

PERFORMANCE AND EVALUATION CRITERIA FOR SIMPLE TES ISOLATED WORD RECOGNITION (IWR) SYSTEMS IN HOSTILE TACTICAL MILITARY ACOUSTIC ENVIRONMENTS

R C Power, R D Hughes and R A King.¹

ABSTRACT

A preliminary investigation into the performance of a simple Time Encoded Speech (TES) isolated word recognition (IWR) direct voice input (DVI) system is described. Experimental conditions included evaluations with four untrained military speakers in severe acoustic background noise (c. 80 - 100dB SPL) with hand-held omni-directional microphones. The limitations of conventional "percentage recognition" scores as a measure of system performance for the military role are discussed.

INTRODUCTION

Essential to all tactical military direct voice input (DVI) systems is an ability to operate effectively in a hostile, degraded and variable acoustic environment. The military system must also remain effective despite variabilities associated with extremes of mental and physical stress imposed upon the human operator. In addition, since the penalties for error in this role may be catastrophic, the operational user seeks 100% system integrity. Under such conditions, the use of "fail-safe" protocols would appear to be mandatory. Such protocols involve verification and feedback as part of system design, with n^{th} choice options available to avoid system lock-up, Ref. [1]. The n^{th} choice routine reduces the likelihood of catastrophic error, at the expense of user interaction with the recogniser. In the "fail-safe" mode, a few errors poorly ranked may incur more exchanges with the recogniser, than a larger number of errors ranked more closely to the first choice of the machine.

This paper describes a preliminary investigation into the performance of untrained military speakers exercising a simple TES DVI system in severe acoustic background noise, and comments upon the limitations of conventional "percentage recognition" scores as performance and evaluation criteria for the military role.

TES CODING

In the subject investigation, speech bandlimited to 4KHz (300 - 4.3KHz) is coded into a 29 symbol TES alphabet. A detailed account of a TES system for IWR, together with an algorithm for simple TES analysis is provided in Ref. [2]. A flexible CMOS LSI embodiment is shortly to be exercised.

THE SYSTEM MODEL

A block diagram of the TES system under investigation is shown in Fig. 1. The current investigation examines isolated words whose end-points are delimited by the action of a pressel switch. A TES symbol stream is generated from the incoming waveform.

Given this sequence of symbols, for each vocabulary word, a two-dimensional "A-matrix" can be generated. This A-matrix is the feature set representation utilised for recognition purposes in this study. Typical A-matrices are shown in Fig. 2.

AIMS OF THE EXPERIMENT

This investigation was intended to provide some insight into the performance of TES systems incorporating "fail-safe" protocols, in operating conditions close to the threshold at which the recognition algorithm presented very high (> 50%) error rates with a view to indicating the relevance or otherwise of simple DVI systems in a hostile tactical military environment.

EXPERIMENTAL PROCEDURES

RECOGNITION VOCABULARY: The vocabulary exercised during this examination

¹Military Communications Research Group, Royal Military College of Science, Shrivenham, SWINDON, Wilts. SN6 8LA.

was limited to the ten digits "ZERO" to "NINE".

END-POINT DETECTION: In view of the use of an omni-directional microphone and the high noise levels under investigation, the start and finish of an utterance were defined by means of a pressel switch. Having been prompted by the system, the user would depress the switch, say the appropriate word, and then release the switch.

SUBJECT SPEAKERS: The four subject speakers used were untrained male Army Officers with no previous experience of DVI systems. All had experience of communicating over a radio link using a pressel switch. Training and familiarisation was limited to a brief verbal description of the evaluation procedures with the first "hands-on" use being the initial training pass.

ACOUSTIC ENVIRONMENT: Tests were carried out in the TES research laboratory at R.M.C.S., an unexceptional room with no special or protective acoustic properties. External acoustic background noise conditions were varied to produce SPL values at the face of the subject speakers as follows:-

| | | |
|--------------|------|-----|
| Medium noise | 82dB | SPL |
| High noise | 94dB | SPL |

A typical plot of the spectral content of both noise and signal + noise for the High noise case can be seen in Fig. 3b. The measurements were taken at the input to the TES coder, that is to say, directly after the microphone. Markers on the diagram at 300 and 4.3KHz are self-explanatory.

A hand-held omni-directional microphone (SONY ECM-170 with pop filter fitted) was used during trials, to introduce some variability in microphone placement and thus, a degree of inconsistency in acoustic coupling such as might be found in operational use.

The background noise used was "cocktail party babble", a mixture of human speech recorded from the BBC sound archives. This mimics, to some extent, the acoustic conditions in a busy forward area command post, and is a most severe test since the use of human speech as background noise effectively precludes the use of noise cancellation strategies as a pre-processing aid to the recognition algorithm. Fig. 3a. shows the frequency spectrum of this noise.

EVALUATION ROUTINES: Two sets of evaluations were taken for each of the four speakers as follows:-

Medium noise A familiarisation stage comprising one training pass followed by two evaluations of 50 random utterances of the words "ZERO" to "NINE". And an evaluation stage of 250 repeats of the same words.

High noise A familiarisation stage comprising one training pass followed by two evaluations of 50 random utterances of the words "ZERO" to "NINE". And an evaluation stage of 500 utterances of the same words.

These procedures took several hours per person, similar to a 'shift' on duty. During this period the subjects experienced boredom, minor discomfort, and fatigue. There were also some complaints of disorientation at the higher SPL level. At times the subjects made mistakes either in speaking the correct word, or in depression of the pressel switch. Errors resulting have been included in the scores.

PERFORMANCE AND EVALUATION CRITERIA

In "fail-safe" systems at high noise levels, conventional percentage recognition scores would appear to be an insufficient measure of system performance. In the present investigation a 'ranking table' of errors was produced for each evaluation set with a view to providing more insight into the performance of the system. Our first limited examination indicates that the "ranking distribution" of errors is likely to relate directly to the performance and stability of

the system when operating under adverse conditions. For convenience, this distribution may be simplified into a single "System Transaction Overhead", T_o , defined as:-

$$T_o = \frac{\text{No. of additional transactions required to achieve correct recognition}}{\text{Minimum number of transactions required}} \quad (1)$$

An example of the difference between this measure and conventional percentage recognition scores is highlighted in Fig. 4. The scores obtained are shown in Table 1a.

These scores show that even though there are twice as many errors in Fig. 4b. as there are in that of Fig. 4a., the value of T_o is lower since less exchanges are needed to achieve 100% system integrity.

RESULTS

Fig. 5a. shows the total percentage recognition scores for each of the four speakers for medium and high noise levels. Fig. 5b. illustrates, for both noise levels, the ranking distribution of errors averaged over all four speakers. Figures have been compiled from three thousand input utterances. Overall results are given in Table 1b.

CONCLUSIONS

A simple TES IWR system has been exercised in very high acoustic background noise using four untrained speakers. An examination of the ranking distribution of the errors suggests that "fail-safe" protocols may permit satisfactory system performance to be achieved using TES systems. This measure for performance and evaluation is commended for further investigation.

ACKNOWLEDGEMENTS

Our thanks are due to Mr M Williams of Bruel & Kjaer (UK) Ltd., and Mr R F Powell of R.M.C.S., for their assistance in the measurement of background noise and in the preparation of Fig. 3. The assistance of 39 and 40 Degree Course in providing the subject speakers is gratefully acknowledged.

REFERENCES

- [1] Power, R C, *et al.* Verification, Archetype Updating, and Automatic Token Set Selection, ... in High Acoustic Noise Backgrounds. **IEE. Conf. Speech Input/Output; Techniques & Applications.** pp 144-151.
- [2] Holbeche, J, *et al.* Time Encoded Speech (TES) Descriptors as a Symbol Feature Set for Voice Recognition Systems. **IEE. Conf. Speech Input/Output; Techniques & Applications.** pp 310-315

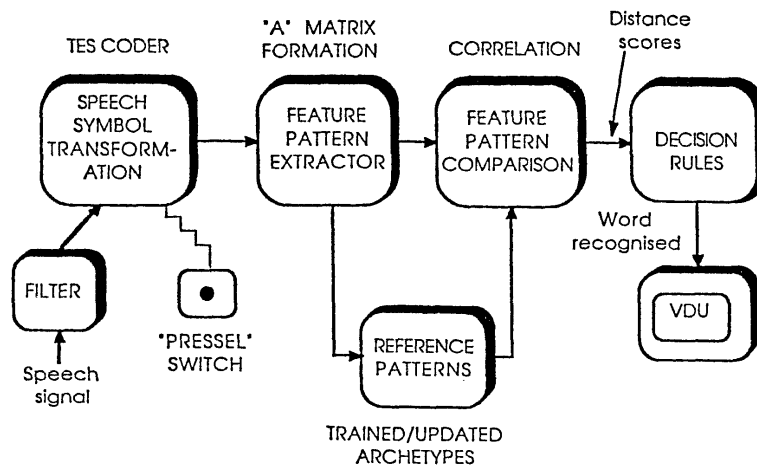


Figure 1. System block diagram.

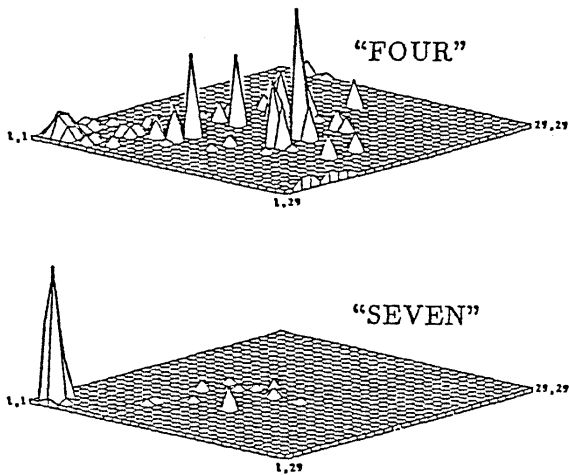


Figure 2. Typical A-matrices for the words "FOUR" and "SEVEN".

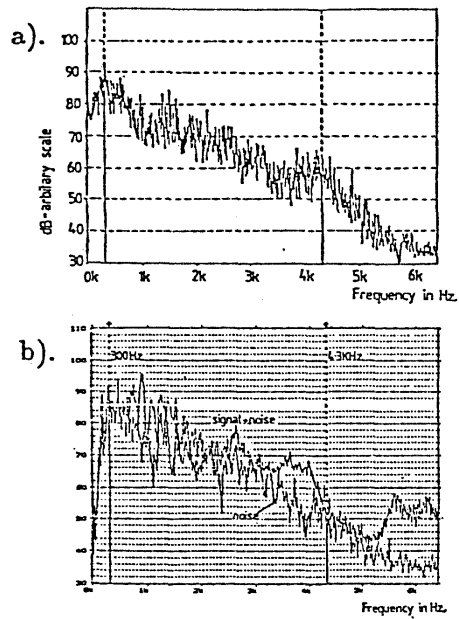


Figure 3. Typical High noise plots for noise and signal + noise.

a). RANKING OF SPOKEN WORD

| | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th | 11th |
|--|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|
| "ONE" | 5 | . | . | . | . | . | . | . | . | . | . |
| "TWO" | 5 | . | . | . | . | . | . | . | . | . | . |
| "THREE" | 5 | . | . | . | . | . | . | . | . | . | . |
| "FOUR" | 5 | . | . | . | . | . | . | . | . | . | . |
| "FIVE" | 5 | . | . | . | . | . | . | . | . | . | . |
| "SIX" | 4 | . | . | . | . | 1 | . | . | . | . | . |
| "SEVEN" | 5 | . | . | . | . | . | . | . | . | . | . |
| "EIGHT" | 4 | 1 | . | . | . | . | . | . | . | . | . |
| "NINE" | 5 | . | . | . | . | . | . | . | . | . | . |
| "ZERO" | 5 | . | . | . | . | . | . | . | . | . | . |
| RANKING TOTALS | 48 | 1 | . | . | 1 | . | . | . | . | . | . |
| MINIMUM NUMBER OF INTERACTIONS REQUIRED TO OBTAIN CORRECT WORD | . | 2 | . | . | 5 | . | . | . | . | . | . |

FOR 50 CORRECT TRANSACTIONS TO OCCUR A MINIMUM OF 67 INTERACTIONS ARE NEEDED.

b). RANKING OF SPOKEN WORD

| | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th | 9th | 10th | 11th |
|--|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|
| "ONE" | 5 | . | . | . | . | . | . | . | . | . | . |
| "TWO" | 4 | 1 | . | . | . | . | . | . | . | . | . |
| "THREE" | 4 | 1 | . | . | . | . | . | . | . | . | . |
| "FOUR" | 5 | . | . | . | . | . | . | . | . | . | . |
| "FIVE" | 5 | . | . | . | . | . | . | . | . | . | . |
| "SIX" | 5 | . | . | . | . | . | . | . | . | . | . |
| "SEVEN" | 4 | 1 | . | . | . | . | . | . | . | . | . |
| "EIGHT" | 4 | 1 | . | . | . | . | . | . | . | . | . |
| "NINE" | 5 | . | . | . | . | . | . | . | . | . | . |
| "ZERO" | 5 | . | . | . | . | . | . | . | . | . | . |
| RANKING TOTALS | 45 | 4 | . | . | . | . | . | . | . | . | . |
| MINIMUM NUMBER OF INTERACTIONS REQUIRED TO OBTAIN CORRECT WORD | . | 4 | . | . | . | . | . | . | . | . | . |

FOR 50 CORRECT TRANSACTIONS TO OCCUR A MINIMUM OF 54 INTERACTIONS ARE NEEDED.

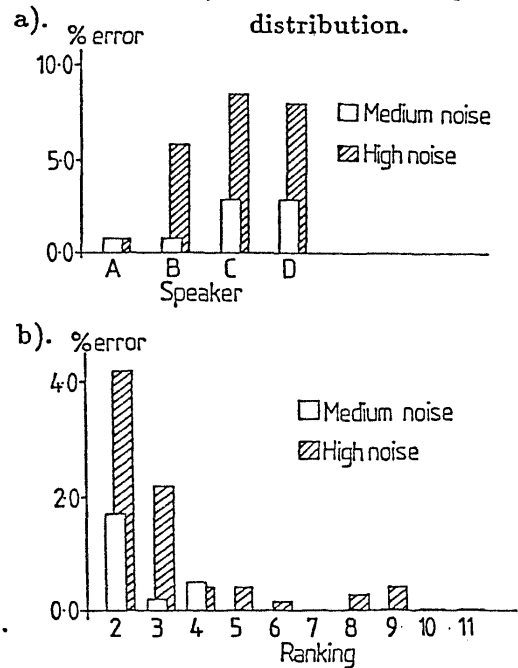
Figure 4. Illustrative "ranking table" of errors.

a).

| | % error | T_o % |
|----------|---------|---------|
| Fig. 4a. | 8.0 | 8.0 |
| Fig. 4b. | 4.0 | 14.0 |

Scores from Figure 4.

Figure 5. a). Average % error rate. b). Average ranking distribution.



b).

| | % error | T_o % |
|------------|---------|---------|
| Med. noise | 1.8 | 2.4 |
| High noise | 5.75 | 8.25 |

Scores from Figure 5b).

Table 1. Performance values.