



PARCEL SORTING BY SPEECH RECOGNITION: A CASE STUDY IN VOCABULARY DESIGN

Clive R. Frankish* and Dylan M. Jones^o

ABSTRACT

Two types of vocabulary were compared in a simulated parcel sorting task using automatic speech recognition. One consisted of place names in current use, the other of equivalent alphanumeric codes. When an isolated word recogniser was used, the rate of code entry was higher for the place name vocabulary. Although the proportion of correctly identified codes was similar for the two vocabularies, the greater redundancy of alphanumeric codes meant that more recognition failures could be detected automatically. These results are discussed in terms of their implications for the development of a practical system for parcel sorting.

INTRODUCTION

Despite the slow growth of successful applications of voice input since the early 1970's, there are several areas where this technology has proved extremely effective. The success of these applications can be attributed to both technical and human factors considerations. Firstly, the requirements for speech recognition must lie within the limited capabilities offered by currently available technology. This generally means speaker-dependent recognition of relatively small vocabularies, and operation either in isolated-word mode, or in connected mode for short phrases. Although a new generation of recognisers now includes devices that have progressed beyond one or more of these limitations, these advances must often be weighed against increases in cost, and/or response time. Secondly, the prospects for voice input are most favourable if there are clearly identifiable advantages over alternative methods of data entry, such as pencil-and-paper, or keyboard. This is most likely to occur if the user's eyes and/or hands are required for some other task component, if there is a need for mobility around the work area, or when manual dexterity is impaired (eg. protective gloves are worn).

These specifications are frequently met in the materials handling, inspection, and quality control tasks that account for the majority of established applications of voice input. All involve data entry by trained staff, using a relatively small vocabulary, with a high rate of throughput. In addition, there is frequently a degree of error tolerance; either the primary task is itself error-prone (eg. inspection), or there are significant error rates associated with existing methods of data entry. Finally, data formats tend to be well-defined, and can be directly mapped onto syntax trees used in speech recognition.

The use of voice input in materials handling is already well established in the USA, where applications include sorting of airline baggage (ref

* University of Bristol, Department of Psychology, Bristol BS8 1HH.

^o UWIST, Department of Applied Psychology, Cardiff CF3 7UX.

1) and parcels (ref 2). In the UK, there is potential for applying this technology in Post Office parcel sorting, particularly in Parcel Concentration Offices (PCO's). Sorting at this stage involves assigning each parcel to one of up to 50 destinations. Parcel sorting machines currently use keyboard input, with a separate key for each place name. The destination for each parcel can be keyed only after the postman has manipulated the parcel so that the address can be read; if the workload is high, this operation may be carried out separately by a "facer". The need for keystrokes to be visually guided further contributes to the serial nature of the reading and keying components of the task.

The application is therefore one which meets the human factors criteria for voice input, and because of the limited number of destinations involved, the recognition vocabulary required is quite modest. However, the selection of place names used for keyboard entry is based purely on operational considerations; as a vocabulary for automatic speech recognition, it may be far from optimal. A given set of place names may contain pairs such as 'Lancaster' and 'Manchester', that for a recogniser are highly confusable (ref 3). If the task is to be converted to voice input, a minimal requirement is likely to be the elimination of such pairs from the vocabulary. An alternative approach would involve to change the entire task vocabulary, with a view to optimising recognition performance. One obvious candidate here is the ICAO alphabet, specifically designed to maximise acoustic distinctiveness. The purpose of the study described here was to compare a vocabulary consisting of place names used for keyboard entry with a vocabulary based on the ICAO alphabet.

The construction of codes for up to 50 possible destinations using the ICAO alphabet means that at least some two-part codes are required. The method chosen for achieving this was to take the first letter of each place name, adding a digit where this was required to distinguish between place names with the same initial letter. This system has the advantage that codes are easily derived from place names and can be quickly learned. Recognition performance and speed of data entry were therefore compared for place names (eg. Cambridge, Bristol), and for alphanumeric code equivalents (eg. Charlie one, Bravo two).

OUTLINE OF THE STUDY

Four representative test vocabularies were constructed by randomly selecting a sample of four PCO's, and listing the destinations used in sorting at these locations. For each set of place names, the corresponding alphanumeric codes were generated by taking the initial letter, and adding a digit where necessary. The average number of destinations for the four selected PCO's was 39.25, of which 15% consisted of two words (eg. "Belfast Delivery"). This resulted in a mean of 39.75 words in the place name vocabulary. The alphanumeric coding system required a greater proportion (80%) of two-word codes, but a smaller number of words in the recogniser vocabulary (22.5).

A Kurzweil Voicesystem speech recogniser (KVS version 1.0) was used. This is a speaker-dependent, isolated word device, with a maximum vocabulary of 1000 words, and a claimed accuracy of 95% or better. The equipment was located in a sound-attenuated room, with subjects seated in front of a keyboard and computer display. For all practice and test

sessions, destination codes (consisting of either one or two words) were read aloud as they were displayed on the screen. To reveal the next code in the test sequence, subjects pressed the '+' key, chosen for its size and prominence on the edge of the keyboard.

A total of sixteen subjects participated in the study, four being tested with each of the sample vocabularies. All were members of the general public, and had no prior experience with speech recognition devices. Each person spent four sessions with the equipment. These consisted of: initial enrollment; practice trials, to provide familiarity with the task; and two experimental sessions, one with each vocabulary. Immediate feedback was given during practice, with recognition failures signalled by the display "No good match". These codes were re-entered, until correct recognition was achieved. No feedback was provided during the two experimental sessions. Speech templates for the vocabulary under test were elicited at the start of these sessions. The complete set of destination codes was then presented 24 times, with the order of presentation randomised independently for each repetition.

RESULTS

Task performance was evaluated in terms of both speed and accuracy measures; the main features of these data are summarised in Table 1.

Table 1: Speed and accuracy measures of task performance

Performance measure	Place names	Alphanumeric
Task completion time (min.)	48.8	61.2
% codes correctly identified:		
Overall	81.8	77.7
1-part codes	85.1	92.2
2-part codes	62.8	75.1
% failures detected	88.2	97.8
% (correct + detectable fail)	98.2	99.6

The experimental task was completed more quickly when place names were used; the observed difference in completion times was statistically reliable ($t(15) = 5.79, p < .01$). This difference is only partly attributable to the greater number of words spoken during entry of two-part alphanumeric codes, since these words tended to be of shorter duration. A major factor in determining the rate of data entry was the use of an isolated word recogniser, which required a substantial pause between the elements of two-part codes.

Average recognition rates for individual items in the two vocabularies were virtually identical; 83% in both cases. One obvious point that must be noted is that identification of a one-part code requires only one correct recogniser response. With two-part codes, the expected level of performance is lowered, since both words must be recognised. Since more two-part codes are required by the alphanumeric vocabulary, we would expect fewer correct identifications of complete codes. This was the case, although the observed difference of approximately 4% was not statistically reliable ($t(15) = 0.90$).

Recognition performance for two-part codes was further depressed as a result of segmentation errors; cases where subjects neglected to pause between words. With an isolated word recogniser, this causes the two word code to be treated as a single utterance, with consequent recognition failure. This speculation was supported both by direct observation of subjects' behaviour and by internal checks on recognition data. When comparisons were made between codes of equal length, the alphanumeric vocabulary was clearly superior. Statistical analysis confirmed that identification rates were reliably higher for one- than for two-word codes ($p < .001$), and that when code length is equated, there was a significant advantage for alphanumeric codes ($p < .05$).

In a further analysis, recognition failures were subdivided into two categories: (i) misrecognitions, ie. instances where the output of the recogniser did not match the input, but did match some other legal code, and (ii) detectable failures, ie. instances where the recogniser output did not correspond to any member of the code set. For example, since each code set contained a limited subset of alphanumeric combinations, the output 'kilo nine' would be classified as a detected error if this code was not a member of that subset. Detectable failures would also include cases where the output consisted only of 'nine', or no acceptable match. In a practical sorting system, these are all cases in which parcels could be re-routed through the system for a second pass, or feedback given to indicate the need for code re-entry.

Although the overall failure rate was marginally higher for alphanumeric than for place name codes, the proportion of detectable failures was significantly higher for alphanumeric codes ($t(15) = 4.25, p < .01$). Assessment of overall system performance in terms of correctly identified codes plus detectable failures indicated a significant advantage for the alphanumeric coding system, for which the incidence of undetected failures was less than 0.4% ($t(15) = 3.54, p < .01$).

CONCLUSIONS

Factors that influence the performance of speech recognition devices include vocabulary size and the distinctiveness of items within the vocabulary. An alphanumeric coding system based on the ICAO alphabet therefore offered two potential advantages over place names; the vocabulary required for the application was virtually halved in size, and the alphabet was specifically designed for its acoustic distinctiveness. Despite this, performance was slightly poorer for the alphanumeric vocabulary, at least when measured in terms of the number of correctly identified codes. However, when code length was taken into account, recognition rates for alphanumeric codes were significantly better than for place names.

Two reasons for this result were identified, both related to the fact that the alphanumeric vocabulary required a greater proportion of two-part codes. The first is that identification of these codes requires two correct recognitions, rather than one; this offsets any advantage that might be gained from higher recognition rates for individual items in the alphanumeric vocabulary. The second reason is that failure to pause sufficiently between words resulted in segmentation errors, further depressing performance with two-part codes. While the first of these problems is unavoidable, the second can be overcome by using a

connected-word recogniser. Elimination of pause time would also reduce the difference in throughput rates for the two vocabularies.

If a connected-word recogniser were used, and correct identification of destination codes were taken as the sole performance criterion, it seems that there is little to choose between the two vocabularies. However, in operational terms there is a substantial advantage for the alphanumeric coding system, by virtue of its portability. Use of the place name vocabulary would require a separate evaluation for each PCO, with further checks if changes were made in the sets of destinations used. On these grounds alone, the use of alphanumeric coding appears to be an attractive long-term option.

Implications for feedback and error correction

The occurrence of recognition errors and failures generally requires the provision of feedback, indicating whether an utterance has been correctly identified. In data entry tasks, feedback must be explicit, with some form of display confirming the recogniser output for each utterance. This in turn detracts from the advantages of using speech, especially in applications with high throughput. If the feedback is visual, then performance of tasks in the 'eyes busy' category will suffer, as users are required to monitor the recogniser display. Auditory feedback, by means of synthesised or recorded speech, is relatively slow, and can be unpopular with users.

Faced with this dilemma, some systems designers have argued that feedback is unnecessary in applications where recognition accuracy is high, and a non-zero error rate is acceptable. Recommendations along these lines have been made for mail sorting, using a vocabulary of ten digits to enter zip codes (ref 4). However, this requires very good recognition performance if the number of sorting errors is to be kept within acceptable limits. A more promising approach is to use code redundancy for internal error-checking. The feasibility of this strategy was clearly demonstrated in the present study, where there was a much greater degree of redundancy in the alphanumeric codes than in place names. A simple check on code validity meant that the overall rate of undetected errors fell below 0.4% when the alphanumeric coding system was used.

If this level of performance were maintained under operational conditions, the use of redundant alphanumeric codes would reduce the need for explicit feedback. Instead, users could rely on a simple tone indicating recognition of a valid code, or detected recognition failure. On balance, it may well prove that the time saved by abandoning explicit feedback more than outweighs any marginal reduction in correct recognitions, caused by the increased use of two-part codes.

REFERENCES

1. Lind A J, Proceedings of Speech Tech '86, New York, 66-67 (1986).
2. Martin T B, Proceedings of the IEEE, 64(4), 487-501 (1976).
3. Visick D, Johnson P & Long J, Proceedings of the First IFIP Conference on Human-Computer Interaction, London, 99-103 (1984).
4. Craft A M, U.S. Postal Technology Research Technical Note PTR-04-81 (1981).