COMPARISON OF VQ AND DTW CLASSIFIERS FOR SPEAKER VERIFICATION

A M Ariyaeeinia and P Sivakumaran

University of Hertfordshire, UK

ABSTRACT

An investigation into the relative speaker verification performance of various types of vector quantisation (VQ) and dynamic time warping (DTW) classifiers is presented. The study covers a number of algorithmic issues involved in the above classifiers, and examines the effects of these on the verification accuracy. The experiments are based on the use of a subset from the Brent (telephone quality) speech database. This subset consists of repetitions of isolated digit utterances 1 to 9 and zero. The paper describes the experimental work, and presents an analysis of the results.

INTRODUCTION

The success of a speaker recognition system depends, to a large extent, on the type of the adopted classifier. To date, extensive research has been conducted in order to compare various classifiers based on their levels of effectiveness and computational complexity [1-4]. The present study attempts to further this investigation by providing a thorough evaluation of a number of classifiers from two main groups of VQ and DTW. Although the classifiers in the latter group are normally expected to be more effective for text-dependent speaker verification, the VQ-based approaches have the advantage of being considerably more efficient in terms of computational cost and memory requirements. This trade-off may, to a certain extent, be overcome by using a combined VQ-DTW approach. It is thought that such a classifier, which was originally proposed and successfully used for speech recognition [5], can be equally effective for speaker verification.

For the purpose of comparison, it is therefore necessary that the relative speaker discrimination abilities of the above classifiers are evaluated accurately. As part of this study attempts are also made to cover the algorithmic issues which affect the performance of the adopted classifiers, and have not previously been considered in a comparative study of this kind.

The following sections introduce the classifiers used in this study, and give a description of the related algorithms. Details of the adopted speech database and the experimental work are also presented.

DESCRIPTION OF CLASSIFIERS

The use of the VQ and DTW techniques in automatic speaker recognition has been the subject of considerable study over several years [1,2,4-9]. The attraction of the VQ approach is due to its capability in compressing the training data efficiently. This in turn leads to a significant saving in computation. The DTW method, on the other hand, provides the possibility for handling the time-normalisation problem associated with variable speaking rates. This section briefly describes the differences in algorithms for classifiers within each category, and presents a discussion on the combined VQ-DTW approach.

A. VQ

Two types of partitioning algorithms for generating VQ codebooks are considered here. These are the standard LBG algorithm [10] and a modified version which is referred to as the distortion driven cluster splitting (DDCS) [11]. Both algorithms are initiated with the centroid of the entire population of the training feature vectors. The main difference between the two algorithms is due to the method used for cluster splitting. In the LBG algorithm, the splitting process is performed simultaneously in all existing clusters until a desired number of clusters is reached. In the modified version, only the cluster with the largest total distortion is split. The process in this case continues until either the mean distortion of all the clusters fall below a convergence threshold or a predefined maximum number of iterations is reached. The advantage of this second algorithm is that it eliminates the problem caused by splitting clusters with too small a distortion.

B. DTW

It has been shown that the DTW algorithm [12] performs best when the ratio of the length of the reference template to that of the test template approaches unity. This performance is found to deteriorate considerably as the above ratio approaches 1/2 or 2. In order to maximise the DTW effectiveness, a linear normalisation technique is adopted which ensures that the above mentioned ratio is always equal to unity [11,12].

The calculation of the degree of dissimilarity between a given test utterance and the reference set of the proposed speaker is carried out using two different methods. The first method is based on time-aligning the reference utterances and then averaging these to obtain a single reference model. This model is then compared against the test utterance to obtain a match score. Due to the use of a single combined reference model (CRM), the complete classifier in this case is referred to as DTW-CRM. The second method involves computing the distances between the test utterance and all the training repetitions of the utterance. The lowest K distances are then averaged to generate the final match score. In this case, the classifier is termed DTW-K.

C. VQ-DTW

This method involves generating a VQ codebook using all the utterance repetitions in the reference set. Each feature vector in the training utterances is replaced with the closest vector in the codebook, and then a matrix of codevector distances is formed. i.e.

$$D(i,j) = d(\mathbf{c}_i, \mathbf{c}_j), \quad 1 \le i \le Q, \quad 1 \le j \le Q \tag{1}$$

where $D(i,j)$ are the distance matrix terms, $d(\mathbf{c}_i,\mathbf{c}_j)$ is the distance between the i^{th} and j^{th} codevectors, and Q is the size of the codebook. In the verification phase, feature vectors in the test utterance are replaced with their closest codevectors. As a result the calculation of distances required in the DTW procedure reduces to a table look-up procedure. i.e. the distance between a test vector represented by the codevector $\mathbf{c}_{i'}$, and a reference vector represented by the codevector $\mathbf{c}_{j'}$ is simply $D(i', j')$.

The main disadvantage of this approach is that of the approximation of the feature vectors to their nearest codevectors. In order to reduce the resultant degradation in the classifier performance an alternative algorithm is adopted [5]. This involves computing and storing the distances between each test feature vector and the nearest codevector. The method eliminates the need for the generation and storage of the matrix of codevector distances during the training phase. It, however, requires producing a look-up table with the following elements for each test.

$$D'(m,q) = d(\mathbf{y}_m, \mathbf{c}_q), \quad 1 \le m \le M, \quad 1 \le q \le Q \tag{2}$$

where $\mathbf{y}_m$ is the m^{th} test feature vector, and M is the number of vectors in the given test utterance. This table is used to acquire the distances needed in the subsequent DTW procedure. The final distance in each test is calculated as the average of the K smallest distances between the test utterance and the reference models of the proposed speaker.

SPEECH DATABASE AND ANALYSIS

The speech data used in the experimental study was a subset of the Brent database [11,13] consisting of 47 repetitions of isolated digit utterances 1 to 9 and zero. The subset was collected from telephone calls made from various locations by 11 male and 9 female speakers. For each speaker, the first 3 utterance repetitions (recorded in a single call) formed the training set. The remaining 44 repetitions (1 recorded per week) were used for testing.

The utterances, which had a sample rate of 8 kHz and a bandwidth of 3.1 kHz, were pre-emphasised using a first order digital filter. These were segmented using 25 ms Hamming windows at 12.5 ms intervals, and then subjected to a 12th-order linear prediction analysis. The resultant linear predictive coding (LPC) parameters for each frame were appropriately analysed using a 10th-order fast Fourier transform, a filter bank, and a discrete cosine transform to extract a 12th-order mel-frequency cepstral feature vector [11,14,15]. The filterbank used for this purpose consisted of 20 filters. The centre frequencies of the first 10 filters were linearly spaced up to 1 kHz, and the other 10 were logarithmically spaced over the remaining frequency range (up to 4 kHz).

In order to minimise the performance degradation due to the linear filtering effect of the telephone channel, a cepstral mean normalisation approach was adopted. The technique involved computing the average cepstral feature vector across the whole utterance, and then subtracting this from individual feature vectors [13].

EXPERIMENTS

The first part of the experimental work was carried out using single digit utterances to compare the relative performance of DDCS-VQ and LBG-VQ. For the latter classifier a codebook size of 32 was selected, as a set of preliminary studies indicated this to be the optimum size for the given utterance durations. For the DDCS-VQ, the convergence threshold was chosen such that its average codebook size also became very close to 32. The distance measure used in these and other experiments described in this paper was a weighted Euclidean metric [11]. The results of the above comparative study are given in Figure 1. These clearly show that DDCS-VQ performs better than LBG-VQ in respect of equal error rate (EER), and therefore confirm that the clustering algorithm used in the former classifier is more effective.

As the next part of the investigation, speaker verification experiments were conducted using DTW-CRM and DTW-K. In the case of the latter classifier, K

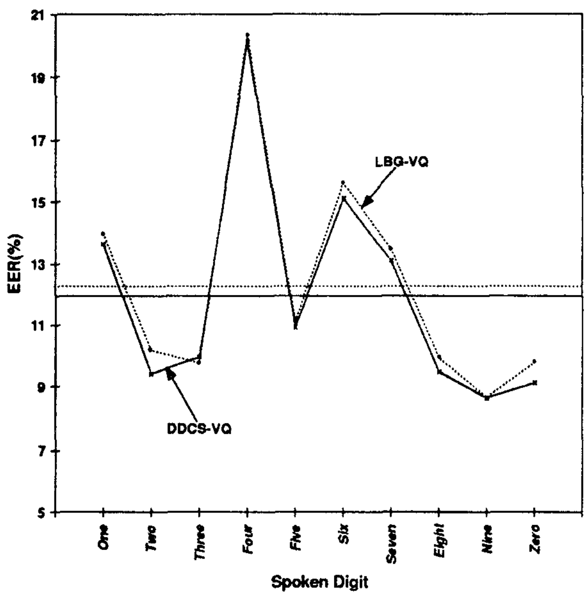

Figure 1. Experimental results for LBG-VQ and DDCS-VQ classifiers.
Horizontal lines represent the average EERs.

was set to 2. Figure 2 gives the results of these experiments together with those obtained earlier for DDCS-VQ.

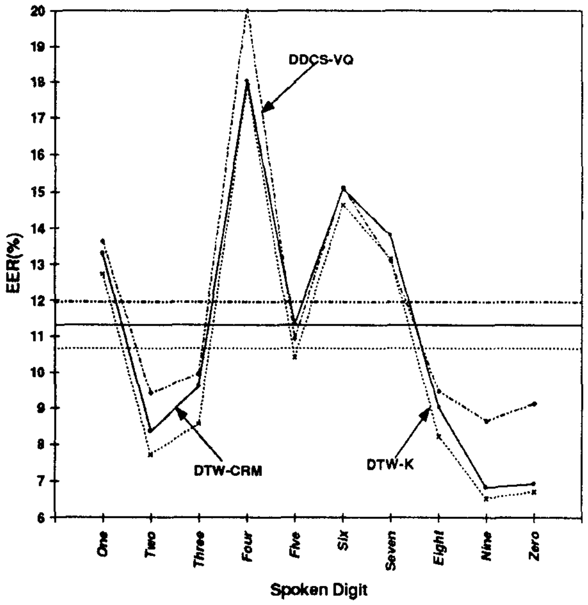

Figure 2. Experimental results for DDCS-VQ and DTW classifiers.

It is observed that, as expected, DTW classifiers have considerably better discrimination abilities than DDCS-VQ. The results also indicate that, with DTW, the use of multiple reference models is more effective than a single combined reference model. This is despite the fact that the training utterances for each speaker were collected in a single call. It is thought that if the training utterances were taken through different telephone calls over a period of time, then the difference in performance of the two types of DTW classifiers would further widen in favour of DTW-K. This is because the training sessions which are considerably different from testing (e.g. in terms of recording conditions, telephone channel characteristics, and speaking behaviour) may seriously corrupt a combined model.

The experimental studies also included an examination of the performance of VQ-DTW which was effectively a combination of DDCS-VQ and DTW-K. Figure 3 compares the EERs obtained for this classifier with those for the other two types of DTW-based classifiers. It is interesting to note that the performance of VQ-DTW is consistently better than that of DTW-CRM. Although, the VQ-DTW appears to be slightly less effective than DTW-K, it is computationally far more efficient.

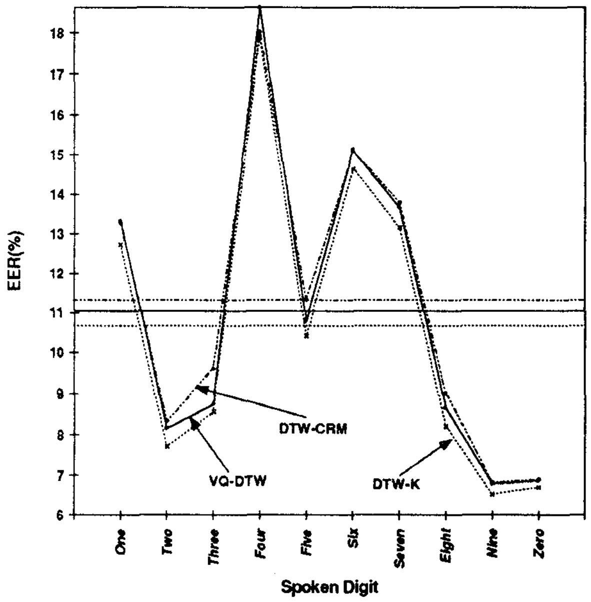

Figure 3. EERs for VQ-DTW and other DTW-based classifiers.

The final part of the investigation consisted of experiments using a sequence of 10 digits (i.e. 1 to 9 and zero). The results of this study together with those obtained for single digits are summarised in Table 1. It is seen that the order of effectiveness of the classifiers is not affected by the duration of the spoken material. The DTW-K and LBG-VQ classifiers are found to be the best and the worst performers respectively. The excellent performance of the former classifier is closely followed by that of VQ-DTW.

Classifier	*EER Based on a Sequence of 10 Digits*	*Average EER Based on Single Digits*
LBG-VQ	4.72%	12.28%
DDCS-VQ	4.34%	11.95%
DTW-CRM	3.53%	11.35%
VQ-DTW	3.40%	11.05%
DTW-K	3.15%	10.66%

Table 1. Summary of the experimental results.

CONCLUSIONS

In this study, the relative effectiveness of a number of classifiers (i.e. LBG-VQ, DDCS-VQ, DTW-CRM, DTW-K, VQ-DTW) for text-dependent speaker verification has been investigated. For the purpose of the experiments a subset from the Brent (telephone quality) speech database was adopted. This subset consisted of repetitions of isolated digit utterances 1 to 9 and zero.

It has been found that, in this group of classifiers, DTW-K and LGB-VQ are the best and the worst performers respectively. It has also been shown that a level of performance close to that of DTW-K can be achieved by using the VQ-DTW classifier which is computationally more efficient. The experimental studies have further indicated that the DDCS-VQ classifier performs relatively better than LBG-VQ.

Based on the experimental results it has been shown that a minimum average EER of about 10.66% can be achieved for single digits. This error rate is found to reach 3.15% when a combination of all 10 digits is used.

ACKNOWLEDGEMENT

The authors would like to express their thanks to Mr. Mark Pawlewski and Mr. Simon Downey of BT Labs for their support and very useful discussions.

REFERENCES

1. T. Matsui and S. Furui, 1992, "Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's", Proc. ICASSP, 2, 157-160

2. D. A. Irvine and F. J. Owens, 1993, "A Comparison of Speaker Recognition Techniques for Telephone Speech", Proc. Eurospeech, 3, 2275-227

3. K. R. Farrell, R. J. Mammone and K. T. Assaleh, 1992, "Speaker recognition using neural networks and conventional classifiers", IEEE Trans. on Speech and Audio Processing, 194-205

4. K. Yu, J. Mason and J. Oglesby 1995, "Speaker Recognition Models", Proc. Eurospeech, 629-632

5. K. Shikano, 1982, "Spoken Word Recognition Based Upon Vector Quantization of Input Speech", Trans. Comm. Speech Res., 473-480

6. S. Furui, 1981, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-29, 254-272

7. F. Soong, A. Rosenberg, L. Rabiner and B. Juang, 1985, "A Vector Quantization Approach to Speaker Recognition", Proc. ICASSP, 1, 387-390

8. F. K. Soong, and A. E. Rosenberg, 1988, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-36, 871-879

9. I. Booth, M. Barlow and B. Watson, 1993, "Enhancements to DTW and VQ Decision Algorithms for Speaker Verification", Speech Comm, 13, 127-133

10. Y. Linde, A. Buzo and R. M. Gray, 1980, "An Algorithm for Vector Quantizer Design", IEEE Trans. on Comm., COM-28, 84-95

11. P. Sivakumaran and A. M. Ariyaeeinia, 1996, "Effectiveness of Various Types of DTW and VQ Classifiers for Text-Dependent Speaker Verification", Technical Report for BT Laboratories

12. C. Myers, L. R. Rabinar and A. E. Rosenberg, 1980, "Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition", IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-28, 623-634

13. M. Pawlewski, B. P. Milner, S. A. Hovell, D. G. Ollason, S. P. A. Ringland, K. J. Power, S. N. Downey and J. Bridges, 1996, "Advances in Telephony Based Speech Recognition", BT Technology Journal, 127-150

14. S. B. Davis and P. Mermelstein, 1980, "Comparison of parametric representation for monosyllabic word recognition in continuously

spoken sentences", IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-28, 357-366

15. R. Deller, J. G. Proakis and H. L. Hansen, 1993, "Discrete-Time Processing of Speech Signals", Macmillan Inc., New York