

## AUDIO FEEDBACK FOR ERROR CORRECTION IN A DIGIT RECOGNITION TASK

W.A.Ainsworth\*.

### ABSTRACT

No matter how much the performance of speech recognisers improves, it is unlikely that perfect recognition will be possible in all circumstances as environmental sounds interfere with recognition. In such circumstances it is necessary to provide feedback so that errors may be detected and corrected. In some situations, such as over the telephone, the feedback must be provided auditorily. The question arises as to whether this feedback should be provided after each word or after a group of words. It is shown that in the case of spoken digits this depends on the accuracy of the recogniser and on the times required for recognising the digits and for changing from recognition to synthesis mode.

### INTRODUCTION

When communicating with a computer by voice, feedback must be provided so that recognition errors can be detected and corrected. Over the telephone, or in situations where the eyes are busy, this feedback must be provided auditorily. If it takes a significant time for the system to change from recognition mode to synthesis mode, it will be more efficient for the feedback to be provided after a number of words have been spoken. However if recognition errors occur, the cumulative probability of an error will increase with the length of the word string. This suggests that a certain string length may be optimum, in terms of the average time required to input a word correctly.

The special case of recognising strings of spoken digits is considered. First expressions are developed for digit recognition time in terms of recognition accuracy and the durations of each of the processes involved, then these are verified by means of experiments with a recognition and synthesis system.

### THEORY

Suppose that the time required for a human to speak a digit and for the machine to recognise it is  $t_1$ , and the time for the machine to synthesise a digit is  $t_2$ . Suppose also that the time required for the human to change from speaking to listening and the machine to change from recognising to synthesising, and vice versa, is  $t_0$  (the overhead time). The time required to speak, recognise and check  $n_0$  digits will be:

$$t(n) = n(t_1 + t_2 + t_0/s) \quad (1)$$

where  $s$  is the length of the digit string between checks.

Suppose that the recognition rate of the machine is  $p$  for each digit. The probability of an error in recognising a string of length  $s$  will be:

$$q = 1 - p^s \sim (1-p)s \quad (2)$$

If a string is repeated every time a recognition error occurs, the number of strings which must be spoken for  $n$  digits to be recognised correctly is given by:

\*Department of Communication and Neuroscience, University of Keele, Staffs., ST5 5BG, U.K.

$$N=(1+q+q^2+q^3+\dots) \quad (3)$$

which may be simplified by substituting equ.(2) to:

$$N=n/s(1-s(1-p)) \quad (4)$$

The time taken to enter n digits correctly will thus be given by:

$$t_d(n)=n(s(t_1+t_2)+t_0)/s(1-s(1-p)) \quad (5)$$

By minimising  $t_d(n)$  with respect to s, it can be shown (ref 1) that the string length which gives the shortest input time per digit is given by:

$$s_m=(r^2+r/(1-p))^{1/2}-r \quad (6)$$

where the overhead ratio  $r=t_0/(t_1+t_2)$  (7)

Fig 1 shows how the optimum string length varies with the recognition rate for various values of the overhead ratio. For a high overhead ratio, string lengths greater than one are worthwhile if the recognition rate is greater than about 75%. For an overhead ratio of one, string lengths of greater than one are useful if the recognition rate is greater than about 85%, and if the overhead ratio is small ( $r=0.1$ ), string lengths greater than one are more efficient only if the recognition rate is greater than about 95%.

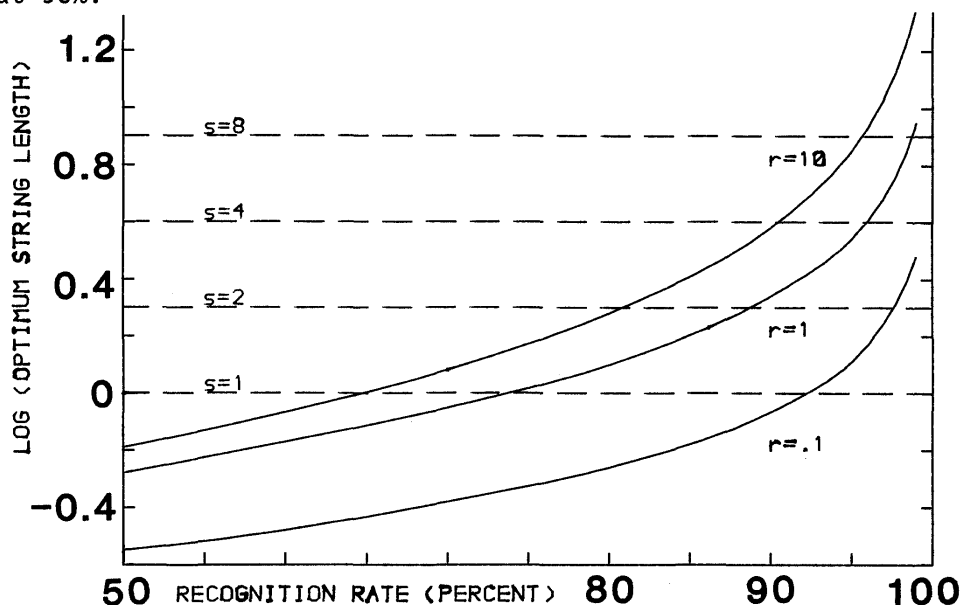


Figure 1. Optimum string length as a function of recognition rate for various values of the overhead ratio, r.

### EXPERIMENTS

In order to investigate whether equations (5) and (6) are applicable in a practical situation, some experiments were carried out with a speech recognition and synthesis system. The system employed was a Texas Instruments TI-Speech board installed in an IBM PC-AT microcomputer. The TI-Speech system consists of a speaker-dependent isolated word recogniser and a text-to-speech system (Ref 2). It is thus possible to interact with the system by speech alone.

The system was programmed so that a string of digits, of predetermined length, was selected randomly and displayed on the screen. The task of the subject was to read the digit string into a microphone. An auditory prompt was given when the system was ready to receive the next digit. When the string was completed, the text-to-speech system repeated the digit string as it had been recognised. If any recognition errors had

been made, the subject said the word 'no', and repeated the digit string. If the string had been recognised correctly, the subject read the next string of digits which appeared on the screen. The elapsed time from the start of the experiment until the last digit had been correctly recognised was measured by the computer.

In laboratory conditions and with a vocabulary consisting of the digits plus the word 'no', the recogniser performed very accurately. (In fact it made 5 errors in 800 digits.) In order to simulate lower recognition rates a random number generator was sampled to produce a number in the range 0 to 1 and if this was greater than the target recognition rate a different digit was substituted for the recognised digit.

In each session a subject was required to enter 40 digits into the computer. The string lengths were set at 1, 2, 4 or 8 digits and the target recognition rates at 80%, 90%, 95% and 100%. The combination of an 8 digit string length and 80% recognition rate was excluded because of the excessive duration needed.

Five subjects took part in the experiments. In the first session they were required to utter each of the words in the vocabulary twice in order to train the system. Their speech patterns were stored and used by the system for subsequent word recognitions. Random combinations of target recognition rate and string length were then chosen until all combinations had been covered.

## RESULTS

The mean digit recognition time for each of the conditions is shown in Fig 2. For 100% recognition rate equation (1) shows that the digit recognition time is a linear function of the reciprocal of string length. A linear regression of  $t_d$  versus  $1/s$  gave estimates of the  $t_1+t_2$  as 2.23 s and of  $t_0$  as 1.58 s, making an overhead ratio of  $r=0.71$ .

From Fig 1 it can be observed that a 95% recognition rate should give an optimum string length of about 4, a 90% recognition rate of about 2, and for lower recognition rates 1. Fig 2 shows that this was found to be the case for recognition rates of 90% and 95%. For 80% the minimum was found to be 2, but the digit recognition time for this is not significantly different from that for a string length of 1.

A comparison between the digit recognition times predicted from equation (5) and the observed recognition times is shown in Fig 2. It will be seen that the difference is generally within one standard deviation, except for a string length of 8 and a target recognition rate of 90%.

## CONCLUSIONS

Experiments have been performed to ascertain whether it is more efficient in a digit recognition task to perform checks after every digit or after a string of digits. The results suggest that the optimum string length can be estimated if recognition rate and the overhead ratio of the system are known.

1. W A Ainsworth, in preparation.
2. TI-Speech Programming Guide, Texas Instruments Inc., 1985.

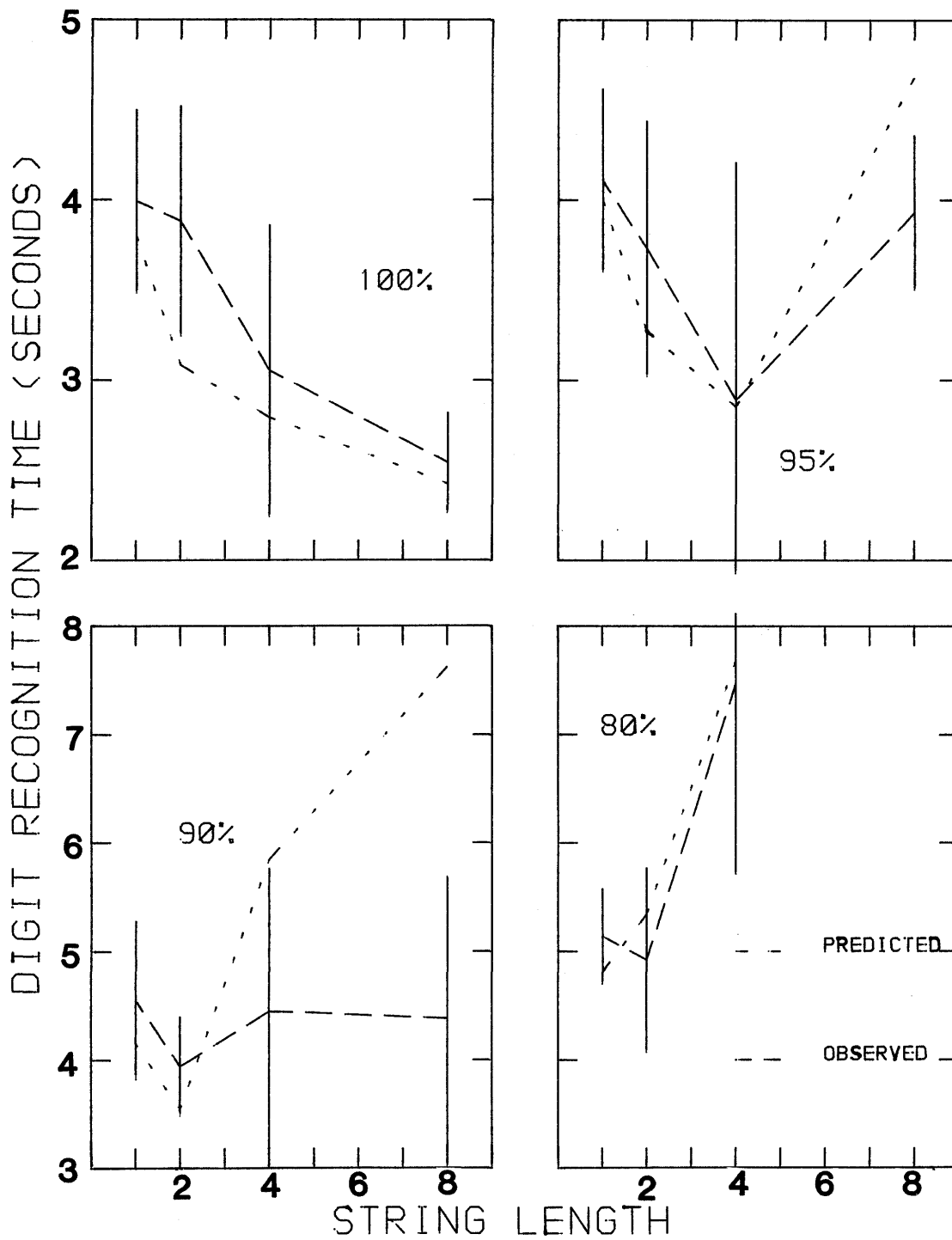


Figure 2. Average recognition time per digit as a function of string length for various recognition rates. The predicted functions are shown as dotted lines and the observed functions as dashed lines. Error bars are one standard deviation.