Proceedings of the Institute of Acoustics

DETECTION AND CORRECTION OF USER AND DEVICE ERRORS IN THE USE OF AUTOMATIC SPEECH RECOGNITION

C. BABER, R. B. STAMMERS and R. G. TAYLOR

Applied Psychology Division,
Aston University,
Aston Triangle,
Birmingham.
B4 7ET

Previous research into the detection and correction of errors, in the use of Automatic Speech Recognition, has tended to concentrate on device misrecognitions. Such an approach neglects the causes and results of human error in ASR use. We discuss a number of user errors in ASR use, and suggest possible remedies. Error correction is considered in the context of control room systems. In such systems, operators rely on high resolution graphical displays for feedback. We suggest that ASR feedback could conceivably be provided using symbols on these displays. However, while this provides information concerning plant operation, textual feedback is required for error handling.

1. TYPES OF RECOGNITION ERROR

Automatic Speech Recognition (ASR) is based upon principles which match human speech with stored representations of spoken words. Devices differ in terms of how the representations are derived and how matching is performed. The majority of the techniques employed in commercially available ASR devices attempt to model dynamic speech signals with static representations. This inevitably results in problems of matching, and will lead to various types of recognition error. Techniques for capturing the dynamics of speech, such as Hidden Markov Modelling, have been refined in the past decade and are beginning to be employed commercially. However, such techniques are still subject to recognition error.

Ringle and Bruce (1982) propose that, in human dialogue, failure of the listener to respond appropriately to speakers' words can arise from three factors (under normal circumstances). They define these factors as:

i.) Perceptual Failures - in which words are not clearly perceived, misperceived, or are constantly misinterpreted.

ii.) Lexical Failures - in which the listener perceives a word correctly but fails to interpret it correctly.

iii.) Syntactic Failures - in which all words are correctly perceived and interpreted, but the intended meaning of the utterance is misconstrued.

Of these, only (i.) can be directly applied to ASR use. Both (ii.) and (iii.) require a degree of intelligence on the part of the dialogue partner, which commercial ASR devices have yet to exhibit.

Typically, when humans make perceptual errors in speech processing, they are able to call upon their knowledge of the language to correct the error, or they can ask for the speaker to repeat the last word(s) spoken. In situations where such errors could prove critical, for example Air Traffic Control, there ahve been attempts to reduce the amount of spoken communication human are required to engage in by using other means of information communication (Matthews and Hahn, 1987). Obviously, the latter method of reducing perceptual error is not vialable for ASR use. This means that perceptual errors need to be dealt with either by repeating the misrecognised word(s) or by using some degree of intellignece to resolve errors.

There are three main types of recognition error that an ASR system can produce (Williamson and Curry, 1984). These can be related to the notion of perceptual error proposed by Ringle and Bruce (1982), in that they all involve some form of misperception on the part of the recogniser. The most common type of recognition error is the **substitution** error. This type of error occurs when an incorrect item is substituted for the spoken one. Brown and Vosburgh (1989) found that over 90% of recognition errors, in their experiments, were due to substitution. One can characterise such errors as the misperception of received speech; the user says one word and the deivce 'recognises' another.

Insertion errors occur when spurious noise is recognised as a legal vocabulary item. These account for between 5% and 6% of recognition errors. **Rejection** errors are the least common form of recognition error. They occur when a legal vocabulary item is spoken by the user and the device does not respond, such as would be expected if a problem exists in the communication between user and device. In their studies, Brown and Vosburgh (1989) found that rejection errors accounted for between 2% and 3% of recognition errors. Insertion and rejection errors can be characterised as errors arising because the device could not clearly perceive the words spoken.

While insertion and rejection errors can be minimised by using adequate communication channels, such as a good quality microphone and noise cancelling techniques, substitution errors are harder to define. They can result from the occurrence of similar sounding words in the vocabulary, or from similar templates being created at enrolment. To some extent the vocabulary can be tailored to reduce the number of confusable words, but the vocabulary will inevitably be determined by the tasks one wishes to perform with ASR. Similar templates are difficult to detect and could result in dissimilar sounding words being confused due to patterns of noise being present at enrolment. This could be reduced by enrolling each word several times, but this could still produce some traces of spurious noise which will lead to confusion. It seems that the only way to deal with such errors, as they cannot at present be designed out of system configuration, is to provide some form of intelligence which can correct them. Such intelligence can be programmed into the device or can be left to the user utilise.

Manufacturers presently claim a recognition accuracy rate of 98%+ for their devices. In the workplace, recognition accuracy varies greatly. We have observed accuracy in the range of 45-90%. This means that recognition errors are highly probable. While errors made by ASR devices in recognising speech have been studied by several researchers (Martin and Welch, 1980; Spine et al., 1983; Little and Joost, 1984; Schurick et al., 1985; Dreizen, 1987; Ainsworth, 1988; Baber et al., 1990), there has been little research into possible causes of user error.

This is somewhat surprising given the current concern in human factors research for design to minimise user error (see Lewis and Norman, 1986). One of the more irksome problems for users of ASR is that recognition errors appear to occur irrespective of user action. Peckham (1986) has noted that a keyboard has a 'standardising' effect on user actions. That is, providing one strikes the correct key, it is unimportant how the key is struck. Using ASR, on the other hand, requires the user to not only speak the correct word, but also to perform the speak the word correctly (within the constraints of the devices matching algorithms). For this reason, the possibility of user error requires investigation. Furthermore, Frankish and Noyes (1990) have shown that, in a task involving entry of strings of digits, while device errors account for approximately two thirds of all errors, approximately one third of errors were due to user error.

2. USER ERROR

The area of user error in ASR systems has not received serious study. As there are relatively few ASR systems in use, and as these have only been in operation for a short period of time, it is difficult to obtain accident statistics relating to ASR. However, it is possible to draw hypotheses from the general literature on human error (Norman,1981; Reason and Mycielska, 1982; Reason, 1986), which can be tested experimentally.

One can propose user errors will result from an error of intention, such as the incorrect selection of an action, or from errors in execution of the action (Berman, 1986). Reason and Mycielska (1982) define errors of intention as 'mistakes' and errors of execution as 'slips'. Mistakes can be defined as errors at the planning level of action, specifically in terms of interpreting the situation. Slips can be described as errors at the production level of action. Given this distinction between two basic types of human error, we will deal first with production level errors, slips, as these are the hardest type of error to predict and have, as yet, little empirical support. There has been limited research into the effects of mistakes in ASR use. Methods of reducing user error are considered in the light of results from these papers.

2.1. Slips in ASR use

In terms of using ASR slips can be related to the production of spoken commands. The simplest type of slip would occur when vocabulary items are mispronounced due to users introducing spurious noise, such as yawning, into their speech. Users could introduce

overlong pauses into their commands, as a result of being distracted by another task. There is an extensive literature concerned with speech errors known as 'slips of the tongue' (Fromkin, 1980), but it is difficult to propose why such slips occur, or how to predict them. Even though it is not possible to predict where slips in speaking will occur, it is important to provide the user with some means of correcting system errors that these slips cause. Error correction is discussed in section three.

Slips can also occur in the interaction between user and computer. Frankish and Noyes (1990) have found that if users receive feedback visually, they are prone to errors in detecting misrecognitions. That is, users do not notice the misrecognitions. This is not the case when feedback is auditory. Ito et al. (1989) found that when auditory feedback was provided, users would wait until their speech had been echoed before speaking the next word. For isolated word recognition devices, at least, forcing users to wait until the device has correctly recognised a word will reduce the likelihood of user error. However, the use of auditory feedback has a number of problems associated with it when it comes to error correction tasks (see below). Further, it will interfere with the use of connected word recognisers. Finally, Thomas and Rosson (1984) found that, given the opportunity, users prefer to interrupt synthesised speech messages. Presumably this allows the user to control the rate at which they receive the message, to make processing easier.

Baber et al. (1990c) utilised a simulation of a connected word ASR device to recognise spoken command strings in a process control task. They found that if feedback was presented in a text window, below a display of a process plant they were controlling, users ignored around 7% of feedback. If the feedback was symbolic and incorporated into the display, users only ignored 4%. These results suggest that if feedback of recogniser performance is incorporated into the users' primary task, they will be less likely to make slips in error detection, than if feedback monitoring constitutes a task in itself.

2.2. Mistakes in ASR use

Mistakes could result from a number of factors in ASR. The user might attempt to use an illegal word to issue a command. This could be due to the user being more familiar with a synonym of the command word than the word employed. For this reason, vocabularies need to be designed which users find easy to use and remember.

Poock (1980) has noted that function keyboards provide memory cues for users, in the labelling or coding of the keys. ASR does not normally offer such cues. Legal vocabulary options could be displayed to the user, but this may require screen space which is not available. If the vocabulary was designed to be not only task specific, but also habitable (Watt, 1968), users would know which words were required for which operations. This suggests that efficient vocabulary and dialogue design could reduce such mistakes.

Another likely source of user errors is the fact that users can often have difficulties when faced with the restrictions imposed on their style of speech by isolated word recognition devices: users try to speak too fast for the device. However, Brown and Vosburgh (1989) examined the occurrence of 'segmentation errors' in the use of an isolated word ASR device. Segmentation errors arose when users either spoke too quickly for the device, coarticulated

words, or spoke too slowly, introducing pauses into words. Of all the recognition errors they recorded, Brown and Vosburgh (1989) found that segmentation errors only accounted for between 0.3% and 0.8%. This shows that users are capable of adapting to isolated word ASR devices, given adequate training and practice. With the increasing availability of connected word ASR devices on the market, one might question the utility of this observation. However, it supports the observation that users are capable of adapting efficiently to ASR use (Baber and Stammers, 1989), even when the ASR device changes the form of interaction (Zoltan Ford, 1984).

Given that ASR is not 100% efficient, it is necessary to provide users with feedback concerning the performance of the recogniser, and a means of correcting errors. Users can also make mistakes in their use of feedback from the device and in correcting recognition errors.

Baber et al. (1990c) found that, in addition to ignoring feedback (see 2.1), subjects were also prone to misreading feedback. If feedback was provided in a text window, adjacent to the process they were controlling, subjects misread almost 10% of feedback. "Misreading" was defined as the inappropriate response to feedback, in terms of confirmation of commands. Symbolic feedback, incorporated in the display next to the valve to which it referred, was misread on only 4% of occasions. Again, this suggests that incorporating feedback into the primary task will reduce the likelihood of user error.

Users can also make mistakes in error correction tasks. Baber et al. (1990a) report a study in which users were prompted with "Did you say <X>", when recognition failed to reach a specified threshold. It was found that, rather than users responding 'yes' or 'no', as required, they would repeat the word. Generally the repetition was louder than the first attempt, resulting in a further increase in the deviation between speech and template. Error correction presents a number of problems for designers of systems using ASR, and is discussed in section three.

3. ERROR CORRECTION

The simplest form of error correction dialogue would require the user to say "yes" or "no" after each item has been recognised and displayed. This will be extremely time consuming and irritating to use (Leiser et al., 1987). Little and Joost (1984) suggest that the user need not respond to correctly recognised words, and simply respond "no" or "delete" to misrecognitions.

This type of error detection dialogue can be extended to cover whole commands. Martin and Welch (1980) propose that a buffer could be used to store words as they are spoken. Verbal commands could then be used to edit the information in this buffer before the command is sent. For example, the user could say "o.k." to verify the whole of the command in the buffer, or they could "erase" individual words, or "cancel" the entire command. Spine et al. (1983) found that where subjects had the option to correct errors on an individual item or whole command basis, they tended to prefer individual item error correction. Schurick et al. (1985) found that individual item error correction was used 48% of the time, and whole

command error correction was used only 12% of the time.

Martin and Welch (1980) point out that an obvious problem in the use of spoken error correction commands is that these commands might themselves be misrecognised. This is shown to be a major problem by Frankish and Noyes (1990). A second problem is that the error correction dialogue imposes an intervening task between the user and the primary task .

This problem can be overcome by displaying the recognised command string and allowing the use of word repetition to correct errors. Such an approach may appear overly simplistic, but is deemed more natural than error correction dialogues (Baber et al., 1990a) as it does not interfere with the primary task of command entry.

However, while such an approach makes sense in human factors terms, it will require some degree of "intelligence" on the part of the host computer to decide which words are being corrected. In limited vocabulary domains, this can be achieved by judicial use of syntax constraints or careful vocabulary design. In larger vocabularies, one needs a more robust approach.

Dreizin (1987) proposes a method of intelligent error detection which employs a combination of well designed syntax with the use of 'second choice' words from the recognition process, to make guesses as to where errors have occurred. While such intelligent error detection is feasible it is necessary to allow the user ultimate decision as to the appropriateness of the proposed corrections. It is perfectly feasible for the error correction algorithms to distort the initial utterance into a well formed but incorrect string.

4. FEEDBACK FOR ERROR CORRECTION

We have argued elsewhere (Baber et al, 1990b) that correcting recognition errors in ASR use, is a form of verbal decision process. Consequently, the user requires sufficient information concerning what the device has recognised to allow a decision concerning its accuracy to be made. Such information should be presented, visually, in the form of a string of text.

While visual information is easy to present to the user, and for the user to attend to, Schurick et al (1985) found that auditory feedback improved data entry time by a factor of three. However, auditory feedback has a number of problems associated with it, not the least of which is the fact that it is transitory. This makes error detection and correction difficult, especially in a long string of words. One could allow users the opportunity to interrupt the message (Thomas and Rosson, 1984). However, this will require error correction to be carried out during the interruption and will inevitably disrupt primary task performance. Alternatively, one could limit the string length to around eight words, which appears to be an optimum length (Ainsworth, 1988). Obviously, the type of feedback used will depend on the situation in which one is using ASR (Schurick et al 1985).

We have pointed out some of the possible human errors which could occur in ASR use. Human error can be reduced by careful system design. This will mean that attention can be

given to developing means of correcting and detecting recognition errors. We have suggested that insertion and rejection errors can be minimised without too much difficulty. Future research in error correction should, therefore, concentrate on the problems and causes of substitution error.

5. REFERENCES

W.A. Ainsworth, 'Optimisation of String Length for Spoken Digit Input with Error Correction', Int. J. Man Machine Studies 28 pp.573-581 (1988)

C. Baber and R.B. Stammers, 'Is it Natural to Talk to Computers? An Experiment using the Wizard of Oz Technique', Contemporary Ergonomics pp.234-239 (1989)

C. Baber, R. B. Stammers and D.M. Usher, 'Error Correction Requirements in Automatic Speech Recognition', Contemporary Ergonomics, pp.454-459 (1990a)

C. Baber, R.B. Stammers and R.G. Taylor, 'Feedback Requirements for Automatic Speech Recognition in Control Room Systems', In Ed. D. Diaper et al Interact'90 Amsterdam: North Holland pp. 761-767 (1990b)

C. Baber, D.M. Usher, R.B. Stammers and R.G. Taylor, 'Feedback Requirements for Automatic Speech Recognition in the Process Control Room',Int. J. Man Machine Studies (Paper submitted) (1990c)

J.V.F. Berman, Speech Recognition Systems in High Performance Aircraft: Some Human Factor Consideration, I.A.M. Report No. 646 (1986)

N.R. Brown and A.M. Vosburgh, 'Evaluating the Accuracy of a Large Vocabulary Speech Recognition System', Proc. Human Factors Society 33rd. Annual Meeting, pp. 296-300 (1989)

F. Dreizen, 'An Alternative Way to Handle Errors: the Sieve Method for Error Detection and Correction', Int. Speech Tech'87 192-195 (1987)

V.A. Fromkin, Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand, New York: Academic Press (1980)

Ito, N., Inoue, M., Ohkura, M. and Masada, W. 'The Effect of Voce Messages on Interactive Computer Systems' In Klix et al Man Computer Interaction Research: Macinter II Amsterdam: North Holland pp.245-253 (1989)

R.G. Leiser, M. de Alberdi and D.J. Carr, 'Generic Issues in Dialogue Design for Speech Input / Output', European Conf. on Speech Technology, vol. 2 pp. 69-72 (1987)

C. Lewis and D.A. Norman, 'Designing for Error', In Norman, D.A. and Draper, S.W. User Centred System Design, Hillsdale, N.J.: L.E.A. pp.411-432 (1986)

A. Little and M.G. Joost, 'A Comparison of Two Error Correction Strategies in Voice Data Entry', Proc. Voice I/O Systems Application Conf. (1984)

T.B. Martin and J.R.Welch, 'Practical Speech Recognisers and Some Performance Effectiveness Parameters', In Lea, W.A.Trends in Speech Recognition , Englewood Cliffs, N.J.: Prentice Hall (1980)

B.G. Matthews and C.K.G.Hahn, 'Voice Communication Errors in the Air Traffic Control Environment', 20th. Annual Conference of the Human Factors Society of Canada pp.167-170

User Error in ASR

D.A.Norman, 'Categorisation of action slips', Psychological Review, 88 pp.1-15 (1981)

J. Peckham, 'Human Factors in Speech Recognition', In Bristow, G. Electronic Speech Recognition, London: Collins (1986)

G. Poock, Experiments with Voice Input for Command and Control: using Voice Input to Operate a Distance Computer, Monterey, Naval Postgraduate School: Report NPS55- 80- 016 (1980)

J.T. Reason, 'Actions not as Planned', In Underwood, G. and Stevens, R. Aspects of Consciouosness London: Academic Press

J. T. Reason and K. Mycielska, Absent Minded? The Psychology of Mental Lapses and Everyday Errors, Englewood Cliffs, N.J: Prentice Hall (1982)

M.H. Ringle and B.C. Bruce, 'Conversation failure', In Lehnert, W.G. and Ringle, M.H. Strategies for Natural Language Processing Hillsdale, N.J.: Erlbaum (1982)

J.M. Schurick, Williges, B.H. and Maynard, J.F., 'User Feedback Requirements with Automatic Speech Recognition', Ergonomics 28 pp. 1543-1555 (1985)

T.M. Spine, Maynard, J.F. and Williges, B.H., 'Error Correction Strategies for Voice Recognition', Proc. Voice Data Entry Systems Application Conf. (1983)

J.C. Thomas and M.B. Rosson, 'Human Factors and Synthetic Speech', Proc. Human Factors Soc. pp. 763-767 (1984)

W.C. Watt,'Habitability', American Documentation, pp. 338-351 (July, 1968)

D.T. Williamson and D.G. Curry, 'Speech Recogniser Performance Evaluation in Simulated Cockpit Noise', Speech Technology (1984)

E. Zoltan Ford, 'Reducing Variability in Natural Language Interactions with Machines', Proc. Human Factors Society 28th. Annual Meeting, (1984)

Proceedings of the Institute of Acoustics

TEMPLATE TRAINING CONDITIONS AND RECOGNISER PERFORMANCE IN SIMULATED VOICE-DIALLING IN A NOISY ENVIRONMENT

W A Ainsworth & S R Pratt

Department of Communication and Neuroscience, University of Keele, Keele, Staffordshire ST5 5BG, U.K.

1. INTRODUCTION

Speech recognition machines are particularly likely to make detection errors in noisy environments where even human listeners often experience difficulty in understanding speech. In circumstances where the accuracy of the message is important listeners check that they have heard it correctly by repeating it and asking for confirmation.

A pattern-matching speech recogniser would be expected to produce optimum performance when the background acoustic conditions for template training and for system evaluation are identical. In practical applications, however, it may not always be desirable, or even possible, for the training to be carried out in the conditions which obtain when the recogniser is in use.

A speech recogniser used for voice-dialling in a car telephone installation is likely to suffer impaired performance owing to noise from a variety of sources. Such a system must provide feedback to the user before it attempts to dial out. This feedback, where the machine repeats instructions, would be "secondary" in the terms described by Schurick *et al.* [1].

A number of strategies can be employed for error correction. The number can simply be repeated to the user by synthetic speech. If an error occurs the user can say "no", "wrong" or "correction" and repeat the number. Ainsworth [2] recently investigated the effect of recognition rate on the optimum number of digits uttered before feedback is given using a strategy of this type. If feedback is given after each digit, often desirable in noisy conditions, more rapid correction can be provided if words already rejected are removed from the active vocabulary of the recogniser. A further possible improvement might be to allow the recogniser to suggest the next most likely digit according to the output of the recognition algorithm. This saves the user the need to repeat the word but denies him the possibility of making a more typical utterance or of choosing an occasion when less environmental noise is present.

This study investigated the effect of template training in silence and in the presence of recorded car engine idling noise on recogniser performance assessed during playback of the noise generated by the same car at varying speeds on the open road. A number of error-correcting strategies were also examined.

2. ERROR-CORRECTING STRATEGIES

2.1 Simple Repetition

In this strategy feedback is given after each digit and when an error is made the user repeats the word

and the recogniser is free to try the same digit repeatedly. This is liable to result in considerable waste of time and much user frustration.

2.2 Repetition with Elimination

This strategy is similar to the above but the recogniser does not suggest words which have already been rejected by the user. This procedure often produces rapid correction of errors.

2.3 Elimination without Repetition

When an error occurs the machine suggests the next most likely word. This strategy can be time consuming if the original utterance was atypical or occurred in the presence of loud background noise. The user does not have the opportunity of assisting the recogniser by repetition of the word which can result in some frustration.

3. METHODS

Experiments were carried out using a speech input/output system consisting of a recognition and synthesis board (Loughborough Sound Images[3]) installed in a PC-AT. The experiments were designed to simulate voice-dialling of telephone numbers in a car but all training and testing took place in a laboratory using recorded car noises. Two training conditions were used: silence and recorded car engine idling noise. Testing took place in the presence of noise recorded in a car driven along main roads at various speeds.

Eight subjects took part in the experiments. They had a variety of English accents and their ages ranged from 21 to 53.

The subjects' task was to train the recogniser and "dial" four 14-digit numbers by voice, correcting recognition errors by the repetition with elimination strategy. A record of the fit of each word in the vocabulary enabled the elimination without repetition strategy to be evaluated. The vocabulary consisted of the digits "one" to "nine" inclusive, the words "oh", "hundred", "thousand", "double", "treble" and also the word "correction". The words "yes" and "no" were also trained but they were not in the active vocabulary during dialling. If the recogniser made an error, the subject said "Correction" and repeated the last word. The synthesiser then asked "Was it ____?" to which the subject replied "yes" or "no" accordingly.

4. RESULTS

Average recognition scores, with no attempt at correction, are shown for the two training conditions in Table 1. The standard deviations were calculated from each subject's mean. There was a wide variation in the scores of individual subjects, but those who scored highly on one condition tended to score highly on the other, indicating a difference between subjects in consistency of pronunciation. Recognition scores were higher for training in idling noise (85.3%) than for training in silence (80.4%). A chi-squared test showed that this difference was significant at the 0.1% level.

TEMPLATE TRAINING CONDITIONS FOR A NOISY ENVIRONMENT

Table 1.

Recognition Rate (% Correct)

Train in:	Silence	Idling Noise
Mean	80.4	85.3
Std. Devn.	15.6	7.6

The acceptability of a speech recognition system is likely to depend as much on ease of error correction as on the recognition score. The number of corrections required to produce error-free recognition can be expressed as the percentage of extra utterances required. These percentages for the repetition with elimination strategy are shown in Table 2.

Table 2.

Repetition with Elimination Strategy (% Corrections)

Train in:	Silence	Idling Noise
Mean	46.5	28.5
Std. Devn.	47.7	19.0

Training in silence required 46.5% more utterances, while 28.5% more were required with training in idling noise. A chi-squared test showed that this difference was significant at the 0.1% level.

Table 3 shows, for each condition, the average number of corrections which would have been required if the elimination without repetition strategy had been adopted.

Table 3.

Elimination without Repetition Strategy (% Corrections)

Train in:	Silence	Idling Noise
Mean	56.7	34.8
Std. Devn.	63.2	22.4

A similar pattern emerges. Training in silence required 56.7% more utterances while only 34.8% more were needed with training in idling noise. A chi-squared test showed that this difference was significant at the 0.1% level.

The repetition with elimination strategy required fewer corrections than the elimination without repetition strategy in both conditions. Chi-squared tests showed that the difference between strategies was

significant at the 0.1% level for both training conditions.

The number of corrections may also be expressed as the average number of trials needed to input each digit correctly. The figures show the relationship between this metric and the probability of a correct recognition at the first attempt. Figure 1 gives the results for training in silence. The regression lines show that the relationship is approximately linear and that for any probability level slightly more corrections will be needed for the elimination without repetition strategy (2.3) than for the repetition with elimination strategy (2.2).

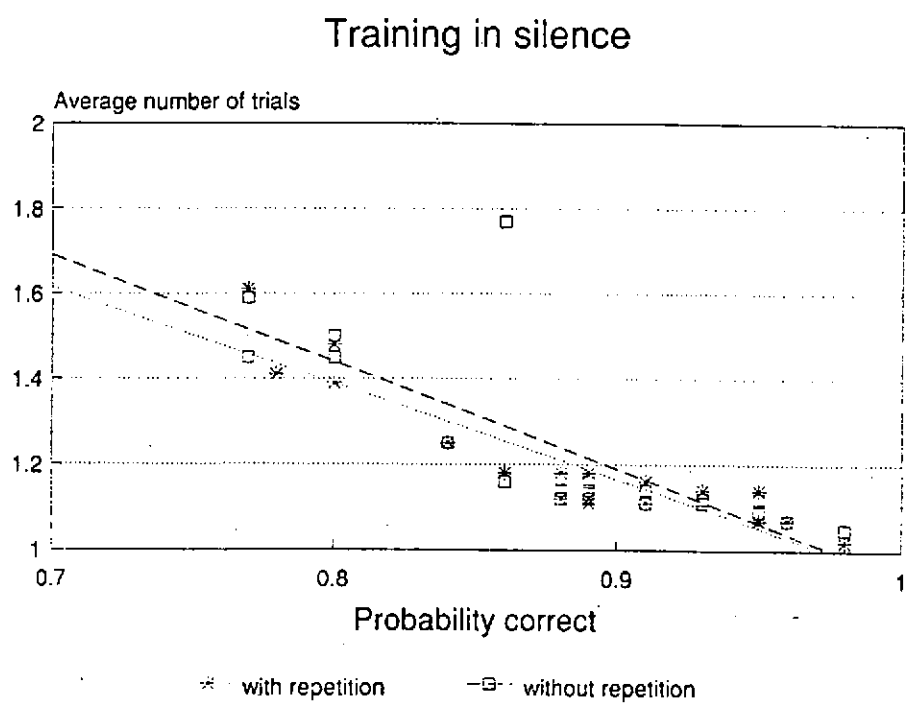

Fig. 1. Relationship between the probability of a correct recognition at the first attempt and the average number of trials needed to input a digit successfully, including corrections. Training in silence.

Figure 2 gives the corresponding results for training in idling noise. The regression lines of this figure show the same pattern although the data for the elimination without repetition strategy are subject to a degree of scatter.

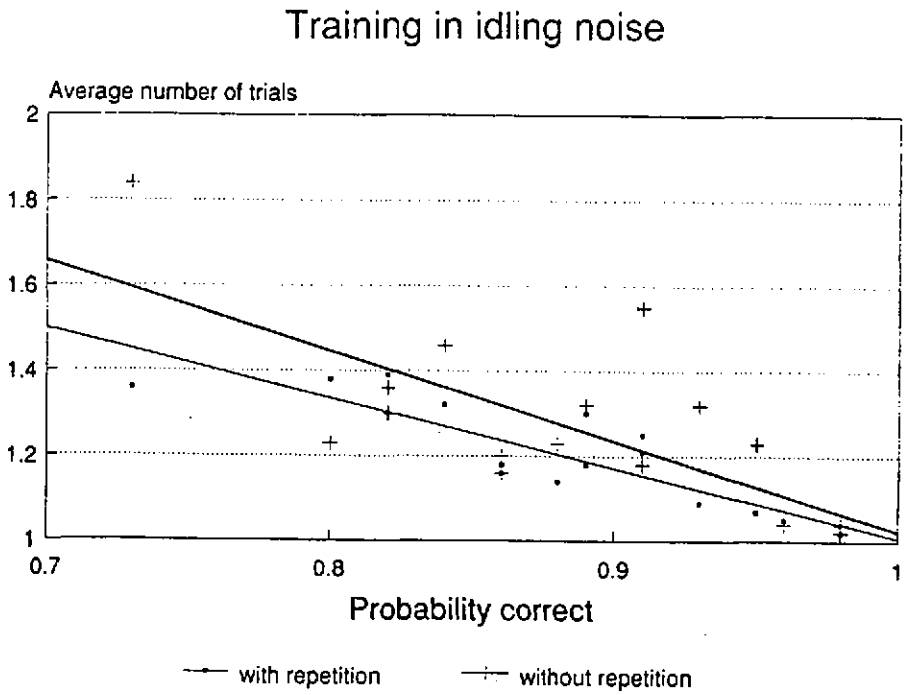

Fig. 2. As Fig. 1. but training in idling noise.

5. DISCUSSION

In spite of the variation in recognition scores between subjects (the "sheep and goats" phenomenon of Doddington & Schalk[4]), the results indicate that it is better to train a speech recogniser in the acoustic conditions in which it is going to be used than in silence. In the case of the car environment it is not practicable to train a recogniser while the car is in motion, but it is preferable to train it with the

car stationary and the engine running than with the engine switched off.

Recognition errors were corrected more easily if the user repeats the word that was misrecognised than if he allows the system to guess on the basis of the pattern matching differences. There are two possible explanations for this. An error may be caused by abnormal pronunciation, in which case when the user repeats it he is likely to be more careful. Alternatively the error may be caused by noise masking the signal. The repetition strategy gives the user the opportunity to repeat the word when the background is less noisy and correct recognition more probable.

6. CONCLUSIONS

If a speech recogniser is to be used in a noisy environment it is better for it to be trained in as close an approximation to that environment as possible.

Fewer corrections are required with a repetition with elimination strategy than with an elimination without repetition strategy.

There appears to be a linear relationship between the number of corrections required and the probability of the system identifying a word correctly at its first attempt.

7. ACKNOWLEDGMENTS

We thank staff and students of the University of Keele for acting as subjects in the experiments. The work was supported by EC ESPRIT Contract 2101 "Adverse-environment Recognition of Speech".

8. REFERENCES

[1] J M SCHURICK, B H WILLIGES & J F MAYNARD, 'User Feedback Requirements with Automatic Speech Recognition', Ergonomics, 28, p1543 (1985)
[2] W A AINSWORTH, 'Optimization of String Length for Spoken Digit Input with Error Correction', Int J Man-Machine Studies, 28, p573 (1988)
[3] LOUGHBOROUGH SOUND IMAGES, 'uPD7763/4 PC Card for Speech Recognition and Synthesis, Issue 3' (1988)
[4] G R DODDINGTON & T B SCHALK, 'Speech Recognition: Turning Theory into Practice', IEEE Spectrum, Sept. p26 (1981)