concurrent tasking by Damper and Wood does not explain the observed superiority of keying over speech in this latter study.)

The remainder of this paper is structured as follows. First, we briefly review our previous work aimed at effecting a fair comparison of the two media. Next, we describe new experiments extending the previous study, principally by the addition of concurrent tasking to the experiments. The results are then detailed together with appropriate statistical analyses, before concluding with some discussion of the implications of these results both for the experimental comparison of input media in human-computer interaction and for future applications of speech recognition technology.

Previous Work

The study of Poock (1980; 1982) is notable in the human factors literature for the degree of superiority which emerges for speech over keying. His subjects entered (simulated) military command and control instructions on the ARPANET. Speech was found to be

17.5% faster than keyboard entry while there were 183.2% more errors for keying than for speech. Also, speech input allowed subjects to transcribe 20% more (weather report) information as a secondary, concurrent task than was possible during manual entry. This work has been widely quoted in support of the view that command and control is one of the specific tasks in which ASR holds clear advantages. By contrast, we have questioned aspects of Poock's experimental methodology, arguing that this embodied an implicit bias towards speech and against keying. The putative flaw is that while commands had to be entered character-by-character when keyed, they were spoken as single (whole-phrase) utterances. There is no sensible reason why keyed commands should be entered in full: acronyms, key assignment or abbreviation-completion constitute the 'natural' language for a keypress interface. Hence, we set out to test the hypothesis that Poock's experimental scheme was deficient in suiting the requirements of speech input but making little or no concession to the requirements of keying.

Our earlier work (Damper & Wood, 1995) compared speech and keying in a situation similar to that employed by Poock, but using terser (and arguably more reasonable) commands for the speech condition. At that stage, no attempt was made to follow Poock's experiments faithfully in all other respects; we wished principally to explore the impact of command length. Where there were differences, we attempted to make these favour speech so as to provide a maximally stringent test of our hypothesis (that speech input was shown in an unduly favourable light). For instance, we used 28 distinct commands in place of Poock's "about 75": this should favour speech as recogniser error rate is expected to increase with vocabulary size. However, for practical reasons to do with their availability, our subjects had considerably less time for prior familiarisation with the speech recognition equipment (10 or 15 minutes as compared to an average of 3.26 hours). Although we used a comparable speech recogniser (Interstate SYS300) to that used by Poock (Threshold T600), ours was hosted on a stand-alone PC (Amstrad 1512) rather than on a distributed

network. Our 12 subjects (c.f. Poock's 24) were prompted by a pre-determined on-screen script of 79 commands, 28 of them being unique (c.f. Poock's "fixed scenario"). In all cases, keyed commands were entered as acronyms. In Experiment 1, subjects also entered spoken commands as acronyms (e.g. *gte* for GO TO ECHO). Exceptionally, however, single-word commands (such as REPEAT) were spoken as whole words in view of the difficulty of recognising single-letter names reliably. In Experiment 2, the same subjects spoke the commands in full (but keyed acronyms). Experiment 2 in particular was felt to effect a fair comparison of the input media since each medium was employed together with something approaching its 'natural' command language. Subjects repeated entry of the script 4 times (referred to as *runs* 1 to 4) for speech and for keying, in each of Experiments 1 and 2. The recogniser was trained at the outset of each experiment.

There were no systematic nor statistically significant differences between Experiments 1 and 2, so the results were pooled for further analysis. (Note, however, that results for one subject who was unable to use speech recognition at all effectively were excluded.) Contradicting Poock's results, we found that speech was 10.6% slower (although this difference was not statistically significant) and 360.4% more error-prone than keying (t-test, $p < 0.1\%$). This we interpreted as strong support for the hypothesis that Poock's methodology was flawed: we concluded that a fair comparison of input media requires that the experimenter explicitly attempts to minimise the number of user actions necessary to elicit a system response *for each medium*, just as one would do in a practical interface design.

There remain, however, alternative explanations of the differences between our experiments and the earlier ones. Poock's subjects were given a secondary task (transcribing weather report information) to perform in system idle time. Since we used a stand-alone PC as host, there were no non-deterministic network delays and so we omitted the secondary task. Thus, it is possible that the absence of concurrent tasking played a part

in the observed differences. This is plausible in view of the general agreement that the advantages of speech should come to the fore in situations of concurrent, manual tasking. To test this possibility, we have performed new experiments featuring a simulated, non-deterministic 'network' delay during which subjects perform a secondary task similar to Poock's.

To summarise, the differences between the new experiments and Poock's (apart from the use of abbreviated keyed commands and the more efficient statistical design) are that our subjects:

- are prompted for data entry by an on-screen script, rather than following a "fixed scenario";
- have considerably less time for prior familiarisation with the speech recognition equipment;
- did not complete questionnaire(s), probing their expectations of ASR beforehand, and attitudes to it following the experiment.

Concurrent Tasking Experiments

Our earlier experiments, without concurrent tasking, compared spoken abbreviated commands with keyed abbreviated commands (in Experiment 1) and spoken full commands again with keyed abbreviated commands (in Experiment 2), as shown schematically in Figure 1(a). This design required subjects to attend the laboratory on two separate occasions, one for each experiment, yet the results were pooled for analysis. Hence, some efficiency was possible – in particular, avoiding the need to repeat the keyed (abbreviated) condition in its entirety – in the design of the new experiments featuring concurrent tasking. Furthermore, the original design did not compare (full) speech input with full keyed input – thus ruling out any more direct comparison with Poock's results.

**** FIGURE 1 ABOUT HERE ****

Again, 12 subjects completed the new experiments, in which entry of a prompted script was repeated 4 times (runs 1 to 4). 8 were male and 4 female with ages ranging from 20 to 46 years, and none were touch typists. 5 subjects were engineering undergraduates, 3 were medical students and 4 were post-doctoral speech scientists (although only 2 of these had any ASR experience). 2 of the 12 subjects had extensive ASR experience. The experimental equipment was identical to that used earlier. Figure 1(b) depicts the three conditions studied which now include full keyed commands. In the experimental design, every subject took part in three sessions and experienced all three conditions, so that comparisons can be made within each of the subjects.

The detailed design employed is shown in Table 1 where I, II and III denote the session order. Subjects were allocated at random to the 12 sequences of conditions. The design is of the type proposed by Williams (1949) in that every condition is preceded by each of the other two conditions the same number of times (twice). This property enables the investigator to determine if there is a contribution to the measurements made in one session by the *immediately preceding* session. Account can be taken of any such 'carry-over' or 'residual' effects in the statistical comparison of the conditions: for further details see Jones and Kenward (1989). An additional advantage of this design, compared to that used by Damper and Wood, is that each subject was only required to visit the laboratory on one occasion.

**** TABLE 1 ABOUT HERE ****

As before, experiments were conducted in a sound-deadened, quiet room with speech commands entered using a Shure SM-10 head-mounted microphone. Subjects first trained the recogniser: this was followed by a short familiarisation phase of some 10 or 15 minutes

during which they were encouraged to vary their delivery of spoken commands and to observe the effect on recognition performance. They were then given a detailed instruction sheet to read, outlining the work to be completed. The concurrent task (to be performed in the simulated delay periods during which the computer system would not respond to inputs) consisted of transcribing weather information from a source data sheet onto a pro-forma by handwriting, as in Poock's study, using pen or pencil. Subjects were instructed to give the primary task (entry of the script) absolute priority over the secondary task.

Simulated network delays were rectangularly distributed in the range 0.2 to 7 seconds. The lower value of 200 ms was selected as just about on the limit at which most subjects would decide to transfer to the secondary task. Accordingly, at least some of the secondary task was completed between most primary data inputs. The upper value of 7 seconds was more arbitrary but allowed a significant amount of secondary material to be transcribed. Although delays were probabilistic, the total delay was held constant at approximately 265 seconds for all runs, so as to treat all experimental conditions equitably. Readiness of the system to accept data entry was signalled by an auditory 'beep' which served to divert attention from the secondary task back to the primary task. Acceptance of input by the system was signalled by a double 'beep', intended to be easily distinguishable from the other auditory signal.

The rejection threshold on the recogniser was set maximally low so that the error rate due to rejections was effectively zero (as confirmed by observation during the experiments – see below); that is, the recogniser accepted almost any input as a within-vocabulary utterance. The reason for this was technical: communication between the SYS300 and computer was such that it was not easy for the computer to ignore an incorrect input and carry on. (This was also the case in the work of Damper and Wood, although we omitted to document the fact in the earlier paper.) Hence, the overall error rate is certainly higher than would otherwise have been the case, since rejection errors could not be traded against

errors of other kinds (deletion, insertion and substitution) to achieve an optimal balance. Also for technical reasons, there was visual feedback for the keying condition but not for the speech condition. We do not believe this difference was important, largely because the high-workload demands of the task were such that subjects paid little attention to the feedback. A consequence of this arrangement is that (unlike Poock's subjects) our subjects did not correct errors entered during the experiment.

Subjects were observed, primarily to distinguish speech errors on the part of the speaker from errors on the part of the recogniser (but also to note any unusual circumstances affecting the results, although there were none). Breaks were given between the three sessions for as long as the subjects wished. This was never more than a few minutes, however, as they were generally keen to complete what was a rather tedious and demanding chore. The complete experiment took approximately 2.5 hours per subject.

Results and Statistical Analyses

The data were analysed using the split-plot-in-time analysis of variance (see Jones and Kenward, 1989, Chapter 6). The degrees of freedom in the F-tests were adjusted using the correction factor of Box (1954) to take account of possible correlations between observations on the same subject.

Figure 2 shows average input times for the 12 subjects under the three conditions (spoken, keyed abbreviated and keyed full) as a function of the run number. Summary entry time statistics averaged across all runs and all subjects are also given in Table 2. It is apparent from the figure and the table that spoken entry is essentially indistinguishable from keyed abbreviated entry in terms of entry time. Speech is actually 1.23% faster, compared to 10.6% slower in our earlier work without concurrent tasking, although these figures are not statistically significant. We were slightly surprised that the speech advantage was not greater: it was observed, however, that most subjects chose to reduce

the overhead of switching between keying and writing by retaining their pen in hand when keying. Figure 2 and the summary statistics of Table 2 also plainly show that full keying of commands slows entry very considerably: speech is 32.7% faster (c.f. Poock's 17.5%) than full keying, and this difference is statistically significant ($p < 1\%$) on the basis of Tukey's (1949) test.

**** FIGURE 2 ABOUT HERE ****

**** TABLE 2 ABOUT HERE ****

There are no evident trends for entry time to increase or decrease as a function of run number. The statistical analysis confirmed that there were no significant differences between different runs for the same condition.

Figure 3 shows the corresponding average number of errors – see also Table 2 for summary figures. The average error rate for speech is 14.11% (11.15 errors in 79 commands), compared to Damper and Wood's 8.68% (6.86 errors in 79 commands) and Poock's 3.20%. The increase relative to Damper and Wood is no doubt due to the introduction of concurrent tasking in the present work. The increase relative to Poock (given that a comparable recogniser was used) can be attributed to:

- reduced familiarity of our subjects with the recogniser and with data entry by speech;
- the fact that we could not adjust the rejection threshold so as to balance errors of different kinds (see earlier).

It is clear, however, that speech is enormously more error-prone (1038%) in the present work than abbreviated keying – to an extent which cannot be explained by a mere failure to optimise the recogniser's rejection threshold. On the basis of the Tukey test, this difference is statistically significant at the 1% level (although there is a strong interaction with run

number for the different conditions – see below). As expected, full keying has a much higher error rate (349%) than abbreviated keying, but this just fails to reach significance at the 5% level. However, full keying also has a considerably lower error rate (60.5%) than spoken data entry, and this difference is significant ($p < 1\%$). This latter finding is in direct contradiction to Poock's result that speech yielded many fewer errors than full keying. (See below for discussion of this discrepancy.)

**** FIGURE 3 ABOUT HERE ****

There was strong evidence from the statistical analysis for different relationships of error rate with run number for the three conditions ($p < 1\%$ in F-test for the interaction between condition and run). As seen in Figure 3, error rate increases with run number for speech, but decreases for full keying while remaining effectively constant for abbreviated keying. This is in contrast to Damper and Wood's earlier findings, whereby the error rates for both speech and abbreviated keying were constant across runs. It seems virtually certain that the demands of the concurrent task are causing divergence of the spoken commands from the exemplars supplied during training. That is, the cognitive load during the experiments is such that subjects are unable to allocate sufficient mental resource to maintaining an adequately constant level of pronunciation. On the other hand, task familiarity means that error performance improves with repetition for full keying. The same effect is not seen with abbreviated keying (either here or in Damper and Wood). This is presumably because the scope for such improvement is a function of the number of keypresses and so is reduced if not eliminated in the abbreviated keying condition.

Figure 4 shows the average number of characters transcribed as a function of run number under the three conditions: again, these figures are summarised in Table 2. As expected, speech allows more of the secondary task to be transcribed than does abbreviated keying. However, the difference (11.3%) is not large. Although there was

no questionnaire, a number of subjects volunteered the opinion that speech data entry produced perceptibly less interference with the secondary task. That is, it was noticeably easier to retain context information relating to the secondary task in short-term memory than in the keying conditions. Again, full keying is worse than either of the other two conditions by some margin.

**** FIGURE 4 ABOUT HERE ****

There was very strong evidence that, as a result of practice and familiarity, the amount of transcribed material increases with run number in a similar way across all three conditions ($p < 0.1\%$ in the F-test for comparing runs; no significant interaction between condition and run).

In general, there was no significant evidence of carry-over effects from the condition in the preceding session. However, there was some indication (see Table 3) that abbreviated keying in the preceding session produces a lower entry time for full keying in the current session than does spoken entry in the preceding session. It seems intuitively reasonable that there should be such a 'priming effect': keying the full commands amounts to practice in the immediately preceding session which may carry over to abbreviated keying.

**** TABLE 3 ABOUT HERE ****

Conclusions and Discussion

We have described experiments modelled on those of Poock in which speech input, full keying and abbreviated keying are compared under conditions of concurrent tasking in a (simulated) command and control situation. As in the earlier study of Damper and Wood, the hypothesis under test is that the observed superiority of speech over keying in Poock's

experiments arises from a methodological flaw. In the present work, however, the test is more stringent than previously as an important difference between our experimental conditions and Poock's has been eliminated – namely, the concurrent tasking of the subjects, who were required to transcribe weather information in system idle time.

We find that speech input is no faster and enormously more error-prone (1038%) than abbreviated keying, but allows somewhat more (11.3%) of the secondary task to be completed. Full keying has no advantages whatsoever, confirming the methodological flaw in Poock's work.

It is notable that our subjects performed less well on speech input than did Poock's under equivalent conditions of full keying. We find that full keying produces 60.5% *fewer* errors than speech, whereas Poock found full keying gave 183.2% *more* errors. Possible causes for this discrepancy are:

1. the failure to optimise the rejection threshold in our experiments;

2. the generally lower familiarity with speech data entry of our subjects, who had only 10 or 15 min practice as opposed to 3.26 hours for Poock's subjects.

3. the keyed commands for Poock's subjects may, on average, have been considerably longer than ours.

The first two of these have already been cited as possible causes for the difference in the *absolute* error rate. As stated above, however, it seems unlikely that the mere failure to adjust the rejection threshold (so as to trade optimally the different types of error) could account for the level of disagreement seen here. The lower familiarity of our subjects with ASR is a plausible cause, but Poock himself (1982, pp. 37–38) writes: "... there was no correlation between practice time and the errors entered".

Some light can be shed on this matter by supposing that reasons 1 and 2 did not apply, and we had obtained precisely the same error rate on spoken input as did Poock – instead

of an error rate which was actually a factor of 14.11/3.20 (= 4.409) higher. Dividing our average error rate (Table 2) by this factor, speech would then have produced 42.5% fewer errors than full keying: however, this is still some way short of Poock's 183.2%. It seems, therefore, that we cannot explain the discrepancy on the basis of the performance under the speech condition alone (factors 1 and 2). Rather, it is necessary to invoke differences under the keying conditions. Accordingly, we conclude that the main source of the observed discrepancy must be factor 3. This interpretation is also consistent with speech being *relatively faster* in these experiments than in Poock's (32.7% c.f. 17.5%).

Damper and Wood argue strongly on several fronts that their work – in spite of focusing on a study from the early 1980's – is relevant to today's technology. Nonetheless, the fact remains that recognisers like the T600 and SYS300 are obsolete. For this reason, we have given some consideration to the following possibilities for future work:

- repeating the experiments with a modern, state-of-the-art recogniser;
- simulating a high-performance recognition device of the future, using the now-popular 'wizard of Oz' technique (Fraser & Gilbert, 1991).

However, the specific details of our experiments (no rejection, subjects not required to correct errors) mean that there is absolutely nothing for a wizard to do. This leads to the idea of a 'virtual wizard', i.e. to simulate a perfect recogniser, we can simply ignore recogniser errors (see also Martin, 1989, p. 369).

By observation of the subjects, the speaker errors were determined to be 7% of the total (37 out of 530) with the remainder being misrecognitions. Hence, the error rate of 11.15 average errors per run would reduce to 0.7805, somewhat below the abbreviated keying figure of 0.98, if recognition errors are ignored. Thus, a *perfect* speech recogniser would outperform abbreviated keying, having a lower error rate and allowing more of the secondary task to be performed (although it would not be significantly faster). This

indicates that speech input has potential for the future – especially for high workload situations involving concurrent tasks – if the technology can be developed to the point where most errors are attributable to the speaker rather than to the recogniser.

Acknowledgements

The authors wish to thank Christine Shadle for constructive comments and criticisms throughout the conduct of this work. Simon Wood solved an intractable interfacing problem for us which enabled the experiments to proceed.

The idea of a 'virtual wizard' emerged in discussion at Vigsø with Louis Pols (Department of Linguistics, University of Amsterdam), Roger Moore (Speech Research Unit, DRA, Malvern) and Fergus McInnes (Centre for Communication Interface Research, University of Edinburgh).

We are grateful to our experimental subjects, who gave their time entirely freely.

References

BERMAN, J.V.F. (1984) "Speech technology in a high workload environment", *Proceedings of 1st International Conference on Speech Technology,* Brighton, UK, pp. 69–76.

BOX, G.E.P. (1954) "Some theorems on quadratic forms applied in the study of analysis of variance problems, II", *Annals of Mathematical Statistics,* **25**, 484–498.

DAMPER, R.I. (1993) "Speech as an interface medium: how can it best be used?", in *Interactive Speech Technology: Human Factors Issues in the Application of Speech Input/Output to Computers,* C. Baber and J.M. Noyes (eds.), Taylor and Francis, London, pp. 59–71.

DAMPER, R.I., LAMBOURNE, A.D. & GUY, D.P. (1985) "Speech input as an adjunct to

keyboard entry in television subtitling", in *Human-Computer Interaction – INTERACT '84*, B. Shackel (ed.), Elsevier (North-Holland), Amsterdam, pp. 203–208.

DAMPER, R.I. & WOOD, S.D. (1995) "Speech versus keying in command and control applications", *International Journal of Human-Computer Studies*, **42**, 289–305.

FRASER, N.M. & GILBERT, G.N. (1991) "Simulating speech systems", *Computer Speech and Language*, **5**, 81–99.

FURUI, S. (1995) "Prospects for spoken dialogue systems in a multimedia environment", *Proceedings of European Speech Communication Association (ESCA) Tutorial and Research Workshop on Spoken Dialogue Systems: Theories and Applications*, Vigsø, Denmark, pp. 9–16.

JONES, B. & KENWARD, M.G. (1989) *Design and Analysis of Cross-Over Trials*, Chapman and Hall, London.

KARL, L.R., PETTEY, M. & SHNEIDERMAN, B. (1993) "Speech versus mouse commands for word-processing: an empirical evaluation", *International Journal of Man-Machine Studies*, **39**, 667–687.

MARTIN, G.L. (1989) "The utility of speech input in user-computer interfaces", *International Journal of Man-Machine Studies*, **30**, 355–375.

MOUNTFORD, S.J AND NORTH, R.A. (1980) "Voice entry for reducing pilot workload", *Proceedings of the Human Factors Society 24th Annual Meeting*, Los Angeles, CA, pp. 185–189.

NEWELL, A.F. (1985) "Speech – the natural modality for man-machine interaction?", in *Human-Computer Interaction – INTERACT '84*, B. Shackel (ed.), Elsevier (North-Holland), pp. 231–235.

NORTH, R.A. (1977) “Task functional demands as factors in dual task performance”, *Proceedings of Human Factors Society 21st Annual Meeting,* San Antonio, TX, pp. 367–371.

POOCK, G.K. (1980) “Experiments with voice input for command and control: using voice input to operate a distributed computer network”, *Naval Postgraduate School Report, NPS55-80-016,* Monterey, CA.

POOCK, G.K. (1982) “Voice recognition boosts command terminal throughput”, *Speech Technology*, **1**, 36–39.

SHARMAN, R.A. (1993) “Speech interfaces for computer systems: problems and potential”, *Displays,* **14**, 21–31.

SIMPSON, C.A., MCCAULEY, M.E., ROLAND, E.F., RUTH, J.C. & WILLIGES, B.H. (1985) “System design considerations for speech recognition and generation”, *Human Factors,* **27**, 115–141.

TUKEY, J.W. (1949) “Comparing individual means in the analysis of variance”, *Biometrics,* **5**, 99–114.

WICKENS, C.D., SANDRY, D.L. & VIDULICH, M. (1983) “Compatibility and resource competition between modalities of input, central processing and output”, *Human Factors,* **25**, 227–248.

WILLIAMS, E.J. (1949) “Experimental designs balanced for the estimation of residual effects of treatments”, *Australian Journal of Scientific Research,* **2**, 149–168.

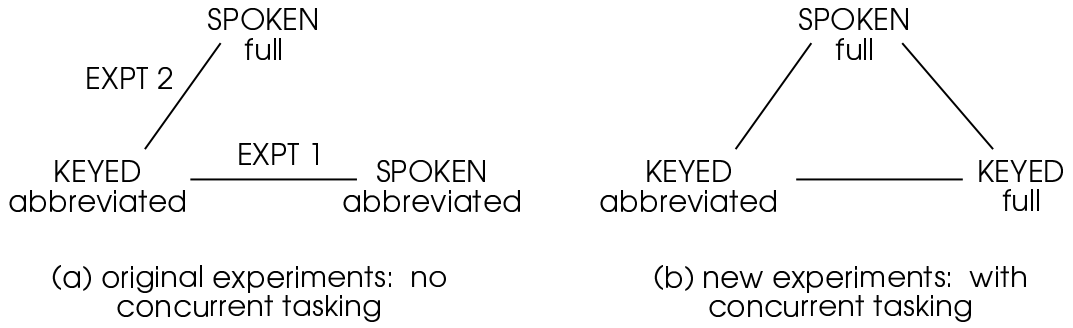

Figure 1: Schematic of experimental conditions for (a) original experiments without concurrent tasking and (b) new experiments featuring concurrent tasking. See text for details.

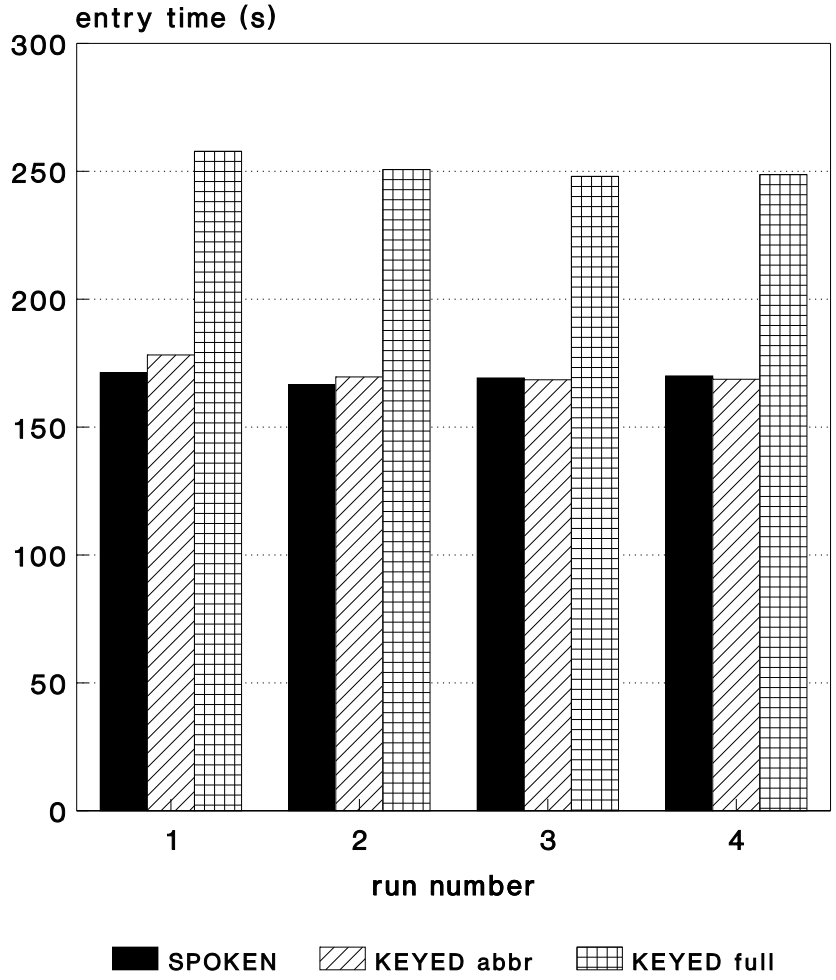

Figure 2: Time for script entry versus run number for the three experimental conditions averaged across all 12 subjects.

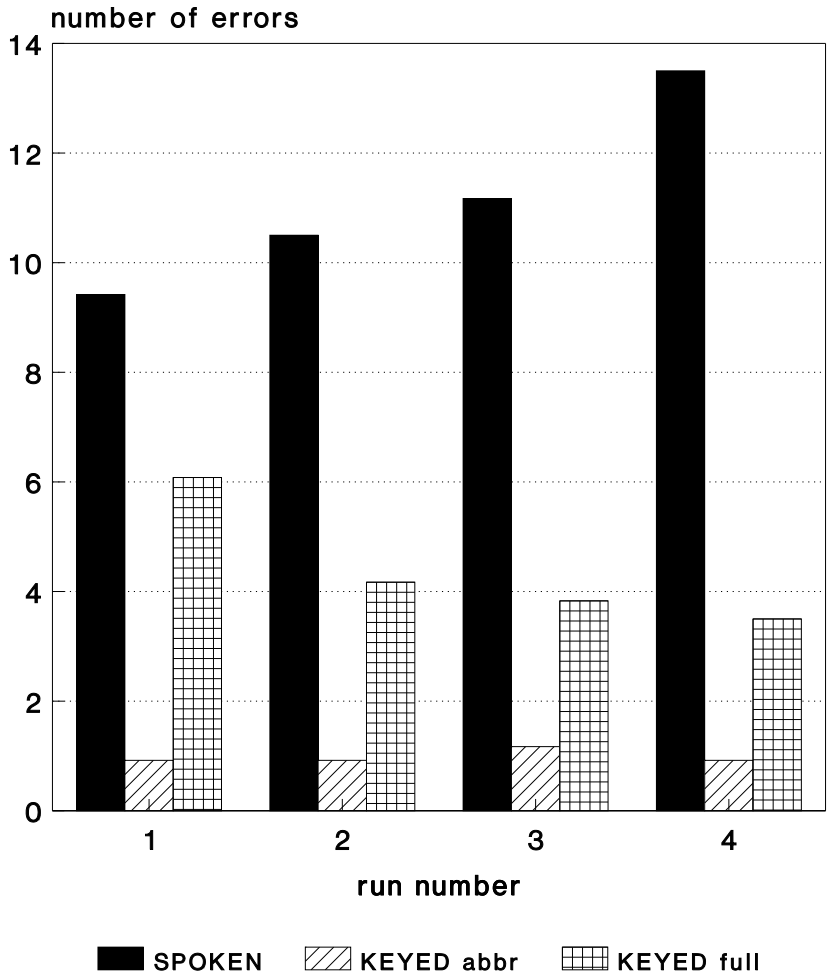

Figure 3: Number of errors for script entry versus run number for the three experimental conditions averaged across all 12 subjects.

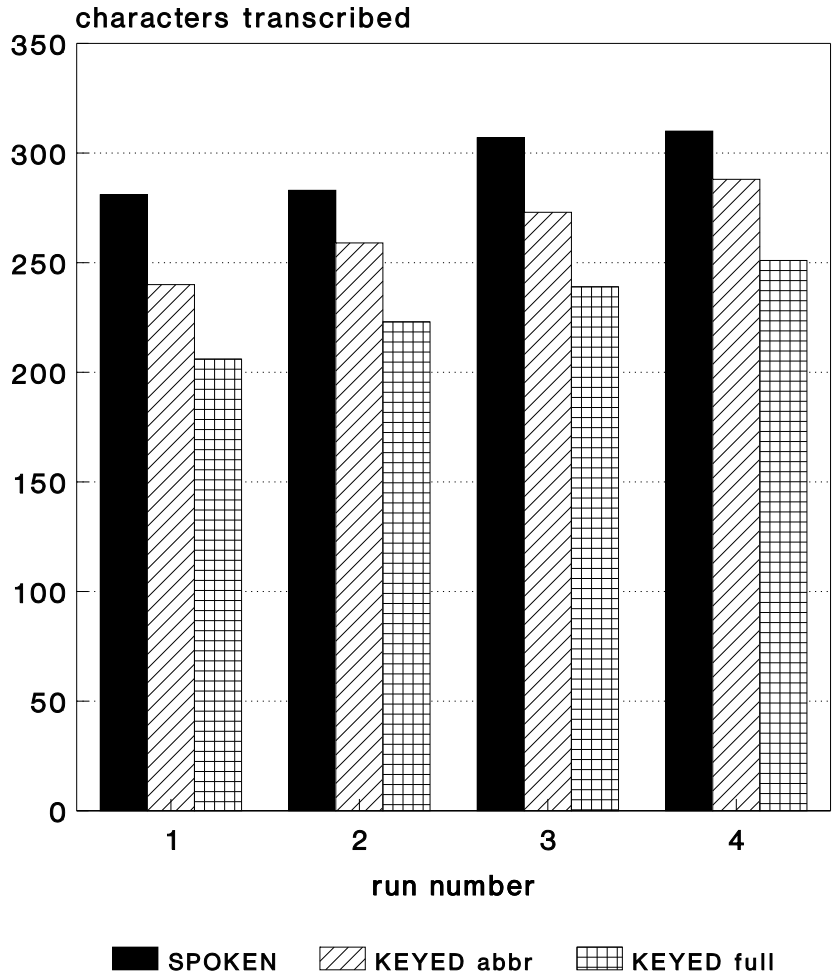

Figure 4: Number of characters transcribed in secondary task for the three experimental conditions averaged across all 12 subjects.

Subject	Session		
	I	II	III
1	S	KF	KA
2	KF	KA	S
3	KA	S	KF
4	S	KA	KF
5	KF	S	KA
6	KA	KF	S
7	S	KF	KA
8	KF	KA	S
9	KA	S	K2
10	KF	S	KA
11	KA	KF	S
12	S	KA	KF

Table 1: Experimental design employed in this study. KEY: S – speech; KA – keyed abbreviated; KF – keyed full.

Condition	Entry time (s)	Number of errors	Characters transcribed
Spoken	169.2	11.15	295
Keyed abbreviated	171.3	0.98	265
Keyed full	251.3	4.40	230

Table 2: Summary of results – figures given are averages per run across all runs and all subjects.

Preceding condition	Current condition		
	Spoken	Keyed abbreviated	Keyed full
Spoken	–	159.2	272.2
Keyed abbreviated	169.2	–	235.2
Keyed full	170.5	167.5	–

Table 3: Entry time (seconds) averaged per run across all subjects for each combination of preceding and current condition.