This is a repository copy of *The Automatic Transcription of Tape-Recorded Data.*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/2346/

Monograph:

Bonsall, P.W., Hardwick, B.A. and Kirby, H.R. (1985) The Automatic Transcription of Tape-Recorded Data. Working Paper. Institute of Transport Studies, University of Leeds , Leeds, UK.

Working Paper 205

Institute of Transport Studies
University of Leeds

This is an ITS Working Paper produced and published by the University of Leeds. ITS Working Papers are intended to provide information and encourage discussion on a topic in advance of formal publication. They represent only the views of the authors, and do not necessarily reflect the views or approval of the sponsors.

White Rose Repository URL for this paper:
http://eprints.whiterose.ac.uk/2346/

Published paper
Bonsall, P.W., Hardwick, B.A., Kirby, H.R. (1985) *The Automatic Transcription of Tape-Recorded Data.* Institute of Transport Studies, University of Leeds, Working Paper 205

THE AUTOMATIC TRANSCRIPTION

OF TAPE-RECORDED DATA

P.W. BONSALL, B.A. HARDWICK and H.R. KIRBY

Contents

ABSTRACT

BONSALL, P.W., HARDWICK, B.A. and KIRBY, H.R. (1985) The automatic transcription of tape-recorded data. Working Paper 205, Inst. Transp. Stud., Univ. Leeds. (Paper presented at the PTRC Summer Annual Meeting, University of Sussex, July 1985). Price: £0.75.

There are a number of situations where, because of the rapidity of events, it is easier to record data by speaking it onto an audio cassette rather than recording it by other means, e.g. paper and pencil recording. The problem with spoken data is that considerable effort is then required to transcribe the data onto a computer for subsequent analysis. This paper describes the use of automatic speech recognisers for transcribing such data recorded under both laboratory and field conditions. Different types of recognisers are compared along with associated equipment (microphones and tape recorders). 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.

THE AUTOMATIC TRANSCRIPTION OF TAPE-RECORDED DATA

Peter Bonsall, Bryan Hardwick, and Howard Kirby

1. BACKGROUND

There are a number of situations where data is efficiently captured by audio tape recording. Typically these are where there is a need to keep one's eyes on the phenomenon being recorded, or to keep one's hands free for some other task. Such situations limit the use of a written or key-punched recording technique.

Audio tape recording methods have been tried out in a number of transport applications, including inventory surveys, manual classified counts, and registration plate surveys. The last of these is a particularly good example because traditional, (pencil and paper) methods are so obviously unable to deal with high traffic flows.

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). The possibility of using speech recognisers to automatically transcribe tape recorded data onto a computer has been investigated under a grant from the Science and Engineering Research Council. This has enabled us to purchase two speech recognisers to assess their usefulness in transport and traffic data collection exercises.

2. SPEECH RECOGNISERS

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.

2.1 Recognisers evaluated

At the outset of the 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.

2.2 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.

3. WORK PROGRAMME

The work in this project falls into three clearly defined sections, namely :

     (i) laboratory tests
    (ii) field tests
   (iii) comparative trials .

The laboratory tests were mainly concerned with getting to know the speech recognition equipment and its associated problems, and also to assess the suitability of the tape recorders and microphones that were available.

The immediate application in mind for the speech recognition equipment was the transcription of tape recorded registration plate data, usually just the digits and the year letter and possibly including a vehicle type as well. For the laboratory tests, programs were written on a BBC micro-computer to display a series of random registration numbers at variable time intervals. 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 accuarcy of both the observer and the transcription process could be determined, and the effect of varying data rates investigated.

Having used the laboratory tests to determine the best combination of equipment and method of use, some field trials 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 a complete comparative study was set up. This involved training a small group of people to use all the equipment, and then carrying out a survey of registration plates 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.

4. RESULTS OF LABORATORY AND FIELD TESTS

4.1 Selection of equipment

As already stated the speech recognisers available for this project were the SYS300 and the Marconi SR128X. 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

4

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.

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
P L A Y E D	B A C K O N	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
P L A Y E D	B A C K O N	HITACHI DECK	58% – 70%	83% – 85%
		HITACHI PORTABLE	48% – 75%	N/A
		SONY WALKMAN	N/A	88% – 90%

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.

4.2 Vocabulary size and content

As stated, our main use for speech recognition was initially to be for recorded registration plate data. 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. The main problem with using a phonetic alphabet is that the time taken to say the multi-syllable words slows down the rate of data entry even further.

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 overcome 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.

4.3 Data entry rates

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 .

6

The second factor applies only to the isolated word recogniser, as a connected word recogniser does not require pauses between words, only between phrases. It was this point that showed up clearly the shortcomings 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.

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 4.3. 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 4.3. Accuracy of transcription as a function of vocabulary size and data rate.

		VOCABULARY SIZE				
		10 (DIGITS ONLY)	15 (DIGITS + VEHICLES)	22 (LETTERS ONLY)	32 (LETTERS+ DIGITS)	37 (ALL THREE)
D	500	100%	100%	100%	100%	100%
A	750	100%	100%	100%	98%	96%
T A	1000	100%	100%	90%	80%	75%
R	1250	100%	100%	80%	70%	
A T	1500	100%	100%	75%		
E	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 4.3 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.

4.4 Comparison between laboratory and field trial results

This brief comparison serves to introduce a topic that will be covered in greater depth in the next section. Using a 37 word vocabulary, (i.e. digits, phonetic alphabet and vehicle types), data rates up to 1000 vehicles/hour were recorded in both laboratory and field conditions and the transcription accuracies were then obtained.

Table 4.4. Comparison of transcription accuracy between preliminary laboratory and field trials.

		LABORATORY TRIAL	FIELD TRIAL
D R	500	100%	88%
A A			
T T	750	98%	77%
A E			
	1000	74%	35%

Table 4.4 shows that the laboratory trials gave considerably better results than the field trials. Much of this can be ascribed to the increased background noise on the tape recordings, even when using the Shure microphone. Furthermore, listening to the recordings made on the field trials it is clear that a typical operator modulates the level of his voice to suit the level of ambient noise. Some allowance

can be made for this when the templates are recorded, this is discussed in the next section.

5. TRAINING PROGRAMME AND DATA COLLECTION SURVEY

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.

5.1 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.

9

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.

5.2 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 5.1 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 5.1   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 5.2 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 5.2   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 5.3 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 5.3   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.

5.3   Comparison of data collection and transcription methods

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.

Table 5.4 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.

5.4 Results of trials

The results of the comparative data collection exercise for both parked and moving traffic surveys are shown in the following tables.

Table 5.4 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 5.5 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 5.5 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 5.6, 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.

When listening to the recorded data it was obvious that there was a lot of blank (redundant) tape, i.e. lengthy breaks in the traffic, 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.

Table 5.6 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%

6. 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 5.6 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 recognisers 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 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.

15

7. FURTHER WORK

Following on from the conclusions drawn above it would seem that an application with a lower data rate would be more suitable for speech recognition. One such application is that of street furniture inventory surveys. At present the surveyor has a distance recording wheel, a bank of tally counters, and a clip board to deal with as he is making the survey. This equipment could be replaced by a single tape recorder and the distance wheel, with all items surveyed being recorded as a single word, this being the sort of data most accurately transcribed by the speech recogniser. Trials of such a system are now being conducted, and results should be available for presentation at the conference

8. 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.