Speech as an Interface Medium: How Can it Best be Used?

R.I. Damper
Department of Electronics and Computer Science
University of Southampton
Southampton SO9 5NH

Abstract

Intuition, and some influential research in the experimental psychology literature, suggests that speech is the human's most natural communication mode. Based on this, there is a widespread assumption that speech represents the ultimate medium for human-computer interaction; yet speech technology remains little-used in real applications. A possible reason for this is ignorance on the part of systems designers of what the technology can do, or how best to employ it. While the art of designing speech-based systems continues to make steady, if slow, progress, there do seem to remain more fundamental problems. One obvious cause might be limitations in the capabilities of current technology (which we can hope to overcome in time). A further possibility is that speech, while offering some specific advantages, is actually a poor medium for many human-computer interaction tasks. This paper attempts to reach a balanced view of the advantages of speech relative to other interface media and, thus, of the likely rôle of speech in future interactive systems. We focus on input to keep the discussion tractable. Analytical and experimental methods for comparing speech with competitor media are reviewed and discussed.

1 Introduction

Over recent years, considerable research effort has been directed at realising major advances in the technology of speech communication with computers. This effort has been motivated by the assumption that speech technology offers the key to dramatic improvements in the effectiveness of the human-computer interface. An important, but usually implicit, thread is the notion that speech is somehow a *universal* medium – good for all situations. For instance, Lea (1980) writes:

> "...you will want to use speech whenever possible because it is the human's most natural communication modality."

while Viglioni (1988) says:

> "With speech recognition and speech response systems, man can communicate with machines using natural language human terminology."

and Lee (1989) states:

> "Voice input to computers offers ... a natural, fast, hands free, eyes free, location free input medium."

While the superiority of speech as an input medium is most often merely assumed, some authorities refer to the experimental psychology literature on problem solving (e.g. Ainsworth, 1988, pp. 3–4). For instance, Chapanis (1975) and his co-workers (Chapanis *et al*, 1977) have shown a clear advantage to the use of speech in cooperative problem solving between humans in terms of solution speed.

However, there are obvious and marked differences between human-human and human-computer interaction which mean that advantages in the former case do not *necessarily* transfer to the latter situation. Ainsworth (1988) gives one reason such transfer might occur:

> "Presumably as speech is a more natural form of communication it is possible to think and speak simultaneously, whereas complete sentences need to be composed before they can be written or typed."

On the other hand, if speech really carries all the advantages that have been claimed for it, one is entitled to ask why it has not been more widely used in interactive systems.

2 Why is Speech Not More Widely Used?

In spite of many years of optimistic predictions to the contrary, speech remains a very little-used medium of human-computer communication. Certainly, there have been successful applications – notably in hands/eyes busy situations, where mobility or computer access via the telephone network are important, or when the user is unable to operate more conventional input devices because of physical disability. However, these remain small-volume, essentially *niche*, markets. In many other cases in which speech has been tried, it has been found that conventional input devices serve as well as, and very often better than, speech.

There are perhaps three possible reasons why conventional input devices seem to serve as well as, or better than, speech in most circumstances. The first is that

successful integration of speech into an interactive system requires a profound understanding of the unique nature of the medium, and the development of new human engineering techniques, which are only now emerging. While there are good reasons for believing this, it does not seem to be the whole of the story. The second possibility is that current technological limitations (which we can realistically hope to overcome in the future) impose a burden on the user which compromises the application. Finally, it may be that speech (far from being 'universal') is actually a rather poor general-purpose interface medium, although it may still offer worthwhile advantages in a restricted set of special circumstances. This last possibility is not entirely independent of the first: if the advantages of speech are peculiar (i.e. restricted and not obvious), then human engineering techniques will need to be evolved to exploit them.

Newell (1985) gives several reasons why speech might be less than ideal as an interface medium. In his words:

> "... a major justification for the use of speech has been that it is the 'natural' method of communication for man ..."

and:

> "... in general, much greater thought must be given ... than is implicit in justifications of this nature."

An interpretation of Newell's view is that speech input should be assessed rigorously – rather than by assumption. To do this, we need measurable goals.

3 Human Factors Goals and Methodologies

Shneiderman (1987) has suggested the following human factors goals as suitable for quantifying the efficiency and usability of an interactive system:

- speed of performance;
- rate of errors;
- subjective satisfaction;
- time to learn;
- retention over time.

These, then, are the yardsticks against which any interface, including one based on speech, should be evaluated.

In this paper, we consider three ways that success in meeting these goals might be predicted, namely:

- by analysis of the input tasks involved;
- by case studies, either 'laboratory-based' or 'application-based';
- by so-called *Wizard of Oz* simulation – see Fraser and Gilbert (1991) for a comprehensive and up-to-date review.

4 Analysis of Input Requirements

If speech really is an effective and universal interface medium, then it would be expected to score highly on all the dimensions listed by Shneiderman. If on the other hand, there are particular applications for which speech is appropriate and others for which it is less good (as assessed by the measurable goals), we need some analytical framework in which to assess the likely success of a specific application.

One very useful classification of input (sub)tasks has been put forward by Foley, Wallace and Chan (1984). They were primarily concerned with graphics input, but their scheme can be usefully applied to other domains. Foley and his colleagues identify six 'primitive' interaction subtasks which they call:

- selection
- string
- quantify
- orient
- position
- path.

The selection function is illustrated, for example, by the common requirement to pick an item from a menu. The string function involves composing a sequence of characters selected from some set as in text composition; hence, it can be viewed as a sequence of selection operations. Quantify calls for the specification of some scalar (uni-dimensional) quantity denoting, for example, the point in some file where editing is to take place. Orient specifies an angular quantity, such as the orientation of a line segment. Position identifies a point in (usually) two-dimensional space –

effectively a dyad of quantify operations. Finally, path describes an arbitrary curve in the applications space and can be seen as a sequence of either position or of orient. Each of these 'primitives' can be seen as implying the existence of some *abstract device* which ideally implements that subtask.

Real, physical devices map onto the abstract devices in vastly differing ways. Thus, whereas the conventional, QWERTY keyboard is essentially a string composition device, it can be used to simulate the other abstract devices with greater or lesser facility. For instance, augmented with cursor keys, it can perform the position function. However, a pointing device such as a mouse is a far better realisation of the abstract position device. In turn, a data tablet used in conjunction with a stylus is a generally better realisation of the path abstract device than is a mouse, even though path can be viewed as a sequence of position, because a stylus has far better handling dynamics than a mouse.

Examination of the way that speech input maps onto the abstract devices gives a useful means of evaluating its strengths and weaknesses. Most obviously, a powerful speech recogniser (used as an automatic dictation machine) would appear to be a very promising text-composition (string) device. For a direct implementation, however, this application requires large-vocabulary capabilities beyond what is currently available commercially. Further, speech input is potentially an excellent means of *1-out-of-N* selection, especially for N (i.e. vocabulary size) of the order of hundreds, since this is very close to the recogniser's real mode of operation. By contrast, a keypress device with several hundred keys would be unwieldy in use. On the other hand, it should be readily apparent that speech recognition is a poor match to the requirements of, for instance, quantify and position. The former calls for an essentially analogue, continuous device (whereas recognisers produce discrete output), while position requires spatial, pointing abilities which speech does not really possess.

Given this analysis, speech certainly does not appear to be a universal medium. Rather, it meets the requirements of selection well, with the promise that future technology will deliver good string devices.

5 Case Studies

While the sort of analysis of input requirements described can be extremely helpful, human-computer interaction is a complex subject; it cannot realistically be reduced to the simple procedure of mapping real devices onto abstract devices. Typical of the many other dimensions which need to be considered are the user's physical situation and skills, the cognitive load imposed by the task and safety criticality. Because we do not fully understand all the factors impacting on performance, it remains mandatory to carry out task-related case studies. The basic methodology which has evolved is to use the insights offered by such means of analysis as do exist to design

a prototype interface which is then evaluated against the measurable goals, listed above.

A number of investigations has attempted to assess the relative merits of speech and keyboard input via case studies. However, according to Simpson *et al* (1985):

> "Research comparing speed and accuracy of voice versus manual keyboard input has produced conflicting results, depending upon the unit of input (alphanumerics or functions) and other task-specific variables."

However, the analytical framework provided by Foley, Wallace and Chan (1984) seems to provide an explanation for much of this conflict without ascribing it to imponderable, "task-specific" factors.

5.1 Small-Cardinality Selection

Consider first the issue of 1-out-of-N selection. Certainly, the entry of numeric data by keyboard in a simple, primary task (i.e. without concurrent, secondary tasking) is faster and less error prone than entry by speech (e.g. Welch, 1977). This is only to be expected as a comparison of speed of keypressing with the time to utter a word makes clear. Hershman and Hillix (1965) found that targeting the finger and depressing a key on a QWERTY keyboard took some 200 ms for a practised typist; the corresponding figure for unskilled typing is some 1000 ms when typing meaningful text (Devoe, 1967). By contrast, it takes some 400–800 ms to utter a typical spoken command with present-day recognisers imposing a further processing delay. In the case of discrete-word input, an additional overhead is introduced by the need for a distinct pause between inputs.

Given these figures, it seems unlikely that spoken entry of primary data of low cardinality (i.e. small N) using isolated words could ever be competitive with keying by even a moderately practised typist, except in special (e.g. hands-busy) situations.

An example of a study contradicting this pessimistic view is that performed by Poock (1980). 24 subjects followed a fixed scenario in which they entered commands (from a set of 90) typical of a naval application, such as *go to echo* and *forward message*, either by speech or keyboard. The recogniser used was the Threshold Technology T600. Poock claims that speech input is vastly superior to keying in this scenario. The stated findings were that speech input was some 17.5% faster than typing, while typing had enormously more errors (183.2%) than speech command entry.

Crucially, however, the spoken commands were entered as single utterances (speaking e.g. "go to echo" as a connected phrase) whereas the typed commands had to be keyed character-by-character on a QWERTY keyboard. This means that *go to echo*, for instance, would require 10 or 11 *separate* input (keypress) acts, each of which is time-consuming and error-prone. Yet for the purposes of selecting between

90 actions, there is absolutely no need for keyed commands to be this verbose; one-, two- or three-letter commands should be perfectly adequate – even allowing for them to have some sort of mnemonic content.

Rather than comparing speech and keypressing as input *media* as such, Poock is in fact comparing:

1. a reasonable interface design in which a single spoken command effects a selection with a much less reasonable design in which an excessive number of keyings is required to select that action, and;

2. the mapping of a real 1-out-of-N device (the speech recogniser) to an abstract 1-out-of-N selection device, with the mapping of a real string device (the QWERTY keyboard) to the same abstract 1-out-of-N selection device.

It is hardly surprising that the former mapping is more successful than the latter, regardless of the difference in input media. Indeed, it would be interesting to see how a 90-key pad allowing commands to be selected by a single keystroke – a much closer physical realisation to the underlying abstract device than the QWERTY keyboard – would have performed.

In an attempt to overcome these methodological objections, we have recently carried out a slightly simplified version of Poock's experiment (Damper and Wood, forthcoming). 12 subjects entered commands (from a set of 40 rather then 90) from a fixed naval search-and-rescue scenario in two different conditions. In both conditions, the keyed commands were identical; they were acronyms such as *GTE* for *go to echo*. In the first condition, spoken commands were 'natural', consisting of the whole phrase (e.g. *go to echo*) as in Poock's study. In the second case, however, they were the spoken equivalent of the keyed acronym (i.e. subjects said "g t e"), so that the requirement for subjects to recode the command string to an acronym was maintained across input media. The recogniser used (Interstate SYS300) was similar in specification and performance to that employed by Poock.

Results did not differ markedly between these two conditions. Overall, speech input was slightly slower (10.6%) than keyed input but not significantly so. However, using our more reasonable coding of the keyed commands, error rates for keying were very significantly lower (78.3%) than for speech. Thus, contrary to Poock's finding, our study indicates that speech is an inferior medium for the sort of small-cardinality selection which is typical of command and control applications.

5.2 String Composition

Recently, it has become possible to perform preliminary, experimental comparisons of speech and keyboard entry of text strings using large-vocabulary recognisers. Brandetti *et al* (1988) used the *TANGORA* recogniser developed by Jelinek and

his colleagues at IBM (Jelinek, 1985) in such a study. According to these authors, the prototype for the Italian language "recognizes in real time natural language sentences built from a 20,000 word vocabulary". Subjects entered a 553-word text on two occasions: "once they used the voice recognition capability of the system, and the other time they used the keyboard only". (Note the implication, confirmed by a personal communication, that the 'speech' condition is actually speech plus keyboard.) For 8 non-typist subjects, it required less time to enter the raw text using speech input than using the keyboard; the relative average figures are 15.5 min for speech and 22.4 min for keyboard. Errors were significantly higher for speech; an average of 7.75% of words in error for speech as against 1.75% for keyboard. The total time to produce *corrected* text was found to be comparable for the two conditions: 29.7 min for speech and 28.6 min for keyboard. (These figures, averaged for all non-typist subjects, are not given explicitly but have been computed from Brandetti *et al*'s published data.)

For 2 professional-typist subjects, total times were 40.0 min for speech as against 22.0 min for keyboard, with errors of 8.8% and 0.5% respectively. Subjectively, it is said that:

> "users found more pleasure and satisfaction in the usage of voice rather than keyboard".

Finally, the authors believe that use of a ...

> "... voice activated text editor indicated that large-vocabulary speech recognition can offer a very competitive alternative to traditional text entry".

6 The Simulation of Future Systems

When interactive systems are based on emerging, developing technologies, as in the case of speech, the usefulness of case studies is limited by the performance of available devices. Thus, much human factors work in this area could be criticised as over-concentrating on the shortcomings of currently-available equipment rather than on more fundamental interaction issues, thereby assisting the development of the next generation of interface technologies. According to Simpson *et al* (1985):

> "By simulating speech recognition hardware, various levels of speech-recognition capability can be controlled and evaluated experimentally".

Since such studies use a human operator hidden from the experimental subject to simulate the recogniser, they are generally called *Wizard of Oz* (WoZ) studies. The

hope is that data collected from such simulations will be invaluable in determining the maximally useful properties of future speech systems.

Gould, Conti and Hovanyecz (1983) performed an important, early WoZ study in which the capabilities of a simulated speech input device were constrained in various ways. They asked the question: "would an imperfect listening typewriter be useful for composing letters?" A human typist entering text on a conventional (QWERTY) keyboard was the basis of their simulated listening typewriter (SLT). This means that the simulated 'recognition' was not real-time. The SLT was compared with baseline performance set by handwriting. The particular imperfections studied were limitations of vocabulary size and the need for artificial pauses between utterances. Thus, the various different versions of SLT used had:

- either 1000-word, 5000-word or unlimited vocabulary;
- either isolated- or connected-word capability.

The vocabulary restriction was simulated by matching the typist's keyboard entries to words stored in a fixed-size dictionary. Words not in the dictionary could be entered in a spell mode. Feedback to the subjects used a visual display unit.

Subjects were professionals used to office work, including groups with and without dictation experience. They were required to compose letters by various means: namely, speech input using the SLT (called the "speech" condition), writing or dictation onto audio tape for later transcription. Note that the form in which the letter was produced is different in the different conditions. Subjects were asked to adopt two separate strategies:

- DRAFT – in which text-entry errors were to be ignored and left to a final editing stage;
- FIRST-FINAL – any errors encountered were to be corrected at the time of entry.

For the writing condition, however, only the FIRST-FINAL strategy was employed.

Results were assessed in terms of speed, subjective preference and the "quality" of the letters produced as judged by a panel of vetters. Figure 1 shows composition time only (i.e. excluding correction time) for the 1000-word and unlimited vocabulary conditions under both DRAFT and FIRST-FINAL strategies.

Overall, except for small-vocabulary, FIRST-FINAL as shown in the Figure, use of the SLT was found to be faster than writing but slower than dictation (although the latter does not produce output in 'real time'). As expected, the isolated-word SLT was slower than the connected-word input and speed generally increased with vocabulary size. The subjects mostly preferred the speech condition, presumably because

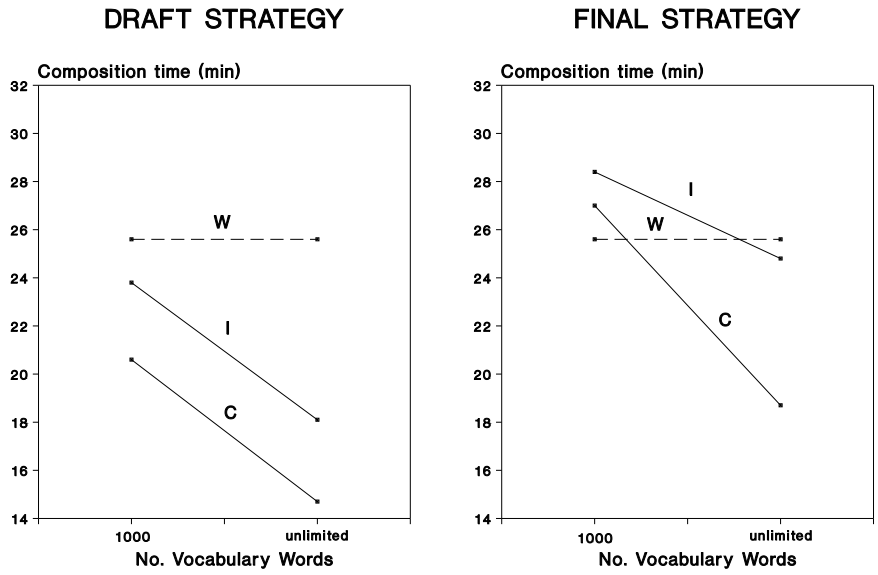

Figure 1: Mean composition times for isolated-word speech (**I**), connected-word speech (**C**) and handwriting (**W**), after Gould, Conti and Hovanyecz (1983).

of the combination of reasonable speed and observable results in 'real time'. Interestingly, large vocabulary size appeared slightly more important than connected-word capability. Finally, and not entirely surprisingly, letter quality was more a function of composition time than of means of input. This latter finding argues that, for creative writing at least, speed of input is not a vital issue since the slow process of composition is the limiting factor.

Gould, Conti and Hovanyecz's work was influential in establishing simulation as a legitimate methodology in the study of the human factors of speech input but does have a number of shortcomings. First, the use of a QWERTY typist sets an artefactual limit on simulated speech-entry speed according to the speed at which this keyboard can be operated. Importantly, it also means that a comparison with (conventional) keyboard entry – a much more serious competitor medium than handwriting, being both faster and yielding 'softcopy' – is not really possible. Second, the SLT does not simulate error patterns (substitutions, rejections, deletions and insertions) at all realistically. In fact, all errors were either rejections (out-of-vocabulary utterances or typographical 'errors' on the part of the typist) appearing as XXXX's on the screen, or simulated homophones (the highest-frequency token of a homophone pair was always selected.) Additionally, Simpson et al (1985) implicate "inconsistent restriction of discrete data entry when the spelling mode was used ..." as a shortcoming. Finally, post-editing of the DRAFT documents was done by handwriting: thus, the comparison of DRAFT plus edit times with FIRST-FINAL composition times is difficult since it involves a mix of media.

More recently, Newell and his colleagues (e.g. Carter, Newell and Arnott, 1988; Murray, Arnott and Newell, 1991) have addressed some of the shortcomings of the Gould *et al* work in their simulation of a speech-driven word processor (SDWP). In particular, they use a Palantype (machine shorthand) keyboard in place of the QWERTY keyboard, much reducing the response time of the simulated recognition system. Their studies have addressed the editing of documents by speech by providing speech as the *sole* input medium.

Input is effectively connected-word with a very large vocabulary of some 13,000 words. Using university students as subjects, composition rates for the SDWP (7.9 words/min) were not as high as those obtained by Gould and his colleagues (11.5 words/min) for the unlimited vocabulary, connected-word SLT using inexperienced dictators and the DRAFT strategy. Also, subjects were less impressed with the SDWP than were Gould's subjects with the SLT, all ranking speech as worse than writing even though the simulation was for full speed, connected speech with a very large vocabulary.

As far as speech editing is concerned, Murray, Arnott and Newell (1991) say:

> "a completely speech-driven word processor ... is an unsuitable system for the fast composition of documents."

Overall, simulation is a powerful way of assessing what would be useful features of future speech systems and, therefore, what ought to be the priorities for development. For instance, the finding of Gould, Conti and Hovanyecz (1983) that a listening typewriter restricted to isolated-word input might prove useful, provided it had a large enough vocabulary, would be difficult to obtain by other means.

7 Concurrent Input Featuring Speech

Broadly, it appears that speech input is not yet competitive with keyed input for primary data entry. However, speech may well hold an advantage when hands and eyes are busy. Thus, it might come into its own in cases of high workload or concurrent tasking when the use of an additional sensory or motor channel becomes important.

As well as examining data-entry performance for simple, random numeric and alphanumeric strings, Welch (1977) also studied performance in a "complex scenario". Subjects had to interpret an English language statement relating to simulated flight data and convert it mentally to a form suitable for entry in restricted fields. In this case, speech entry was faster than keyboard (for inexperienced subjects) and had a comparable "operational" error rate. However, the speech condition showed a substantially higher error rate before correction. Note the similarity to the results of Brandetti *et al* (1988), and of Damper and Wood (forthcoming) as described above.

Welch also added a button-pressing secondary task to the primary data-entry task. Although the secondary task did not impact significantly on speed or error rates for the simple data-entry scenario, the complex scenario revealed that the input speed using speech was degraded less severely than that of either keyboard or lightpen entry (Figure 2).

In similar vein, Mountfield and North (1980) studied a dual task in which (simulated) pilots had to keep their aircraft on course using a joystick while, at the same time, having to select radio channels either by speech or keypress. They found that tracking performance using the joystick was degraded very little when radio channels were selected by speech; however, errors were much increased with keyboard selection. Also, radio channel selection errors were higher for the keyboard condition.

It is tempting to explain the speech-input advantage in the dual-tasking studies described above on the basis that the speech channel is additional and does not interfere with the motor channel employed for keypressing in one or other of the tasks. Indeed, we have implicitly taken this to be true in justifying the utility of speech in the hands and eyes busy situation. However, on the basis of experimentation, Berman (1984) cautions:

> "the assumption that ... freeing a particular channel for inputs or re-

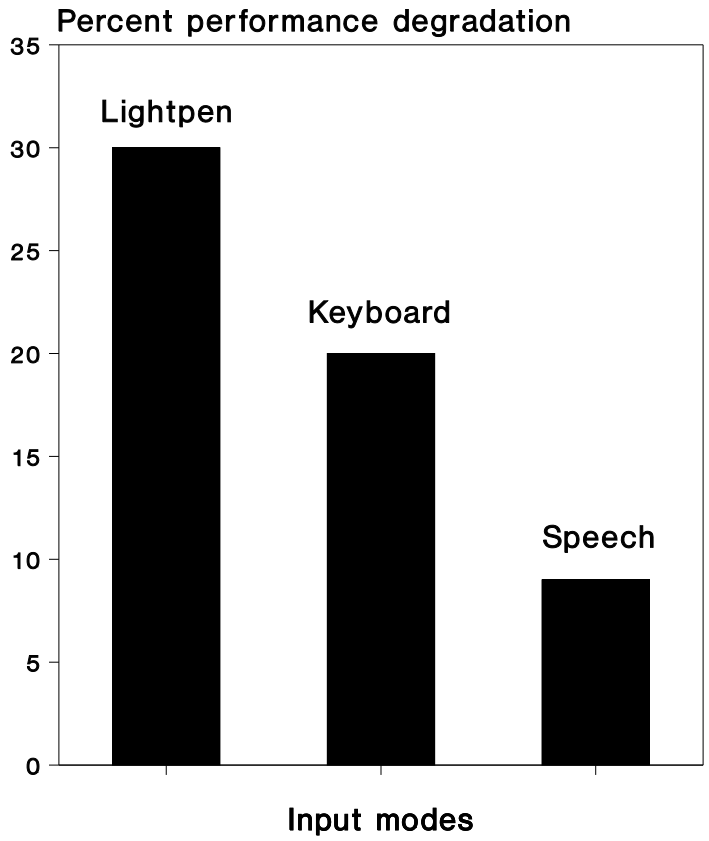

Figure 2: Percentage input speed degradation for lightpen, keyboard and speech entry when a secondary task is added to the primary data-entry task (after Welch, 1977).

> sponses must necessarily increase the number of tasks or items that can be attended to is not always true".

One relevant finding is the common observation that speech recognition performance under single-task (e.g. list-reading) conditions in the laboratory is always significantly higher than in real or simulated multi-task conditions (e.g. Biermann *et al*, 1985; Damper, Lambourne and Guy, 1985). The implication is that task stress competes for information processing resources which would otherwise be allocated to speech production, even when the concurrent task is manual in nature.

Berman (1984) considers a range of models of human information processing and their implications for dual-task interference when speech I/O is used in conjunction with other modalities. His favoured model is that of McLeod (1977), which extends Kahneman's (1973) model by replacing the "undifferentiated" single reservoir of processing resources by multiple reservoirs of resources. Berman states:

> "The implications for speech recognition are that, if it permits the use of a previously less utilised pool of resources, then the total processing resources involved will have effectively increased. However, it also raises the possibility that the change in response modality and, potentially, in encoding structures, may act to overload a previously-less burdened pool of resources, and hence cause an ... increase in task interference."

In a study of the use of speech recognition in television subtitling, Damper, Lambourne and Guy (1985) obtained results which can be interpreted in exactly this way. Speech was compared with keypad for the entry of "style" parameters (colour, on-screen position etc.), with a keyboard being used for subtitle-text entry in both conditions. It was found that speech input of style increased total preparation time by some 9% in spite of the fact that the time spent transferring between text and style entry was reduced. Significantly, and counter to our initial hypothesis of diminished task interference, there was a more than offsetting increase in the "time for apparently unrelated activities as evidenced by longer text-entry, 'between-subtitle' and idle times". This lends weight to Berman's belief that: "One should not be misled into anticipating workload reductions by adherence to an inappropriate information processing model" and reinforces the view that task-specific case studies remain essential given the current state of our knowledge in this area.

8 Summary and Prognosis

Simplistic thinking about the virtues of speech as an interface medium is no substitute for rigorous, scientific investigation.

Contrary to apparently widely-held belief, speech is a poor candidate as a universal input medium. It can act as a useful 1-out-of-N selection for reasonably large N,

and may offer advantages in situations of dual-tasking, but is otherwise a generally-poorer medium than keypressing at the present time. In the future, however, speech recognition should emerge as a powerful means of fast text (string) entry, but will need to be used in conjunction with a non-speech editing mode.

Wizard of Oz simulation can help define useful characteristics and rôles for future interactive systems featuring speech. The emerging generation of recognisers, such as *TANGORA*, can also form the basis of useful experimentation. In time, as the technology and human engineering know-how mature, it is to be hoped that the two empirical approaches of simulation and case study will converge.

References

AINSWORTH, W.A. (1988) *Speech Recognition by Machine*, Peter Peregrinus, London.

BERMAN, J.V.F. (1984) "Speech technology in a high workload environment", *Proceedings 1st International Conference on Speech Technology*, J.N. Holmes (ed.), pp. 69–76, IFS and Elsevier (North-Holland).

BIERMANN, A.W., RODMAN, R.D., RUBIN, D.C. AND HEIDLAGE, J.F. (1985) "Natural language with discrete speech as a mode for human-to-computer communication", *Communications of the ACM*, **28**, 628–636.

BRANDETTI, M., D'ORTA, P., FERRETTI, M. AND SCARCI, S. (1988) "Experiments on the usage of a voice activated text editor", *Proceedings Speech '88, 7th FASE Symposium*, Edinburgh, pp. 1305–1310.

CARTER, K.E.P., NEWELL, A.F. AND ARNOTT, J.L. (1988) "Studies with a simulated listening typewriter", *Proceedings of Speech '88, 7th FASE Symposium*, Edinburgh, pp. 1289-1296.

CHAPANIS, A. (1975) "Interactive human communication", *Scientific American*, **232**, 36–42.

CHAPANIS, A., PARRISH, R.N., OCHSMAN, R.B. AND WEEKS, G.D. (1977) "Studies in interactive communication: II. The effects of four communication modes on the linguistic performance of teams during cooperative problem solving", *Human Factors*, **19**, 101–126.

DAMPER, R.I., LAMBOURNE, A.D. AND 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), pp. 203–208.

DAMPER, R.I. AND WOOD, S.D. (forthcoming) "Speech versus keying: a human factors study", to appear in *Proceedings of Joint European Speech Communica-*

tion Association (ESCA) and NATO Research Study Group 10 Workshop on Speech Technology Applications, Lautrach, Germany, September 1993.

DEVOE, D.B. (1967) "Alternatives to handprinting in the manual entry of data", *IEEE Transactions on Human Factors in Electronics*, **HFE-8**, 21–31.

FOLEY, J.D., WALLACE, V.L. AND CHAN, P. (1984) "The human factors of graphics interaction techniques", *IEEE Computer Graphics and Applications*, **4**, 13–48.

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

GOULD, J.D., CONTI, J. AND HOVANYECZ, T. (1983) "Composing letters with a simulated listening typewriter", *Communications of the ACM*, **26**, 295–308.

HERSHMAN, R.L. AND HILLIX, W.A. (1965) "Data processing in typing; typing rate as a function of kind of material and amount exposed", *Human Factors*, **7**, 483–492.

JELINEK, F. (1985) "The development of an experimental discrete dictation recognizer", *Proceedings of the IEEE*, **73**, 1616–1624.

KAHNEMAN, D. (1973) *Attention and Effort*, Prentice-Hall, Englewood Cliffs, NJ.

LEA, W.A. (1980) "The value of speech recognition systems", in *Trends in Speech Recognition*, W.A. Lea (ed.), Prentice-Hall, Englewood Cliffs, NJ, pp. 3–18.

LEE, K-F. (1989) *Automatic Speech Recognition: the Development of the SPHINX System*, Kluwer, Dordrecht.

MCLEOD, P. (1977) "A dual task response modality effect: support for multiprocessor models of attention", *Quarterly Journal of Experimental Psychology*, 185–189.

MOUNTFIELD, S.J. AND NORTH, R.A. (1980) "Voice entry for reducing pilot workload", *Proceedings of the Human Factors Society*, 185–189.

MURRAY, I.R., ARNOTT, J.L. AND NEWELL, A.F. (1991) "A comparison of document composition using a listening typewriter and conventional office systems", *Proceedings of Eurospeech '91*, Genova, Italy, pp. 65–68.

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.

POOCK, G.K. (1980) "Experiments with voice input for command and control", *Naval Postgraduate School Report, NPS55-80-016*, Monterey, CA.

SHNEIDERMAN, B. (1987) *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, Addison-Wesley, Reading, MA.

SIMPSON, C.A., MCCAULEY, M.E., ROLAND, E.F., RUTH, J.C. AND WILIGES, B.H. (1985) "Systems design for speech recognition and generation", *Human Factors*, **27**, 115–141.

VIGLIONI, S.S. (1988) "Voice input systems", in *Input Devices*, S. Sherr (ed.), Academic, San Diego, CA, pp. 271–296.

WELCH, J.R. (1977) "Automated Data Entry Analysis", *Rome Air Development Center Report, RADC TR-77-306*, Griffiss Air Force Base, NY.