AUTOMATIC DIAGNOSTIC AND ASSESSMENT PROCEDURES FOR THE COMPARISON AND OPTIMISATION OF TIME ENCODED SPEECH (TES) DVI SYSTEMS

V. V. Vu and R. A. King

Royal Military College of Science (Cranfield)
Shrivenham, Swindon, Wilts, ENGLAND

Abstract

The development of simple automatic diagnostic and assessment procedures for the comparison and optimisation of TES-based DVI systems is presented. The use of a "Diagnostic Matrix" [1] to assess the variability of input acoustic events purporting to be the same word and to indicate the degree of orthogonality between the acoustic events which form the vocabulary under examination, is discussed. The process of adapting the diagnostic procedures, so that system performance may be evaluated against incremental system parameter changes is described. An initial set of "Assessment procedures" used to predict likely system performance under varying noise conditions is presented. Experimental evidence is offered to support the discussion.

1 Introduction

The performance of Time Encoded Speech (TES) systems has recently reported upon and the coding process involved appears to commend itself in fields such as Isolated Word Recognition (IWR) in the tactical military arena [2, 3, 4]. Current investigations are indicating that this coding format may also show utility in the recognition of whispered speech [5] and respirator speech as well as in a variety of military application areas associated with the recognition and classification of non-speech acoustic events.

TES systems are multi-parameter and thus may be tractable to a range of optimisation procedures to cater for different acoustic targets and widely varying acoustic background conditions. This flexibility may be considered an attractive feature of TES DVI systems but it poses a requirement for the identification of an optimum feature set for each different but relatively stable operating domain, eg. normal speech, whispered speech, respirator speech, etc.

To meet this objective, a set of Diagnostic and Assessment procedures are currently under development. These, together with an in-house Acoustic Database Management System (ADMS) enable the generation of Diagnostic Matrices for subsequent analyses. The ADMS provides controlled and repeatable acoustic signals, from existing databases or from locally generated signals which have been digitally recorded on disc using 20 KHz sampling rate and 12-bit resolution A/D conversion either in a real noisy environment or in a quiet room. Instrumentation has been developed to enable the System-Under-Test (SUT) to access the database by sending commands to the ADMS over a serial link and to modify the software-based parameters of the SUT. By this means, predetermined incremental changes to specified system parameters of the SUT may be made to assess their effects on system performance. The ADMS can also output a number of different noise waveforms at specified levels and can add these, if required, to acoustic signals previously recorded in a quiet environment before output. Assessment procedures are then used automatically to measure the performance of the subject recogniser, according to a range of performance measures currently under investigation.

2 TES Processing and Recognition Algorithms

2.1 TES Coding

Time Encoded Speech (TES) is a form of speech waveform coding first proposed by King and Gosling [6]. In this approach, the speech waveform is broken into segments between successive real zeros. For each such segment of the waveform the code consists of a single digital word, derived from two parameters of the segment, its quantised time duration and its shape. For economy, the symbols produced by this process are then mapped on to a very small Alphabet of about 30 code descriptors.

As reported in [2], the simplicity of implementation and of the TES numerical coding format commends it as a vehicle for a variety of tasks in the speech recognition arena. A more detailed account of a TES system for IWR is contained in [2]. Those features relevant to the subject investigation are summarised briefly below. Some duplication will be inevitable.
2.2 Whole Word Segmentation and "A"-Matrix Formation

TES analysis is usually performed on whole word utterances, the end points of which are defined by a voice operated switch, or the manual operation of a pressel switch. The result of this analysis is a sequence of TES symbols, \( \tau(n) \), where \( n \) refers to the \( n \)th symbol of the sequence. Given this sequence, for each isolated whole word, we accumulate the following two dimensional histogram, called the A-Matrix, whose entries are defined as follows:

\[
a_{ij} = \sum_{n=1}^{N} \chi_{ij}(n)
\]

where \( N \) is the total number of symbols in the utterance.

The A-Matrix contains no information on amplitude or energy of the input acoustic signal, and also no information on positions of its segments on the time axis. The definition of \( a_{ij} \) may be modified to incorporate into the matrix adjustments for amplitude and time.

2.3 A-Matrix Compression

A-Matrices are found to be sparsely populated and an equivalent description of an in terms of its non-zero entries is used. Further, it is found convenient to limit the extent of this description by including only the highest entries of \( A \). Systems utilizing between 40 and 60 entries are currently being investigated.

Specimens of three-dimensional representations of the A-Matrices of a normal YES and a normal NO are given in figs. 1 and 2.

2.4 Archetype Formation

Given a number, \( U \), of utterances of the same word the corresponding matrices \( A(u) \) \( (u = 1, 2, \ldots, \ U) \) may be formed. From this set it is possible to define an Archetype, \( A' \), having the same form as these matrices but in some sense representative of them all.

\[
A' = F(A(u)) = F(A(1), A(2), \ldots, A(U))
\]

The role of the Archetype is similar to that of the template in conventional word recognition systems.

2.5 "A"-Matrices Comparison

A measure of similarity between an Archetype and an utterance or between two utterances is given by a correlation score. If \( A \) and \( B \) are their A-Matrices, in the current embodiment a correlation \( c(A, B) \) is computed as follows:

\[
c(A, B) = \frac{(\sum a_{ij} b_{ij})^2}{(\sum a_{ij}^2)(\sum b_{ij}^2)} \times 100
\]

VARIATIONS OF THIS MEASURE ARE ALSO USED AS AN INDICATOR OF SYSTEM PERFORMANCE.

3 The Diagnostic Matrix

For each individual utterance an A-Matrix is formed and stored during training. A-Matrices derived from the input tokens of each similar word in the training set are then merged together to form an Archetype. Finally, a Diagnostic Matrix is formed by comparing each of the Archetypes so produced and input tokens against

- The tokens from which the Archetype was derived,
- All other tokens and
- All other Archetypes.

Performance of the TES-based recognition system for any selected vocabulary may be assessed by studying Diagnostic matrices for the vocabulary concerned. System performance is analysed by comparing the words of the specimen vocabulary in pairs. The minimum (worst-case) separations of all possible pairs of comparisons are computed against variations in the system parameters of the SUT. These are then combined via a range of simple combination formulae which favourably weight similarity and magnitude, to indicate parameters likely to provide optimum performance.

A specimen Diagnostic Matrix is shown in fig. 3. This compares five different tokens of the words YES and NO normally spoken.

4 An Illustrative Examination

In this illustrative investigation speech, band-limited to 4kHz (0.3-4.3kHz), is coded into a 29 symbol TES Alphabet. For this bandwidth, and with normal speech, symbols are generated at an average rate of about 2500 per second.

For simplicity, the acoustic variability of normal versions of the words YES and NO in the TES measurement space, when subjected to incremental system parameter changes, is presented. The Diagnostic Matrix is used to examine five acoustic utterances of each word produced by a single cooperative speaker against a small
range of assessment measures. The parameters under investigation are Integration (I), a signal conditioning factor which progressively tills the frequency spectrum of the input signals towards full integration, i.e. 6dB per octave, and Amplitude Adjustment (AA), a weighting factor which preferably enhances the significance of A-Matrix data associated with voiced sounds. Both these measures provide a balance between the voiced and unvoiced sounds, in the A-Matrices and resultant Archetypes.

It may be seen from fig. 3 that the Diagnostic Matrix provides measures of separation between the "U" individual input utterance A-Matrices and the Archetypes of the word-set, in the measurement space of the TES recogniser. The Diagnostic Matrix can also provide measures of consistency/variability of the input utterances of each word within and across word boundaries, using Archetype-Token (A-T) and Token-Token (T-T) comparisons. The Diagnostic Matrix also produces correlations which compare the similarity of each Archetype against the Archetypes of all the other members from the vocabulary set. These Archetype-Archetype (A-A) scores may provide a simple guide to the confusions likely to exist between the consistent features of the individual acoustic tokens representing the words which form the vocabulary.

From the data assembled in the Diagnostic Matrix, a range of different comparison scores in the range 0 - 100 may be computed. In our current investigation the Archetype-Token Minimum Distance score A-T(mDis) is used for the comparison of word pairs, thus,

\[
A-T(mDis), \text{Word 1 vs Word 2} = \min[A(\text{Word 1}) vs T_u(\text{Word 1})]
\]

\[
- A(\text{Word 2}) vs T_u(\text{Word 1})
\]

(4)

where \( u = 1, \ldots, U \)

A-T(mDis) scores indicate the minimum (worst-case) separation existing between Archetypes and Tokens of the word pairs compared. When applied to fig. 3 it may be seen that

\[
\]

Other measures related to A-T Spread, A-T Average and A-T Deviation form part of the comparison inventory. A similar range of Token-Token (T-T) comparisons may also be used.

For the purpose of this examination, the A-T Minimum Distance score A-T(mDis) is exemplified as a performance indicator against two system parameters Integration factor "I" over the range 0.0 - 0.9 and Amplitude Adjustment factor "AA" over the range 0 - 16. The combination formula adopted for this examination is:

\[
\text{Combined A-T Min Distance Score YES vs NO} = \frac{[A-T(mDis) \text{ YES vs NO}]^2}{[A-T(mDis) \text{ NO vs YES}]^2}
\]

(5)

Results obtained from the YES and NO comparisons for five normal speech utterances are summarized and shown in figs. 4 to 6. These figures would indicate an optimum performance for parameter settings of AA = 0 over values of I ranging from 0.5 - 0.8.

5 Comment

The development of automatic Diagnostic and Assessment procedures for the comparison and optimisation of TES-based recognition systems has been presented utilizing A-Matrices and Archetypes of the words YES and NO normally spoken. A limited initial set of Assessment measures has been described and the Archetype-Token Minimum Distance score has been used to indicate optimum settings for "AA" and "I". The system parameters under investigation, for these two words. The vulnerability of these results to variations in background noise, in combination formula and to extended vocabularies is currently under investigation with a view to providing an indication of optimum system parameters for larger representative word sets and realistic operational conditions.

Acknowledgements

Thanks are due to the following: -

- The Australian DOD (Navy Office) for the support of Mr. Vu in these investigations.

- The Principal of the Royal Military College of Science (Cranfield) for permission to publish this paper.

References


---

**Figure 1**: A-Matrix of Normal YES

**Figure 2**: A-Matrix of Normal NO

**Figure 3**: Diagnostic Matrix YES vs NO - A Specimen
Figure 4: A-T Minimum Distance Scores for Normal YES vs Normal NO @ AA=0,4,8,16

Figure 5: A-T Minimum Distance Scores for Normal NO vs Normal YES @ AA=0,4,8,16

Figure 6: Combined A-T Minimum Distance Scores for Normal YES and Normal NO