
Published paper

Hardwick, B.A., Bonsall, P.W., Kirby, H.R.(1986) *Applied Evaluation of Speech Recognisers with Respect to Tape Recorded Data.* Institute of Transport Studies, University of Leeds, Working Paper 213

Working Paper 213

January 1986

APPLIED EVALUATION OF SPEECH RECOGNISERS WITH RESPECT TO TAPE RECORDED DATA

Bryan A. Hardwick
Peter W. Bonsall
Howard R. Kirby

APPLIED EVALUATION OF SPEECH RECOGNISERS WITH RESPECT TO TAPE RECORDED DATA

ABSTRACT

HARDWICK, B.A., BONSALL, P.W. and KIRBY, H.R. (January, 1986), Working Paper 213, Institute for Transport Studies University of Leeds.

In several transport planning and maintenance survey applications data can be efficiently captured by audio tape-recording. Such recording is useful in situations where the rapidity of events, or the need for the operator to keep his eyes on the objects being recorded, make it difficult to use other recording methods such as paper and pencil, or keyboards. Examples of the sort of situations considered in this report are car registration plate surveys (moving and parked vehicles), and street furniture inventory surveys.

The transcription of such tape recorded data has, up until now, been a time consuming and labour intensive process, and it was therefore decided to investigate the possibility of using a speech recogniser to automatically transcribe the data. Whilst laboratory studies of the efficacy of the speech recognition process have of course been carried out, there have been very few field trials to evaluate the technology in specific application areas. This paper describes such an evaluation, for the case in which the speech recogniser is used for transcribing data recorded on an audio cassette. Different types of recogniser were tested on data recorded under both laboratory and field conditions. The suitability of associated equipment, (microphones and tape recorders), is also assessed for use with speech recognisers, and the recommended models are described. A training programme is suggested for new users and the results of the equipment trials and training programme are presented. Comparisons are also made with other more traditional methods of data collection and transcription.

1. BACKGROUND

In many types of survey carried out by transport planners and operators and highway authorities, the need for the observer to keep his eyes on the events being recorded makes recording techniques such as pencil-and-paper, banks of tally counters, or hand-held micro-computers difficult to use. In these cases data can often be efficiently captured by recording on audio cassette. The problem with audio tape recording, however, has been the high labour cost associated with the process of transcribing the data from the audio tape into the computer for analysis. Transcription times as high as twice the recording time have been experienced - (Montgomery 1984).

These types of survey would appear to be ideal for the use of speech recognition equipment to automatically transcribe the tape-recorded data, but while many laboratory investigations into the efficacy of speech recognisers have been made, very few field trials for specific applications have been carried out.

The possibility of using speech recognisers to automatically transcribe tape recorded traffic and inventory survey data onto a computer has therefore been investigated, initially under a grant from the UK Science and Engineering Research Council (SERC). This has enabled us to assess the usefulness of two different types of speech recogniser for a variety of transport and traffic data capture situations, and to make some recommendations as regards ancillary equipment and the training necessary for succesful use of the equipment.

This initial work showed that the recogniser's transcription accuracy seemed to be affected by changes in the voice, arising perhaps due to stress or fatigue. Similar findings having been obtained at the US Army's Construction Engineering Research Laboratory at the University of Urbana Champaign, Illinois, further work has been commissioned by the European Office of the United States Army Research, Development and Standardization Group. This work is being carried out jointly by the Institute for Transport Studies and the Department of Linguistics and Phnetics at the University of Leeds. It has the aim of identifying the factors affecting the degradation of the voice and of suggesting ways of compensating for this when using speech recognisers. The speech-related aspects of this investigation are not covered in the present paper, though results of further applications of speech recognisers to traffic surveys are.

2. APPLICATIONS OF SPEECH RECOGNITION

2.1 Registration plate surveys

Surveys of vehicle registration plates can provide the transport planner with several very important pieces of information that are difficult to obtain in other ways. Examples of these are journey times, routes taken, and length of stay in car-parks. Traditional methods of collecting this data are "pencil and paper", where the rate of traffic flow allows it, or by tape recording the spoken data;

transcription into computer-readable form is then done manually thereafter, preferably using an audio typist for the tape recorded data.

The main problem with pencil and paper recording is that the observer is continually looking down to his record when writing. This produces problems of eye-strain due to the rapidly changing focus as the observer looks from pad up to moving vehicle. There is also the risk that vehicles could be missed completely. Also, in wet weather it may be difficult to read any of the data at the end of the survey!

Tape recorded data is far less prone to the problems described above, but the transcription process can be becomе tedious, and hence error prone, for the typist over long periods. In the SERC project, particular attention was paid to the effectiveness of using speech recognisers to transcribe registration plate data automatically.

2.2 Road inventory surveys

Road inventory surveys typically consist of two parts, (i) the recording of fixed items in the street, such as drains, manholes, junctions, crossings etc., and (ii) the noting down of any areas of road (or footpaths) in need of repair. The surveys are traditionally carried out by two people, one on each side of the road, making a single pass down the road. Data is recorded on a bank of tally counters and after the survey both sets of results are combined and entered on a standard inventory survey sheet along with details of site location etc. (Note that the 'standard' inventory survey sheets can vary considerably even within a single local authority). Back in the office the data is typed into the computer from the inventory sheets.

The main drawback of this system is that heavy demands are placed upon the operators hands. He not only has to cope with a bank of tally counters for data recording, but must also push a distance measuring wheel, and be ready to write down anything unusual that needs recording. The operator also runs the risk of missing data items due to having to look down to his clip-board while operating the counters.

With such draw-backs to the traditional method, there is clearly a good case for the automation of some, or all, of the data gathering process.

A first move towards this automation has been made by West Yorkshire Metropolitan County Council in the use of a number of "Husky Hunter" data loggers. In principle, the operator now enters the data on the key-board while walking down the road, the data being stored in the machine's memory. When the survey is complete, the data logger is linked to the mainframe computer and the memory contents are transferred to the mainframe. This transfer is done by programs resident either on the data logger or on the mainframe, and will be error free and very fast.

Initial trials of the data logger were not very encouraging for two main reasons. (i) The process of typing the data is taken away

from the skilled typist in the office and given to the unskilled surveyor in the street, leading to errors in typing and delays due to unfamiliarity with the keyboard; (ii) the operator still has to look down at the keyboard to enter data, and still has a lot for his hands to cope with during the survey. In practice it was found that several of the operators carried out the survey using the traditional inventory sheets and then transferred the data to the data logger back at the van, thus defeating the object of the exercise.

Road inventory surveys would appear to be an ideal application for speech recognition. The data rate is relatively slow, and the operators hands are left free for other tasks, e.g. pushing the distance measurement wheel. The operator would wear a head-set microphone connected to a portable cassette recorder on a belt strap.

2.3 On-train surveys

Since train tickets are not (usually) specific to particular times of trains, market research on the use of different types of ticket by different types of traveller has to be conducted on board the trains themselves. The use of paper-and-pencil techniques and even the use of hand-held data loggers suffers from the same kind of limitations as for road inventory surveys; indeed, the swaying of the carriage and the movement of people makes accurate data capture even more difficult. Trials conducted by Institute staff and British Rail staff showed that tape recorded data capture was very much simpler, and was less effort, than conventional methods.

3. WORK PROGRAM

The work in the SERC project fell into several clearly defined sections, namely:

(i) selection and testing of the speech recognition and other equipment

(ii) training subjects to use all the equipment

(iii) comparative data collection trials.

The laboratory tests and all the training of the users were carried out using test data generated on a micro-computer. This data took the form of random registration numbers appearing at controllable rates, this being a far simpler simulation than road inventory survey data. Observers then read out these numbers either directly into the recogniser for immediate transcription, or into a tape recorder for subsequent transcription. Since the numbers displayed were known, the accuracy of both the observer and the transcription process could be determined, and the effect of varying data rates and equipment investigated.

Having used the laboratory tests to determine the best combination of equipment and method of use, some field trials in collecting registration plate data were carried out. These were to

determine the differences, if any, between using the equipment in the laboratory and using it in the field. These trials consisted mainly of road-side surveys of moving vehicles and car-park surveys of stationary vehicles.

Once the operation of the equipment had been sorted out, two comparative studies were set up. These involved training small groups of people to use all the equipment, and then carrying out surveys in as realistic a manner possible.

The first was a survey of registration plates of moving vehicles using different methods of data collection and transcription. These were:-

Data collection : audio tape recording
pencil and paper
hand held micro-computer.

Transcription : speech recogniser
audio typist
normal typist.

The second set of surveys were road inventory surveys involving:

Data collection : audio tape recording
pencil and paper

Transcription : speech recogniser
typist.

4. SPEECH RECOGNISERS AND ANCILLARY EQUIPMENT

There are three main types of speech recogniser:

(i) Isolated word recognisers

(ii) Connected word recognisers

(iii) Continuous word recognisers.

All recognisers developed so far require that each user trains the machine to recognise a set vocabulary of words, the machine having to be re-trained for each different user. This training provides the "templates", against which incoming data is subsequently matched.

Isolated word recognisers have a serious limitation in that each word must be spoken in isolation from the rest, usually with a gap of about 200ms between each word. This severely limits the speed at which the data can be spoken and also requires the data to be spoken in an unnatural, stacatto manner. Preliminary trials with an isolated word recogniser showed that this limitation would be unacceptable for the data rates envisaged in registration number surveys.

Connected word recognisers can recognise short connected phrases

of words by storing the incoming data until a long enough break in speaking is encountered, and then processing the stored phrase. They do not require each word to be separated from the next, but there is a limit to the length of phrase that can be stored for processing. There is also a limit to the amount of data that can be stored while processing is being carried out on the previous phrase.

Continuous word recognisers process the data as soon as it is entered, and should be able to cope with phrases of any length.

4.1 Recognisers evaluated

At the outset of the SERC project, (Kirby 1983), only one continuous word recogniser was available, the LOGOS developed by Logica. It was, however, too expensive to acquire for this project, and with plans for a cheaper version being in abeyance, the project has had to limit its evaluation to isolated and connected word recognition technology. Those selected for evaluation were the Interstate Electronics Corporation SYS300 Voice Recognition System, and the Marconi SR-128X Speech Recogniser.

The IEC SYS300 is an isolated word recogniser (costing about £1500) with a maximum vocabulary size of 200 words. The minimum between-word pause is 160ms and the response time (following end-of-word detection) is (50 + n)ms, where n is the active vocabulary size.

The Marconi SR-128X is a connected word recogniser (costing about £10500) with a maximum vocabulary size of 200 words. The end-of-phrase silence is set at 600ms and the response time is approximately 650ms, which includes the end-of-phrase silence. The number of words active at any time can be restricted by using a syntax facility. This will usually enhance the recognition accuracy, but must be used with great care as mis-recognition of a word can cause the input data and the syntax to get out of step. A score value is also available for each word recognised and output. This score gives the degree of match between the data and the stored template, and could be used to check the veracity of any word recognised. In practice it was found to be of only very limited use as wrongly recognised words sometimes had higher scores than correctly recognised words, this being due to the peculiarities of the recognition software.

4.2 Programming, training, and use

Before either of the speech recognisers can be used they must be programmed with the required vocabulary and syntax. In both cases programming is initially carried out using a VDU connected to the appropriate port. Once the programming is complete, the recognisers must be trained to recognise the operator's voice. Each machine has a training mode which prompts the operator to say each word in the vocabulary into the microphone and stores the template so produced. The Marconi SR-128X makes one pass through the complete vocabulary, but the SYS300 makes three passes and stores the average template of each word so produced.

Once both the programming and training are complete, the data

from either machine can be stored, and recalled at a later time without the need for further programming or training. The SYS300 must be connected to a main-frame computer for this process to be carried out as it involves a complete memory-image dump in binary form on to the main-frame. The Marconi has a built-in mini-cassette recorder on to which both programs and templates can be dumped. When re-loading this data the Marconi can be used as a stand alone machine with no further need for a VDU, whereas the SYS300 always needs the VDU connected.

For either machine, having programmed and trained it, speech recognition is achieved by setting it into its recognition mode, and then either speaking into the microphone or playing back a previously recorded tape (at a suitable level). On the Marconi the output is visible on the front panel display, and on the SYS300 the output appears on the VDU. Both machines have a further serial interface which allows other equipment to read the output.

4.3 Selection of ancillary equipment

The training manuals for both of these machines recommended the use of Shure SM10A microphones. These microphones cost about £95 and come complete with a headset, thus leaving the hands completely free. The pick-up pattern of this microphone is uni-directional and extends to only about 1" from the front of the microphone and therefore it has excellent noise reducing properties.

A cheaper microphone was also tried in order to compare its performance with the Shure. This was a hand-held Altai EM-506 and it cost about £20. In use the quality was close to that of the Shure, but the user had to be extremely careful not to knock it or brush against it, and it was difficult to keep it at an optimum distance from the mouth when moving around to get a good look at some of the registration plates.

To complete the comparison, the built-in microphones on the portable cassette recorders were tested. These microphones usually have a wide pick-up pattern and are also very prone to picking up noise from the handling of the recorder, the result being a very noisy recording. The use of these built-in microphones was therefore completely ruled out.

The tape recorder recommended by Marconi was the Sony Walkman Professional WM-D6, costing around £200. This machine has Dolby noise reduction facilities and a manual recording level control. Two small cheap portable cassette recorders were tested as well, these were made by Hitachi and Panasonic and cost about £25 each. These machines have automatic recording level control and no noise reduction facilities.

Problems were experienced in obtaining repeatable results when playing back the tapes on the cheaper machines, with nominally identical runs through the tapes giving variations of up to 50% in the number of data items correctly recognised. This problem was most severe when the tapes were recorded in noisy field conditions. The problem was reduced to a 10% variation in the number of correctly

identified items by playing back the tapes on a high quality cassette deck. The probable cause of these problems is the mis-alignment of the recording/play-back heads and the variability of tape speed commonly found on small portable cassette recorders. Although mis-alignment of the heads would be a reason for playing back a tape on the same machine that it was recorded on, problems due to variability of tape speed would be exacerbated by playing back on the same machine. The improvement in recognition rates noted when the tapes were played back on the high quality cassette deck would indicate that tape speed variability was the major problem, and therefore all the tapes recorded on the cheap portables were played back on the Hitachi cassette deck (costing around £80).

Playback repeatability tests were carried out for both types of recorder, with data recorded under both laboratory and field conditions, and the results are summarized in Tables 4.1 and 4.2.

These results show that the Sony Walkman Professional is the better machine to use for collecting data for entry into the speech recogniser. Because of its high cost, only one Sony Walkman was available to the project, and the tests described in the following sections were carried out with the cheaper recorders to provide a fair comparison between trainees. The tapes for the following tests were run through the recogniser several times to eliminate the worst of the variability, and the results must also be taken with the knowledge that an increase of up to 20% in accuracy would have been possible if the Sony recorder had been used.

Table 4.1 Comparison of playback repeatability of portable cassette recorders and mains cassette decks for data recorded under laboratory conditions.

			RECORDED ON		
			HITACHI CASSETTE DECK D-E10	HITACHI PORTABLE RECORDER TRQ 340	SONY WALKMAN PROFESSIONAL
PLAYED BACK ON		HITACHI DECK	96% - 100%	82% - 92%	87% - 89%
		HITACHI PORTABLE	N/A	45% - 92%	N/A
		SONY WALKMAN	N/A	N/A	85% - 89%

N/A = not applicable

Table 4.2 Comparison of playback repeatability of portable cassette recorders for data recorded under field conditions.

			RECORDED ON	
			HITACHI PORTABLE RECORDER TRQ 340	SONY WALKMAN PROFESSIONAL
PLAYED BACK ON		HITACHI DECK	58% - 70%	83% - 85%
		HITACHI PORTABLE	48% - 75%	N/A
		SONY WALKMAN	N/A	88% - 90%

5. PRACTICAL CONSIDERATIONS

5.1 Vocabulary size and content

5.1.1 Registration plate surveys

The details actually recorded are usually the digits, year letter, and a vehicle type. Assuming that there are five commonly used vehicle types, (CAR, BUS, VAN, LORRY, MOTOR-CYCLE), and leaving out the letters that have not been used as year letters, (I,O,U,Z), we are left with a vocabulary size of 37 words. Montgomery (1984) found it necessary to use the ICAO phonetic alphabet, (ALPHA,BRAVO,..,ZULU), even for audio typists, since many letter names are very similar in sound, e.g. B,C,D,E,G,P,T all sound similar and appear to be more or less inter-changeable on the speech recogniser. The ICAO phonetic alphabet is probably the best known, and this was used for all our trials.

A few problems became obvious while using the ICAO phonetics. These were the similarities in sound between the following words:

(i) 5,9, and MIKE

(ii) 8 and ECHO

(iii) VICTOR and 6+some other word.

Most of these problems can be overcame by clear ennunciation, but this is quite difficult to do in high traffic flow conditions. On the whole it is thought that introducing our own modifications to the ICAO system would produce even more problems, since most people already know some of the phonetics from popular usage. If any changes were to be made then the whole phonetic alphabet should be reviewed, and this would be a complete project in its own right, requiring co-operation by many interested authorities.

5.1.2 Road inventory surveys

As stated earlier the data actually recorded in a road inventory survey depends very much on the local authority in charge where the survey is carried out. The following list is a typical selection of the words and phrases required, but it must not be considered an exhaustive list. Digits 0 - 9, LEFT, RIGHT, ROAD, INNER-VERGE, OUTER-VERGE, FOOTPATH, KERB, LENGTH, WIDTH, MANHOLE, CROSSING, SURFACE-CODE, JUNCTION, GULLY, TIED, NOT TIED, LONG-JOINT, TRANS-JOINT. Most of the items are self-explanatory, but the last four maybe need some clarification. "TIED" and "NOT-TIED" refer to whether kerb and step heights are determined by road surface level or not, and "LONG-JOINT" and "TRANS-JOINT" refer to the joints in a concrete surfaced road.

5.2 Data entry rates

This section applies only to data about moving vehicles, since in all other surveys considered the data rate is determined by the operator at a rate suitable for them.

There are three main factors which affect the rate at which data can be entered into the machine. These are :

(i) vocabulary size

(ii) between-word pause length

(iii) size of the input buffer .

The second factor applies only to the isolated word recogniser, as a connected word recogniser does not require pauses between words, only between phrases.

The main restriction on data entry rates for the Marconi was the vocabulary size. Tests were carried out using various vocabulary sizes and these are summarised in table 5.1. As expected the smaller the vocabulary size, the faster the data could be entered and still obtain an acceptable accuracy. Accuracy was measured by the correctness of complete 5-character sequences, rather than the correctness of individual characters, this being the important factor in a registration plate survey.

The three parts of the 5-character registration plate are the digits, the year letter and the vehicle type. The numbers of words in each section is 10 digits, 22 letters (I,O,U,Z not used), and 5 vehicle types.

Table 5.1. Accuracy of transcription as a function of vocabulary size and data rate.

DATA RATE	VOCABULARY SIZE: 10 (DIGITS ONLY)	15 (DIGITS + VEHICLES)	22 (LETTERS ONLY)	32 (LETTERS+ DIGITS)	37 (ALL THREE)
500	100%	100%	100%	100%	100%
750	100%	100%	100%	98%	96%
1000	100%	100%	90%	80%	75%
1250	100%	100%	80%	70%	
1500	100%	100%	75%		
2000	100%	100%	75%		

The data rates are for five character sequences per hour, (i.e. equivalent to vehicle registration plates per hour), and the percentage accuracies are the average values of several runs. The blank boxes indicate that the results obtained were low values of accuracy and not repeatable.

Table 5.1 shows that up to about 750 vehicles per hour, the accuracy of transcription is good enough to allow the digits, year letter, and vehicle type to be specified (i.e. 37 word vocabulary). Above this data rate the accuracy of transcription decreases very rapidly. If the year letter could be omitted then data rates of 2000 vehicles per hour can be attained with high accuracy. At data rates above 2000 registration plates per hour the observer has great difficulty in actually saying the data and results could not be obtained for these higher data rates.

The other factor that limits data entry rates is the size of the input buffer on the speech recogniser. If the machine is occupied with processing some data that has been entered, and some more data is entered, it is digitised and stored in a buffer to await processing. The Marconi SR-128X has an input buffer that will store up to about 8 seconds of speech. Although this might sound an adequate amount the effect of having a large vocabulary is to slow down the processing and cause the incoming data to pile up. Once the buffer is full no more data can be accepted and attempts to enter more data cause the machine go into an error condition where all stored data is lost, and if the syntax facility is being used then this will also probably be left at the wrong point causing further errors. There is unfortunately no warning that this is about about to happen. A future version of the machine should include the ability to output some form of flag to indicate that the buffer is becoming full, and this could be used to control the tape recorder to prevent buffer overflows.

Preliminary trials made with the IEC SYS300 isolated word

recogniser showed it to be incapable of transcribing phrases of "isolated" words, such as would be encountered when collecting data from moving vehicles. Data collected from rows of parked vehicles was then transcribed by the SYS300 and the Marconi SR128X, and the following results obtained.

Table 5.2 Comparison between SYS300 and SR128X using parked vehicle data

	DATA COLLECTION TIME	TRANSCRIPTION ACCURACY
	min sec	
IEC SYS 300 (1)	6 : 33	53 %
(2)	16 : 40	95 %
MARCONI SR 128X	4 : 28	98 %

(1) Fastest time possible consistent with obtaining over 50% transcription accuracy.

(2) Collection time consistent with highest possible accuracy.

As expected, the length of the pause necessary between words proved to be the downfall of the SYS300. In order not to lose information by speaking at a faster rate than the machine could cope with, it was found necessary to speak in a most unnatural manner, with long forced breaks between words. This not only broke up the natural rhythm of saying a registration number (making it harder to actually say the number correctly), but also reduced the effective data rate to about 200 registration plates per hour, an unacceptably low figure. With this restriction in mind it was decided to concentrate on using the Marconi recogniser for the future trials and leave the SYS300 for single word command applications.

6. TRAINING PROGRAMME

This series of experiments was designed (i) to test several different aspects of training people to use speech recognition equipment and (ii) to test the speech recogniser against other methods of data collection.

6.1 Training the users

A training programme was drawn up which involved the user in two distinct types of training. These were : (i) use of the machine in a quiet environment; (ii) use of the machine in an environment where data would actually be collected, e.g. at the side of a road. It was necessary to record separate templates for both these situations as templates recorded in a quiet environment bore little resemblance to

the words spoken whilst actually collecting data on site. The noise of the traffic, for example, caused the user to compensate by raising the level of their voice both in volume and pitch. In order to record templates for a noisy environment, traffic noise was played to the user through headphones while they were recording the templates. Templates so recorded were a far better match for the data spoken in a noisy environment.

Much of the training took the form of sitting with the speech recogniser and watching the results as data was spoken, i.e. getting immediate feed-back as to the accuracy of transcription. This immediately high-lights one of the main problems with using a tape recorder for data collection in this manner, which is that there is no feed-back as to how well the user is doing while he is recording the data. Without this vital feed-back the user is unable to see when anything is going wrong and hence try to correct it.

Great emphasis had to be placed on the fact that the user is ultimately speaking to a machine and therefore clarity of speech is of overriding importance. This is particularly important in order to avoid problems caused by co-articulation, where the user runs one word into the next.

The training started with getting each of the trainees to learn the ICAO phonetic alphabet. This usually took only about 15-20 minutes. The next stage was to get the trainees to record a list of registration numbers before telling them anything about the speech recognition equipment. The same list would be recorded after each training session to show how much they had improved, using the first recording as a reference.

After explaining how the equipment worked each trainee had a session of "hands-on" training, which gave them some idea as to how sensitive it was to variations in the voice etc. Templates for both laboratory and field use were then recorded.

Next there was a period of actual data collection, recording the registrations of both stationary and moving vehicles. These recordings were then run through the recogniser in order to give each trainee some idea of how they had managed. These sesssions were particularly valuable in helping to identify individual problems such as co-articulation.

The indoor training lasted for about 2 hours, and the data collection lasted for about 1 hour; the rest of the 3.5 hour training session was taken up with recording the set list of registrations. This program was then repeated on the second day and any other problems that might have arisen were sorted out.

The outdoor data collection was conducted as follows.

Parked vehicle surveys were conducted along a row of parked cars. The trainee was able to take as much time as they required (within reason) for gathering the data.

Moving traffic surveys were conducted by the side of a main road with traffic flow rates of approximately 600 vehicles per hour. It was important to choose a place where the trainee could see both the front and rear registration plates as the vehicle went past. It was also important to make sure that there were no objects behind the observer that could reflect sound into the microphone.

The transcriptions of these two types of survey were checked against a paper and pencil record taken at the same time (the traffic flow rates being sufficiently low to allow this), and the accuracy of the observer's record on the tape was also checked.

6.2 Results of the training programme

The following tables show the results obtained for each of the trainees after each of the training sessions.

The test data was a list of 50 random registration plates and vehicle types generated by a BBC micro-computer. Table 6.1 gives a direct comparison of transcription accuracy obtained by recording this list after successive training sessions. Strictly speaking the comparisons are relevant only for indoor conditions.

Table 6.1 Transcription accuracy of the test data before and after training sessions.

	BEFORE TRAINING	AFTER 1 DAYS TRAINING	AFTER 2 DAYS TRAINING	AFTER 3 DAYS TRAINING
TRAINEE A	33%	63%	63%	59%
TRAINEE B	82%	93%	98%	98%
TRAINEE C	63%	71%	70%	66%

As expected the greatest increase in accuracy came after the first training session. The probable explanation for the fall off in accuracy for trainees A and C after the third training session was over-confidence, and given another period of training the accuracy would probably have started to rise again.

Table 6.2 gives the transcription accuracy for the parked vehicle surveys. The second set of data for trainees B and C was recorded at a row of parked cars alongside a road with a steady flow of traffic along it. This increased the background noise on the tape to the point where it started to interfere with the speech recogniser, (possibly because the observer is sideways on to the traffic instead of facing it as in a survey where the observer is stationary), hence the fall in accuracy between the two surveys.

Table 6.2 Transcription accuracy for the stationary vehicle surveys

	AFTER 1 DAYS TRAINING	AFTER 2 DAYS TRAINING
TRAINEE A	47%	
TRAINEE B	90%	75%
TRAINEE C	44%	41%

(Trainee A carried out only one stationary vehicle survey).

Table 6.3 gives transcription accuracies for three traffic surveys carried out on a main road in Leeds. The severity of the traffic noise varied considerably and this would account for some of the variations in the transcription accuracies. Trainee B obtained better accuracy after an initial period of settling down. He admitted that for the first 3 or 4 minutes he was still settling down and making mistakes, but after this he was relaxed and speaking in a more natural manner. The other two seemed to vary between trying hard to speak slowly when the traffic flow was light and running the words together when the traffic was heavier.

Trainee B had also been a radio operator in the war, and was quite used to the clear "clipped" type of speech necessary for this type of work.

Table 6.3 Speech recogniser transcription accuracy for the surveys of moving vehicles.

	AFTER 1 DAYS TRAINING	AFTER 2 DAYS TRAINING	AFTER 3 DAYS TRAINING
TRAINEE A	63% (500)	70% (700)	59% (780)
TRAINEE B	88% (480)	77% (850)	74% (510)
TRAINEE C	44% (520)	49% (930)	31% (510)

Approximate traffic flow rate given in brackets in vehicles/hour.

7. COMPARATIVE DATA COLLECTION EXERCISES

7.1 Data collection and transcription methods in registration plate surveys

Three different methods of data collection and three different methods of data transcription were then compared. The data was collected by two observers using tape recorders, one using pencil and

paper, and one typing the numbers into a hand-held micro-computer, (a Tandy TRS-80 Model 100 portable computer). The surveys on parked vehicles were carried out in car-parks on the campus of the University. The surveys for moving vehicles were carried out on one lane of a dual-carriageway in Leeds, with a traffic flow rate of approximately 510 vehicles per hour, and lasted for 18.5 minutes, (extreme cold prevented a longer survey being carried out). The observers were spaced out so that none could be influenced by hearing any of the others.

The transcription of the tape recorded data was carried out in two ways. The first was by the speech recogniser and the second was by an audio typist listening to the tape and typing directly into the computer. The paper and pencil record was typed directly into the computer by a standard typist. The data in the hand held microcomputer was down-loaded directly into the computer using standard programs.

Comparisons were then made between accuracy of recording the data, time taken to transcribe the data onto the computer, and the accuracy of the transcription.

7.2 Data collection and transcription in road inventory surveys

The road surveyed in the first trial was 330 metres long and was quiet enough in terms of traffic to be completely surveyed by one operator in one pass for both methods of data collection. The operator was unfamiliar with the standard inventory sheet method, but was experienced in using the speech recognition equipment.

In the second survey the operators were surveyors from WYCC, used to using the inventory sheets but unfamiliar with speech recognisers. The road used in the second survey was 200 metres long and was surveyed in one pass by both operators together using inventory sheets and then in one pass by one operator using the speech recognition equipment. No formal training in the use of the recogniser was given to this operator, he was simply told to speak clearly and naturally. The template for his voice was also recorded at the same time.

When using the speech recognition equipment all data items must be specified completely if the data is to be collected in only one pass down the road, e.g. "left inner-verge manhole", "road length one zero zero metres", "right footpath crossing", etc. If the survey is done in two passes, i.e. down the road on the left-hand side, and back along the right-hand side, then the designation left or right could be omitted fron the description as long as the change in direction is made clear. The exact form of the data items will depend very much on the software used to collect the data output by the speech recogniser, the examples given above being typical descriptions.

7.3 Results of trials

The results of the comparative data collection exercise for both registration plate and road inventory surveys are shown in the following tables.

Table 7.1 Comparison of data collection and transcription methods for parked vehicle surveys, (after 2 days training).

	TAPE 1 SPEECH RECOGNISER	TAPE 2 SPEECH RECOGNISER	TAPE 1 AUDIO TYPIST	TAPE 2 AUDIO TYPIST	PENCIL AND PAPER TYPIST	HAND-HELD COMPUTER
SURVEY TIME	min sec 5 : 07	4:50	5:07	4:50	4:36	4:24
TRANSCRIPT. TIME TAKEN	5:07	4:50	5:07	4:50	1:52	0:20
TRANSCRIPT. ACCURACY	75.0%	41.8%	100%	100%	100%	100%

N.B. Tape 1 was recorded by trainee B, tape 2 was recorded by trainee C, and all trainees recorded parked vehicle numbers with 100% accuracy.

Table 7.1 shows that, as expected, when making a record of parked vehicle registration plates the extra time available allows better use to be made of direct techniques of recording such as the hand-held computer. This method is particularly good as a transcription accuracy of 100% is guaranteed, and the transcription time is also very low.

Table 7.2 shows the statistics of the comparative data collection exercise for a road-side survey of moving traffic. An "agreed list" of 157 registration plates that actually passed the survey point was drawn up from a careful comparison of the four records obtained. A number appearing in 2, 3, or 4 of the lists obviously passed the survey point, and a number appearing in only one of the lists was also included if it was clearly not a mistaken version of another number. The observers subsequently confirmed that the "agreed list" so drawn up was an accurate record of the vehicles actually passing the survey point. This list was then used as the basis for the accuracy calculations.

Table 7.2 Statistics for moving traffic data collection exercise.

	NUMBERS RECORDED	NUMBERS CORRECT	NUMBERS WRONG	NUMBERS MISSED	% OF AGREED LIST RECORDED	% OF AGREED LIST CORRECTLY RECORDED
TAPE 1	145	143	2	12	92.3%	91.0%
TAPE 2	143	139	4	14	91.7%	88.5%
PEN+PAPER	146	136	10	11	93.0%	86.6%
COMPUTER	132	121	11	25(*)	84.0%	77.0%

(*) Had a proficient typist been used to record the field data, better results than these would have been expected as a consequence

of not having to look at the keyboard while recording the numbers.

Table 7.3 Comparison of time taken to transcribe data and accuracy of transcription for all methods

	TAPE 1 SPEECH RECOGNISER	TAPE 2 SPEECH RECOGNISER	TAPE 1 AUDIO TYPIST	TAPE 2 AUDIO TYPIST	PENCIL AND PAPER TYPIST	HAND-HELD COMPUTER
SURVEY TIME	min sec 18 : 30	18:30	18:30	18:30	18:30	18:30
TRANSCRIPT. TIME TAKEN	18:30	18:30	18:30	18:30	20:30	0:30
TRANSCRIPT. ACCURACY	73.8%	31.7%	98.6%	95.8%	100%	100%
% OF AGREED LIST TRANSCRIBED CORRECTLY	67.1%	28.0%	89.7%	84.7%	86.6%	77.0%

Table 7.3, transcription times and accuracies, shows that the time taken by the audio typist to transcribe the tape-recorded data appears the same as the survey time. This is because the typist was able to keep up with the flow of data and did not need to stop the tape to check on any of the numbers. Approximately 11 of the 18.5 minutes of the tape was silence (i.e. redundant tape), and the audio typist could easily have transcribed data at twice the rate that appeared on this tape.

The time taken to transcribe the pencil-and-paper list includes the time taken to write it out neatly after the survey, the actual typing time being only 9 minutes.

Table 7.4 Results for first road inventory survey

	SURVEY TIME	TRANSCRIPTION TIME	TOTAL TIME	TRANSCRIPTION ACCURACY
INVENTORY SHEET DATA	min sec 13 : 00	2 : 30	15 : 30	100 %
TAPE RECORDED DATA	5 : 10	5 : 10	10 : 20	98 %

Table 7.5 Results for second road inventory survey

	SURVEY TIME	TRANSCRIPTION TIME	TOTAL TIME	TRANSCRIPTION ACCURACY
	min sec			
INVENTORY SHEET DATA	8 : 00	approx 3 : 00	11 : 00	?? %
TAPE RECORDED DATA	4 : 20	4 : 20	8 : 40	86 %

No individual error rates were available for the inventory sheets typed in at WYCC. The procedure for error checking is that after the results have been typed into the computer, a printout is given to the surveyors who check any errors indicated against the original inventory sheets. The corrections would then be handed back to the typist for re-input to the computer.

In the case of the speech recogniser, the software for the data transcription process allows any data that is recognised but which is out of context, (e.g. noise causing a number to be falsely recognised in the middle of a phrase such as "left footpath crossing"), to be placed in a "rubbish bin" at the end of the print-out. If the items in the rubbish bin can be easily and unambiguously recognised, then the operator can replace the corrected item in the correct classification.

When listening to all the recorded data it was obvious that there was a lot of blank (redundant) tape, i.e. lengthy breaks in the traffic, or long pauses between inventory items, and these periods of blank tape mean that the transcription time using a speech recogniser or an audio typist will always be as long as the survey time and possibly longer. The advantage of using the hand-held computer or pencil-and-paper methods is that these blank areas are invisible in the transcription process, thus cutting down the transcription time. It might be possible to pre-process the tapes and edit out the blank patches, but this will increase overall the time spent on the transcription process. One obvious development in the speech recogniser would be to have it automatically wind through the blank areas quickly to the next area of data, thus reducing the transcription time.

8. CONCLUSIONS

The amount of training given to users of the speech recognition equipment is obviously of great importance in determining the accuracy of the transcribed data. From the results presented above it can be seen that a minimum of about 7 hours training is required to achieve reasonable rates of accuracy, and in some cases this would need to be extended considerably. The fact must also be faced that certain people find it very difficult to acquire the necessarily precise way of speaking, e.g. trainee C found it extremely difficult to avoid running the words together. While the tapes recorded by these people seem perfectly intelligable to the ear, the speech recogniser was unable to

separate the individual words. Unfortunately this would not be apparent until after some of the training had taken place, resulting in wasted time and money.

The results shown in table 7.3 show that the traditional methods of data collection and transcription give higher overall accuracies for data transcription. For road-side surveys of moving vehicles, recording the data onto cassette and then using an audio typist to transcribe it gives the best overall results at the flow rates specified, with pencil and paper (transcribed by a proficient typist) coming a close second. The results obtained from the hand-held computer are probably of lower accuracy than would be expected if a proficient typist was the operator. On the parked vehicle survey, (with more time available for the operator), the hand-held computer would appear to be the best option, since very little time is required for subsequent transcription of the data and with a guaranteed transcription accuracy of 100%.

When the cost of the equipment is taken into account along with the cost of training and the complexity of use, we must conclude that for the application of recording registration plate data in moving traffic, speech recognition is not the best solution. If certain conditions could be met, e.g. low traffic flow rates, and assuming that the price of speech recgnisers will fall rapidly in the near future, then it would be fair to re-assess their usefulness for this task. A continuous word recogniser might perform better in this application, but we did not have access to one in time to include an appraisal in this report.

For the task of recording registration plate data of stationary vehicles, the results show that, given adequate training of the users, speech recognition could be used to transcribe the data with accuracies of >90%, including the increase in accuracy obtained by using the Sony Walkman tape recorder.

The results presented in Tables 7.4 and 7.5 show that road inventory surveys can be carried out in less time using the tape recorder than by traditional methods, mainly because the operator does not have to keep on looking down at his instruments. The data item in the first survey that was wrongly transcribed by the speech recogniser was due to vehicle noise, and could easily have been moved from the rubbish bin to it's correct position. The transcription errors in the second survey were due to one word twice being recognised as another. This problem could probably be overcome by more training for the operator, but more likely another set of templates would cure the problem.

The selection of equipment is obviously of vital importance to achieving good results. The microphone recommended by both speech recogniser manufacturers was the Shure SM10A and this was found to perform well under all conditions. Being a head-set mounted microphone, no problems are experienced with keeping it at an optimum distance from the mouth, and the hands are always free for other tasks. The tape recorder should also be of good quality, the main restriction here being that it must be battery powered. It should not

have an automatic gain control, which tends to boost the background noise in quiet periods. The Sony Walkman fulfils all these criteria and has the additional advantage of being very light in weight.

9. REFERENCES

MONTGOMERY, F.O. (1984) Accurate travel times using portable tape recorders. Traff. Eng. and Control. 25(6) 310-313.

KIRBY, H.R. (1983) The automated transcription of spoken traffic data. Project Description, Technical Note 124, ITS, University of Leeds.

MARCONI COMPANY LTD (1983) SR-128 Speech Recogniser Handbook.

INTERSTATE ELECTRONICS CORPORATION (1983) Voice Recognition System Model SYS300 Maintenance Manual.