Proceedings of the Institute of Acoustics

EXPERIMENTS WITH A FULL-SPEED SPEECH-DRIVEN WORD PROCESSOR

I R Murray, J L Arnott, A F Newell, G Cruickshank, K E P Carter & R Dye

Microcomputer Centre, Dept. of Mathematics and Computer Science,
The University, Dundee DD1 4HN, Scotland.

ABSTRACT

The evaluation of a speech-controlled word processor or "listening typewriter" is described. The speech recognition part of the system was simulated using a machine shorthand transcription computer and trained human operator. The system was tested in a number of experiments which evaluated the performance of various features of the system. Guidelines for the performance of automatic speech recognition systems being developed for text composition tasks, and for the dialogue processors for such systems, are proposed.

1. INTRODUCTION

The development of automatic speech recognition machines has been an active research area for the speech community. One goal that is of great practical and theoretical interest is the possibility of using so called "listening typewriters" in creative writing tasks. Although speech has been claimed to be a "natural" method of input to word processors (Lea [1]), there is little empirical evidence comparing the relative merits of speech versus other input methods. As the performance of current speech recognition machines is not sufficiently well advanced, the technique of simulating such systems has been developed. The simulation is based on a concealed skilled typist operating a word processor acting as a speech recognition machine.

Gould, Conti and Hovanyecz [2] simulated a listening typewriter using a QWERTY keyboard in an extensive study, and concluded that "People will probably be able to compose letters with listening typewriters at least as efficiently as with traditional methods". It is often suggested that speech would be a "natural" way of controlling such systems, but Underwood [3] and Newell [4] have questioned the "natural" argument as a valid justification for choosing speech as the most effective and efficient method of interaction for word processing. Gould's editing facilities were very primitive, however, and thus the dialogue structure of the commands of a speech driven word processor remains an important research question. A research project was instituted as a continuation of Gould's work on simulating the ASR part of a voice-controlled text editing system. However, it was to have two main advantages over Gould's system: the interpretation of the user's commands was to be performed by a natural language interface within the computer system (rather than by the secretary) which allowed complex editing instructionsto be used, and the user was to be allowed to use natural speech rates (up to 200 wpm), rather than speech rates constrained by the use of a QWERTY keyboard as input device (up to 60 wpm). To permit these natural speech rates, the input system was based on the commercial palantype machine shorthand transcription system marketed commercially by Possum Controls, who collaborated on the project. This system, with a trained operator, can produce orthography from verbatim speech (with approximately 95% accuracy); verbatim rates are not possible on a QWERTY keyboard (the system used by Gould).

EXPERIMENTS WITH A SPEECH-DRIVEN WORD PROCESSOR

2. THE LISTENING TYPEWRITER

The listening typewriter simulation was implemented using a Sun 3/160 workstation and an IBM PC-AT executing the transcription software. The Sun workstation operated the natural language parser for processing the user's dialogue, as well as the text editor which the user viewed on the Sun monitor. Output from the transcription system and the subject's voice were recorded using a high quality PCM recording system to enable experimental sessions to be replayed. A video camera and recorder were used to record the user's movements when required. Transcripts of the palantyped speech were produced directly by the transcription system.

The speech-driven word processor system was to be used to:
(i) examine man-machine interface implications of the use of the speech modality
(ii) produce guidelines for speech input requirements
(iii) assess performance requirements of automatic speech recognition machines by realistic testing of the use of the speech modality
(iv) develop dialogue structures and design guidelines for continuous natural language speech input systems using an iterative cycle of development followed by trials in realistic environments and
(v) design a suitable human interface for voice composition of text.

3. EVALUATION OF THE LISTENING TYPEWRITER

The "listening typewriter" was tested extensively in a series of experiments to evaluate its performance and characteristics. Two pilot and eight formal experiments were performed, which tested different hypotheses about the system, as well as evaluating the changes which were made to improve the system. Full details of the system hardware, software and evaluation experiments are given in the project report [5]. The hardware is described by Dye and Cruickshank [6].

For all of the formal simulation experiments, the subject's spoken words were transmitted to the palantypist in another room, where they were entered into the transcription system. The orthography of the dialogue was passed to the natural language parser which interpreted the commands and passed corresponding instructions to the text editor. The output from the editor was displayed on the monitor in front of the subject. All subjects were asked to complete a questionnaire about the experiment at the conclusion of their session.

3.1 EXPERIMENTS 1 AND 2 - PARTIAL REPLICATIONS OF GOULD'S SIMULATION EXPERIMENT USING FULL-SPEED SPEECH RATES

Experiments 1 and 2 were carried out to replicate the earlier work by Gould et al. (op. cit.). In order to investigate any effects caused by using a simulation, half of the subjects were told about the existence and purpose of the palantypist (overt group), and for the others, it was implied that the system was a fully operational speech recognition machine (covert group). As with Gould, only primitive editing commands, such as available on a dictaphone, were made available in the experiment.

The composition rate achieved in the first experiment was 7.9 wpm, which was lower than those achieved by Gould (despite the faster transcription speed), with only 38% of the words spoken appearing in the final document, and the subjects all rated the system as worse than writing, and were less impressed by the system than those who had used Gould's system. Most notably, the subjects who knew they were using a simulation (the overt group) rated the system higher and were more impressed than those who were led to believe that they were talking to a machine.

It was hypothesised that the poor quality of the editor resulted in subjects dictating only one or two words at a time (despite the potentially high transcription speed) in order to avoid going back to make corrections. Thus, in an attempt to increase the composition rate, experiment 2 repeated the covert case of experiment 1, but the 22 subjects were told to ignore any errors made; it was expected that speech rates would improve if no editing was performed. One document summary was handwritten (as a control condition), one was done using the unrestricted editor, and one was done using the editor restricted to formatting commands (ie. no editing was possible).

The speech rates were found to be higher than in experiment 1, but higher for the unrestricted editor (27.0 wpm) than for the restricted editor (21.5 wpm). The hypothesis that speech rates would be higher when no editing was required was thus not supported; the lower rates indicate that the subjects became more cautious in their behaviour to avoid making errors with the restricted editor. The average composition rate for the unrestricted editor (10.9 wpm) was found to be similar to that achieved by Gould with inexperienced dictators (despite his subjects being office workers, and the subjects in the current experiment being University students). The average composition rate for the handwritten text (17.9 wpm), however, was substantially faster than the composition rate for the current unrestricted editor, and that of Gould's editor. The system was rated as similar to writing for the ease with which changes could be made, although half of the subjects said that it required a lot more concentration than writing. Experiments 1 and 2 are further described by Newell et al. [7].

3.2 EXPERIMENT 3 - GATHERING EXAMPLES OF TYPICAL USER DIALOGUES

The purpose of experiment 3 was to gather examples of subjects' verbal corrections of text as a precursor to developing the natural language pre-processor for the text editor. The 20 subjects were divided into two groups, naïve computer users and those familiar with using computers, and it was expected that the former group would use more verbose (and hence less efficient) commands. The subjects were presented with a series of paragraph pairs on the screen, and were asked to speak editor commands to change one paragraph to make it identical to the other. The edits were not performed on the screen (ie. no feedback), hence no cursor control stimuli were presented.

Both groups of subjects reacted positively to the idea of editing text using speech, and thought that it would be relatively easy to learn to use. Most subjects, however, found the task of initiating oral commands to an unresponsive machine quite difficult, being put off by the lack of feedback, and most evaluated their own commands as being unclear and at times ambiguous. Few of the subjects in either group said that they had used "natural language", most using a standard command format, and others trying to give the simplest commands to bring about the change. The freedom of choice of commands appeared to be a hindrance to many subjects, as they were unsure what commands they could use, despite being told that there was no restriction. Furthermore, the wide choice was seen to be very inefficient, and some of the subjects thought that voice editing would be slower than conventional methods, although they conceded that it might improve with practice.

3.3 EXPERIMENT 4 - LONGITUDINAL STATIC EXPERIMENT

Experiment 4 was a longitudinal static experiment to perform an in-depth case study using the listening typewriter, which now included the automatic natural language parser. The 5 subjects were all retired male executives unfamiliar with word processors, although 3 were experienced at dictating to a secretary. The subjects performed precis tasks and letter composition tasks, with a practice task and experimental task in each session. The system used for the experiment had three modes controlled by spoken natural language: text entry mode, command (edit) mode entered by saying "System" and exited by saying "Okay", and symbol (spelling) mode entered by saying "Symbols" and exited by saying "Okay". By comparing measurable performance and impressions of the system (by

questionnaire, before the first session and after the last), it was hoped that the experiment would show how easy the system was to learn to use, and show the general acceptability of speech as an input modality.

The natural language interface permitted the use of more complex editing commands than the earlier Gould-type system. For example, the following commands are typical of those that could be understood by the system:

"Delete this word"
"Delete the third word on this line"
"Capitalise the first letter of the second paragraph"
"Replace the next two words with the word spelt P E R F O R M"
"Insert an apostrophe after the fourth letter of the third word on the first line of the second last paragraph"

The average composition rate achieved was 4.6 wpm, with little difference between the experienced and inexperienced dictators, although the average speech rate of the experienced dictators (18.2 wpm) was slightly above that of the inexperienced dictators (15.4 wpm). The natural language parser performed well, with 87.6% of all commands being parsed correctly. Of those that failed, 3.4% were due directly to the inadequacy of the parser, with a further 1.1% caused by the parser timing out as the command would have taken too long to process. 2.1% of the errors were recognition (ie. transcription) errors, and 2.9% were mode errors. On average, only 80% of commands were correctly executed in the subjects' first session, but this increased to 94% by the final session.

Between the initial and final questionnaires, the subjects' rating of the system improved slightly from "a little better than writing" to "better than writing". The 3 subjects who performed all 10 sessions enjoyed the challenge of mastering the speech-driven word processor, and were generally impressed by the ability to see the spoken word almost instantly. However, it took them some time to grasp the restricted syntax, partly due to the need for precision (to avoid ambiguity) and for restraint in the complexity of commands, although the long time between sessions may have led to some commands being forgotten.

The experiment indicated that the "natural language" interface led to the use of commands which were not precise enough for the parser to handle correctly. The subjects indicated that natural language commands were a slow way of editing text, and in particular that cursor movements with the voice were difficult to perform. To offset the slowness, some subjects tended to rush some editing tasks, and so failed to check their edit command before sending it to the parser, hence producing another error and paradoxically compounding the slowness. Overall, the subjects were positive towards the system, the experienced dictators rating it as adequate, and the non-dictators rating it slightly higher.

3.4 EXPERIMENTS 5 AND 5A - THE EFFECT OF DIFFERENT FEEDBACK STRATEGIES

As composition rates in previous experiments had not exceeded 12 wpm, experiment 5 was carried out to compare the effect of various feedback strategies on composition rates. Word-by-word feedback (ie. isolated word) had been proposed (Martin [8]), although it had also been argued (Witten [9]) that feedback should be at the sentence level for connected speech recognisers, and it remained unclear how feedback should be presented to minimise interference with the subject's task. For the current experiment, 10 subjects were asked to compose documents using systems with normal feedback (ie. words appearing on screen as quickly as possible), feedback on syntactic marker

(ie. text appearing at the next full stop, period etc.), feedback on request (ie. text appearing when the subject said "Display") and mixed feedback (combining the latter two modes). It was hypothesised that speed would increase if a non-normal strategy was used. The screen layout of the editor was also improved from that used in experiment 4, as the subjects' performance in that experiment indicated that some aspects of the interface were not ideal.

Initial results indicated that the feedback strategy used had an effect on the composition rate. However, it was possible that the extensive editing performed during composition of some of the documents had biased the results, and a second experiment (5A) was performed to overcome this problem. This experiment was identical to experiment 5, except that different document outlines were used to limit the amount of formatting required by the subjects. The subjects rated normal feedback as the most favourable, with syntactic marker feedback the least favourable; statistically, only the syntactic marker strategy was significantly poorer than the other systems. Most subjects thought normal feedback a hindrance if they were sure of what they wanted to say, but noted that with the other strategies it was possible to "loose the thread" of what they were saying in mid-sentence. A sub-sample of subjects who generated and then edited their documents was selected, and the corresponding timings noted. This revealed that the subjects spent more time dictating than editing their documents, although subjectively in this and previous experiments, the subjects thought the opposite to be the case.

In neither experiment did composition rates exceed 10 wpm, and the average speech rate was less than 33 wpm. The efficiency of dictation was generally low, but improved slightly in the second experiment. Overall, the feedback strategies did not greatly affect composition rates, and it may be concluded that increasing the speed of a listening typewriter system is a non-trivial task, and that feedback changes are unlikely to have a major effect. Experiments 5 and 5A are further described by Carter et al. [10].

3.5 EXPERIMENT 6 - THE EFFECT OF DIFFERENT CURSOR MOVEMENT MODALITIES

This experiment was carried out to compare the effect on composition rate of various cursor movement modalities as part of the speech-driven word processor. The systems used were a touch screen, mouse and voice input. Six subjects took part in the experiment. The results showed that the subjects' speech rates were approximately equal in all three conditions. However, the average composition rate with speech for cursor control (5.11 wpm) was lower than with mouse control (6.02 wpm), and using the touch screen produced the highest rate (7.85 wpm); no subject exceeded 10 wpm for any modality. The composition efficiencies of the systems showed the same performance rankings (17.4%, 22.4% and 27.9% respectively). The subject's own preference ratings of the systems indicated that the touch screen was the preferred system, and speech-controlled cursor was least favoured.

3.6 EXPERIMENT 7 - ADDING VOICE INPUT TO A CONVENTIONAL TEXT EDITOR

This was a pilot experiment conducted with only 2 subjects to investigate whether the addition of speech input to an otherwise standard text editor would improve subjects' performance. A multi-modality editor was constructed, which allowed cursor movement by touching the text window on the screen, touching cursor keys on the screen, pressing the cursor keys on the Sun keyboard, and pointing with the mouse. Other buttons were available on the touch screen, including delete character, delete word and undo. Speech could only be used for text entry, punctuation and formatting - no spoken editing commands were available. Thus, no natural language interface was required for this system.

The average "without speech" composition rate was 13.9 wpm, exceeding the "with speech" rate of 12.8 wpm, although one of the subjects achieved a slightly higher rate in the "with speech" condition. The subjects spent approximately equal times using speech and using the keyboard. However, it is possible that the higher speed of spoken text entry could have been offset by the editing time necessary due to the larger number of errors than would have occurred with keyboard entry. It can only be concluded from these results that documents produced using speech were composed at about the same rate as reasonable "hunt and peck" typists, with the accuracy of the speech recognition system being approximately 95%.

3.7 EXPERIMENT 8 - COMPARISON OF THE SPEECH-DRIVEN WORD PROCESSOR, KEYBOARD-DRIVEN WORD PROCESSORS AND A DICTATING MACHINE

Experiment 8 was carried out to compare composition rates of documents produced using one of four systems: the speech-driven word processor (exclusively speech input), a dictation (into a dictating machine) and editing process, a simple keyboard-driven text editor, and the subjects' normal word processor. It was expected that the document composition rates would be dependent on the composition system used. 11 subjects, including one naïve word processor user, took part in the experiment.

The results showed that the speech-driven word processor was significantly slower for composition than the other systems. It also had the lowest composition efficiency of 21% (of words spoken in the final document) compared to 68% (of key presses in the final document) for the simple keyboard-driven text editor; an average of 43.9 edit operations were performed on each document on the speech-driven word processor, compared to 14.6 on the text editor. Most subjects used a "dictate-then-edit" strategy rather than an "edit-mistakes-as-you-go-along" strategy, and, by timing these separate modes, it was noted that these subjects spent considerably more time editing than dictating (approximately 17 minutes editing, on average, compared to 5 minutes dictating).

4. OBSERVATIONS AND GUIDELINES

The series of experiments reported here produced a great deal of data and experience of the practical problems of using speech-driven word processors. On the basis of this, some general observations can be made:

4.1 General Comments

- The human interface and dialogue characteristics are a vital part of any speech input system. Inadequate design of either of these will lead to a very inefficient system which is unlikely to be used in the long term.
- For listening typewriter and similar tasks, the recognition accuracy for speech-to-orthography must be very high (better than 95%).
- The operators of speech input word processors will need significant training in order to use speech efficiently in document creation tasks. The inclusion of natural language command structures does not reduce this training requirement, and may even increase it.
- Unconstrained natural language is too ambiguous and inefficient to be appropriate for tasks such as text editing.
- Speech-only listening typewriters are slow and inefficient, and are thus likely to only be acceptable in situations where hands-free input is absolutely essential.

EXPERIMENTS WITH A SPEECH-DRIVEN WORD PROCESSOR

4.2 Natural Language Input

- Making cursor movements and describing locations in the text ("pointing with the voice") is very difficult to do with speech alone. It is particularly difficult to perform character-level and formatting operations using the voice. The addition of a mouse or touch screen for cursor control to a speech-driven word processor can substantially ease this aspect of the editing task.
- When faced with a natural language understanding system, operators tend to try to develop a subset of commands, these commands often being similar to a computer command language. Operators tended not to use articles, conjunctions or prepositions, could be sloppy in their usage of tenses, and repeated words unnecessarily (eg. "in this .. this paragraph").
- Both computer naïve and experienced users found it difficult to invent appropriate spoken commands for editing operations. Few subjects even claimed to be using natural language structures.
- Using a speech-based practical natural language command system can be stressful, particularly in the early stages of learning, as the operator is often not certain whether a particular command will be interpreted correctly, incorrectly, or not at all. This has been likened to "Russian Roulette" where, after issuing a command, the operator waits with some trepidation to see what will actually happen.

4.3 Human-Computer Interface

- Visual feedback of spoken words tends to slow down the operator's speech, regardless of the way such feedback is initiated. Long term training may reduce this effect.
- When talking to the listening typewriter, subjects tended to concentrate on the area of the screen where their words were appearing, and did not notice changes in other parts of the screen. This caused a significant number of mode errors. Careful design of the screen layout must emphasise the mode in which the system is operating.
- Some subjects tended to make spoken asides (without using the microphone cut-off switch); these then appeared as text on the screen, often causing confusion, and had to be deleted by the subject. It is not easy to see how the effects of these could be eliminated without increasing the number of modes of the system, but it is possible that subjects would learn to avoid these with practice.

4.4 Subjects' Responses

- The response to speech input can be very polarised, with some subjects being very positive and some very negative. Some highly computer-literate visitors who have visited the project have been wary of using the system, either with or without a demonstration.
- Despite the listening typewriter's low performance in terms of speed and efficiency, many of the subjects enjoyed using the system in the experimental situation. However, the performance characteristics may well be a much more important factor when using such a system in real situations.
- In the Gould replication, those subjects who knew that they were using a simulation were more impressed by the simulated system than those who thought they were talking to a machine.

4.5 Other Considerations

- All of the experiments in this project were performed with the subject alone in an office. The attitudes of potential operators of the system in a real office environment, and those of others in the office, was not investigated, but might be inhibiting to the use of such a system.
- The "hands-free" system frustrated many subjects because they had nothing to do with their hands.
- The fully speech-driven word processor was significantly slower for the task of creative writing than either a simple keyboard-driven text editor or a dictation process (using a dictation machine to enter text, and then editing the typed-up dictation).
- As the ASR system is likely to make more errors during text entry than would a keyboard-driven editor, any speed advantage gained by dictating the text could thus be offset by the increased editing time to correct the larger number of errors.

5. CONCLUSION

The simulation of a full-speed listening typewriter has proved to be a very valuable research tool for the investigation of the human factors of speech and natural language input systems. The results from the formal experiments have been important as benchmarks of what could be expected of future speech-operated systems. The individual experiences of the subjects during these experiments and their attitudes towards the system have also provided valuable data concerning the acceptability of such systems, and the potential pitfalls in the design of speech-operated and natural language-based systems. In addition, giving visiting scientists the opportunity of actually using a listening typewriter in informal situations was an important contribution made by this project; all the visitors commented that this had been a very valuable experience.

A fifteen-minute video presentation describing the project and showing the speech-driven word processor in operation will be shown during the conference.

This work was carried out between 1986 and 1990, under the Alvey Directorate project number MMI/SP/079 (SERC numbers GR/D 3009.9 and GR/F 7059.4).

7. REFERENCES

[1] W A LEA (Ed), "TRENDS IN SPEECH RECOGNITION", Prentice-Hall, Englewood Cliffs, NJ, (1980).
[2] J D GOULD, J CONTI & T HOVANYECZ, "Composing Letters with a Simulated Listening Typewriter", COMMUNICATIONS OF THE ACM, 26, pp. 295-308 (1983).
[3] M J UNDERWOOD, "Intelligent User Interfaces", in G G SCARROTT (Ed), "THE FIFTH GENERATION COMPUTER PROJECT, STATE OF THE ART REPORT", Pergamon Infotech, 11.1, pp. 135-144 (1983).
[4] A F NEWELL, "Speech - the Natural Method of Man-Machine Communication?" PROCEEDINGS OF THE 1st IFIP CONFERENCE ON HUMAN COMPUTER INTERACTION, Imperial College, London, (North Holland) pp. 231-238 (1984).
[5] "MACHINE SHORTHAND AS A FULL-SPEED SPEECH RECOGNITION SIMULATION", ALVEY PROJECT FINAL REPORT, Dundee University Microcomputer Centre (1990).
[6] R DYE & G CRUICKSHANK, "A System for Composing and Editing Text Using Natural Spoken Language", PROCEEDINGS OF SPEECH '88, THE 7th FASE SYMPOSIUM, Edinburgh, (Institute of Acoustics) pp. 4:1321-1328 (1988).
[7] A F NEWELL, J L ARNOTT, K CARTER & G CRUICKSHANK, "Listening Typewriter Simulation Studies", INTERNATIONAL JOURNAL OF MAN-MACHINE STUDIES, 33, pp. 1-19 (1990).
[8] T B MARTIN, "Practical Applications of Voice Input to Machines", PROCEEDINGS OF THE IEEE, 64(4), pp 487-501 (1976).
[9] I H WITTEN, "PRINCIPLES OF COMPUTER SPEECH", Academic Press, New York (1982).
[10] K E P CARTER, S COOKSON, A F NEWELL, J L ARNOTT & R DYE, "The Effect of Feedback on Composition Rate Using a Simulated Listening Typewriter", PROCEEDINGS OF THE EUROPEAN CONFERENCE ON SPEECH TECHNOLOGY, Paris, (CEP Consultants Ltd.) pp. 1:402-404 (1989)