Proceedings of the Institute of Acoustics

EVALUATION OF SPEECH RECOGNITION IN TRANSPORT ENGINEERING

Peter Roach[1], Andrea Dew[1,2], Peter Bonsall[2] and Howard Kirby[2]

([1]) Department of Linguistics & Phonetics, University of Leeds
([2]) Institute for Transport Studies, University of Leeds

1. INTRODUCTION

An important requirement for progress in increasing the industrial use of Automatic Speech Recognition is scientific information about the performance of recognisers and speakers in a variety of real-life applications outside the laboratory. The SPRITE (Speech Research in Transport Engineering) project, conducted jointly in Leeds University by the Institute for Transport Studies and the Department of Linguistics & Phonetics, has tested a number of commercially-available speech recognisers in applications related to transport engineering, and devised standard tests that can be run on all such equipment. This paper presents results on success rates, estimated costings and user reactions. The project was funded by SERC (Grant No. GR/E 38184).

2. DESIGN OF A STANDARD TEST

2.1 As speech recognition technology has developed over the last few decades it has become increasingly important to devise reliable, repeatable and meaningful measures of system performance that will allow comparison of one recogniser with another. Ainsworth[1] points out that evaluation of speech recognisers is necessary to several groups, including

1. the manufacturer who is interested in developing the product
2. the end user who needs to develop an application
3. the prospective purchaser who needs to be aware of the market.

It is well known that a simple percentage score pertaining to recogniser accuracy is of little value by itself; the confusion implicit in such a metric is highlighted in papers by Russell, Deacon and Moore[2], Thomas and Winski[3] and Nusbaum and Pisoni[4]. Baker, Pallett and Bridle[5] present a set of recommendations and evaluation guidelines for test procedures for speech recognisers, and list the major variables to be controlled in testing; these include speaker population (age, sex, dialect, experience, mental state), acoustic environment, channel features (s/n ratio and channel noise), task environment, vocabulary (number of items, use of syntax, phonetic make-up of words), training and test procedure, equipment configuration and method of evaluation (counting of errors).

2.2 The SPRITE test devised for our research is described in detail in Dew and Roach[6] and the present paper presents only a brief overview. The test was devised to enable evaluation of different speech recognisers with a relatively small amount of data in the form of isolated words. Two critical factors of interest were the rate at which data could be input and the level of noise at which each device could function effectively. The main features of the test were that it was portable, flexible, recorded under controlled conditions and could be set up quickly on premises other than our own. It was designed so as to allow a longer or a shorter version to be performed according to circumstances. The results we obtained do not represent the optimum performance for each machine: we deliberately used untrained speakers, and no special tuning of recognisers to suit the conditions was carried out.

2.3 The recordings were made in a sound-treated studio using Sony digital recording equipment and Sony digital tape (16 bit samples); a Shure SM10 noise-cancelling microphone was used which was placed at 35mm off the centre of the lips and 25mm to the side of the mouth. The speakers recorded without interaction with the recognisers, since feedback can alter a speaker's performance quite profoundly (Dew et al[7]). The test consists of two main parts, the first being a comparatively short one consisting of phonetically confusable items (the "E-set" of consonant labels) and the second comprising several lists of phonetically disparate items such as the I.C.A.O. alphabet, the names of digits 0 - 10 and two different small vocabularies representative of typical applications in our field. The lists were recorded six times each in pseudo-random order at different rates, being presented to the speaker by VDU. Three different noise conditions were added, one being white noise produced by loudspeaker in the studio, another being white noise presented over headphones and the third being noise added to the tape after the recording session was finished.

2.4 The test was administered to five different speech recognisers and the individual performances are presented in Dew and Roach[[6]. In overall success rates we found surprisingly little difference between recognisers, but they differed in the degree of their susceptibility to noise and the rate at which data could be input. While this test does not give us any better ability to quote an *absolute* success rate, it does provide a good basis for deriving comparative figures.

3. RECOGNITION OF RECORDED DATA

3.1 Many data-gathering tasks that could be suitable as ASR applications are carried out in environments where conventional speech recognition systems, being bulky and requiring mains power, are not available, and we have therefore devoted considerable attention to the possibilities for off-line audio recording of data and subsequent input through recognisers. A description of an evaluation of this technique in comparison with other available methods has been published by Bonsall et al[8] and concluded that, although ASR showed great potential for assisting with data capture, the systems then available on the commercial market were not sufficiently accurate in these conditions to warrant their widespread use. Since that study we have carried out further testing of audio tape input in two field trials based on widely-used data gathering techniques in the study of car use in urban transport: one was the roadside recording of the registration plates of moving vehicles (the "*regno*" survey) and the peripatetic recording of parked vehicles' registration plates (the "*parking*" survey). Alphabetic letters in the registration plates were recorded using the I.C.A.O alphabet (Alpha, Bravo, Charlie etc.). A single recogniser was used for this test: this was a Marconi SR128, a machine which is no longer produced but whose performance, according to our benchmark tests described above, is still fully competitive with more recent products.

3.2 The data was recorded in a variety of conditions: the regno survey was carried out (a) overlooking an urban motorway (high flow rates and high ambient noise levels), (b) midway along a fairly busy urban road (fairly constant flow and ambient noise levels), (c) midway along the urban road with the observer seated in a car (slightly reduced visibility and lower ambient noise levels) and (d) adjacent to a busy urban junction (ambient noise and flow very variable). The parking survey was recorded (a) in a multistorey car park (variable acoustics), (b) on street (medium ambient noise) and (c) off street (low ambient noise levels). A staff of five (two men and three women) were specially trained for this work. One of the principal purposes of the study was to evaluate speech input in relation to other data-gathering techniques; in both surveys the data was recorded both by paper and pencil and by voice on to audio tape. Subsequent processing of the data was carried out

in three ways: the paper and pencil data was keyed in by an experienced typist; the audio tape recordings were transcribed by an experienced audio typist and the tape recordings were also used as input to the automatic speech recogniser. Full details of the results are given in Dew and Bonsall[9], but a summary is given below.

3.4 The regno survey suffered from the problem that the data rate (the number of vehicles passing) was not under the control of the person recording the data, while the parking survey allowed the enumerator to select data at her/his own pace. We measured both the completeness of the data (how many vehicles were missed out) and the accuracy (how many were wrongly transcribed). In the regno survey the completeness of the pencil and paper data declined from around 95% at 570 vehicles per hour to around 20% at 1100 vph; the decline in completeness on audio tape is less marked (99% at 570 vhp and 87% at 1100 vph). Overall transcription accuracy scores across all conditions were as follows:

Regno survey:	ASR 65%	Audiotyping 91%	Typing 63%
Parking survey:	ASR 69%	Audiotyping 95%	Typing 93%

An attempt was made to quantify the relative costs of the different methods, taking into account cost of equipment (discounted over different periods) and staff costs. Discounting equipment costs over 1000 hours of use, the figures were as follows:

Regno survey:	ASR £13.31	Audiotyping £7.91	Typing £6.31
Parking survey:	ASR £12.67	Audiotyping £6.47	Typing £5.81

3.5 It seemed clear from this study that in field conditions the ASR option did not offer satisfactory performance either on accuracy or on cost. It seems likely that many potential users would be prepared to take on the additional cost of ASR if the accuracy were comparable to the best alternative, and we therefore considered carefully what possibilities existed for improving the accuracy. It is clear that the major problem for the recording of data in difficult conditions for ASR is the lack of feedback to the user, and this can only be provided by on-line data input. In without-feedback working there is an overall effect of errors caused by faulty recognition, and in addition our experience suggests strongly that without the sustaining effect of the feedback, the speaker's performance degrades over time, often becoming significantly less successful after half an hour to an hour. Feedback is therefore required. In theory this could be made available by two-way radio link, but we feel that the recent commercial availability of hand-held portable speech recognisers has completely changed the scene and made such complex arrangements unnecessary. In the next section we describe an experiment similar to those described above but using a portable speech recogniser.

4. DATA INPUT WITH A PORTABLE SPEECH RECOGNISER

4.1 The recogniser chosen was the PTVC (Portable Transactions Voice Computer) by Voice Connexion, based on the Telxon PC-compatible hand-held data logger. Although this battery-powered device works as a stand-alone computer it is found more convenient to tailor the software for a specific application on a desktop machine with full-size keyboard and screen and more memory, then to download the software to the PTVC. This is an isolated-word, speaker-dependent recogniser with a vocabulary of up to 500 words. Feedback is via a speech synthesiser (with General American pronunciation). This device was used in an experiment to evaluate on-line ASR in field

trials. Other techniques looked at were paper and pencil, hand-held computer for keying-in of data, and video-tape (research is in progress in the Institute for Transport Studies on the automatic reading of registration plates by computer from TV pictures, and we have also looked at the possibility of a human operative reading numbers from the TV screen into the speech recogniser). On the parking survey, a team of three experienced enumerators worked together recording the same data by speech input to the PTVC, by paper and pencil and by keying-in on the hand-held computer, respectively.

4.2 Analysis of the results of this study has not yet been completed, but will be reported in a forthcoming paper (Bonsall and Dew[10]). It is, of course, to be expected that the accuracy rate with a feedback ASR system will approach 100% if the speaker is able to make repeated tries until the correct response is received: this is possible in a car-parking survey but not in a moving-car survey where the data rate is not under the enumerator's control. First results indicate accuracy figures in excess of 95% for speech input in the parking survey despite adverse weather conditions at the time of the data recording. Revised costing figures will show a reduction over those quoted above, possibly in the region of £11 per hour.

5. CONCLUSIONS

5.1 It is a truism to say that ASR has still to establish itself in real-world applications. One development that is needed if it is to gain widespread acceptance is an objective way of assessing the performance of a given system relative to its competitors, and we feel that the testing procedures that we have developed make some progress in this direction.

5.2 While the idea of recording data off-line on audio tape and subsequent processing by speech recogniser seems attractive from the point of view of convenience and (if data from a number of operatives is being processed) cost, we have established that in any conditions that fall short of ideal, the error rate rises to an unacceptable level. It is therefore essential for work such as we have been studying that feedback be provided to the operative. A new generation of fully portable speech recognisers makes this easily available for the first time, and preliminary results appear to show a dramatic improvement in accuracy using this technology, to the point where it may well be no less accurate than traditional ways of gathering data in the field given a moderate vocabulary size. The system we have worked with requires separation of words by pauses, but if this presents problems it is possible to acquire a portable connected-speech recogniser.

5.3 One factor that we feel is important is that all the staff who have worked on our test surveys have expressed strong preference for working with the ASR systems if the accuracy can be made acceptable. It was found more convenient and comfortable, and particularly suited for working in bad lighting conditions and in weather cold enough to make writing and typing difficult.

6. REFERENCES

[1] W.Ainsworth, *Speech Recognition by Machine*, (1988)
[2] M.Russell, J.C.Deacon and R.K.Moore, 'Some Implications of the Effect of Template Choice on the Performance on an Automatic Speech Recogniser', *Proceedings of the Institute of Acoustics* (1984)
[3] T.J.Thomas and R.Winski, 'Speech Recogniser Assessment in the Laboratory, not in the Field', *Speech Technology*, 3(4), 88-93 (1987).
[4] H.C.Nusbaum and D.B.Pisoni, 'Automatic Measurement of Speech Recognition Performance', *Computer Speech and Language*, 2, 87-108 (1987)
[5] J.Baker, D.Pallett and J.Bridle, 'Speech Recognition Performance Assessments and Available Databases', *ICASSP Proceedings*, pp. 527-530 (1983)
[6] A.M.Dew and P.J.Roach, 'Assessing Speech Recognition Equipment Using the SPRITE Test', *Technical Note 255*, Institute for Transport Studies, University of Leeds (1990)
[7] A.M.Dew, B.A.Hardwick, P.J.Roach, M.A.Shirt and H.R.Kirby, 'Voice Degradation Problems in Using Automatic Speech Recognition', *Proceedings of the International Conference on Speech Input/Output*, Institution of Electrical Engineers (1986)
[8] P.W.Bonsall, F.Ghahri-Saremi, M.R.Tight and N.W.Marler, 'The Performance of Hand-held Data Capture Devices in Traffic and Transport Surveys', *Traffic Engineering and Control*, 29(1), pp.10-19 (1988)
[9] A.M.Dew and P.W.Bonsall, 'Evaluation of Automatic Transcription of Audio Tape for Registration Plate Surveys', in E.S.Ampt, A.Meyburg and A.Richardson *Selected Readings in Transport Survey Methodology* (1990)
[10] P.W.Bonsall and A.M.Dew, forthcoming Technical Note of Institute for Transport Studies.