ENHANCEMENTS TO DTW AND VQ DECISION ALGORITHMS FOR SPEAKER RECOGNITION

Ian Booth, Michael Barlow, and Brett Watson

Speaker Verification Group
Department of Electrical Engineering
University of Queensland

ABSTRACT - Dynamic Time Warping (DTW) and Vector Quantisation (VQ) techniques have been applied with considerable success to speaker verification. In this paper we develop two enhancements involving statistical weighting and distance normalisation. Speaker verification results on a population of 42 are reported.

INTRODUCTION

Automatic speaker recognition, the technique of recognising people by their voices, has application in a number of areas:- access control, telephone banking, and forensic speaker identification. Considerable research has been carried out in the field over the last thirty years (Smith, 1962) and a number of different techniques have been explored.

Two techniques which have been shown to be very useful for speaker and speech recognition are Dynamic Time Warping (DTW; Doddington, 1971) and Vector Quantisation (VQ; Gersho & Gray, 1992). Considerable work over the years has shown that both DTW (Doddington, 1971; Doddington, 1985; Barlow, 1991) and VQ (Soong et. al., 1985; Matsui & Furui, 1992) are highly efficient decision algorithms for speaker recognition.

Recently Barlow (1991) has shown that the warp path calculated during the dynamic time warping process encodes time-dependent information useful for speaker recognition. This paper furthers the previous work by applying statistically derived linear weighting to combine the warp path parameters with the DTW distance score. Similarly linear weighting is applied to the distortion scores between input frames and individual VQ codebook entries to determine possible improvements in the technique.

A further enhancement to both techniques for speaker verification, in which output scores for the claimed identity are "likelihood normalised" against scores for other identities (Higgins et. al., 1991), is reported upon.

EXPERIMENTAL DATA

A common database of isolated digits (Barlow, Booth, & Parr, 1992) was used in all experiments; thus ensuring ease of comparison between results. A total of 42 adult speakers of English were recorded uttering the isolated digits (zero to nine) a number of times over a period of 2 months. A subset of eleven (11) young adult male speakers were designated "customers" and twenty one (21) instances of each digit from each speaker were used for training and testing. The remaining thirty one (31) speakers were designated "impostors" and three (3) instances of each digit from each speaker were used to test the system.

Speakers were recorded individually in a quiet room using a close talking microphone directly onto digital audio tape. Speech samples were down-sampled to 16 kHz and quantised at 16 bits. A 10'th order cepstral analysis with a frame size of 32ms and a frame shift of 8ms was carried out, with 1'st and 2'nd order delta cepstrals also being extracted (Furui, 1981).

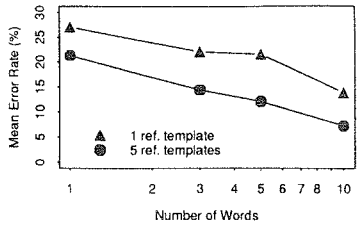

Figure 1: DTW speaker verification performance. Performance is plotted as a function of number of words used when 5 reference templates per word are used versus a single composite utterance per word.

For all experiments mean error rates are given when one, three, five and ten isolated digits are used as inputs. In order to reduce the number of combinations when less than ten digits are used as input the digits zero, three, five, seven, and nine alone have been used; with the others excluded. Only in the 10 digit case are all ten digits included.

METHOD

The following subsections report baseline system performance for both the DTW technique and the VQ distortion technique. In all cases text-dependent speaker verification results are reported as mean error rates. For DTW, system performance as a function of number of reference templates is examined, while VQ experiments examine the impact of varying codebook size.

Dynamic Time Warping

A staggered array dynamic time warping algorithm (Furui, 1986) with equal weighting applied to the three possible warp transitions (see Figure 3) was employed to provide a baseline measure of DTW speaker verification performance.

The 21 repetitions of each digit from the members of the customer set were split into a reference-threshold set of 15 repetitions and a test set of 6 repetitions. Figure 1 shows the results of speaker verification trials for the dataset under two conditions. Under one scheme five (5) separate reference templates are stored for each word occurrence. DTW comparisons are performed between all 5 reference templates and the input speech; the minimum DTW distance score of the 5 being selected for decision making. The second scheme uses a single composite reference utterance derived from the five original reference utterances. As can be seen from the figure the performance obtained using five individual reference templates is significantly superior to that when using a single composite template.

Vector Quantisation

A vector quantisation distortion algorithm (Soong et. al., 1985), in which codebooks for each customer are built and the accumulated VQ distortion (difference between closest codebook element and input frame) used to make verification systems has been implemented as a baseline system. Codebooks of varying sizes are constructed using nine (9) repetitions of each of the ten digits; resulting in a codebook for each speaker composed from 90 utterances.

Figure 2 shows the results of speaker verification trials, using a single word, as the codebook size is varied from 10 to 100 elements in steps of 10. As can be seen from the figure verification performance initially improves as codebook size increases. However above a codebook size of 50 verification performance degrades. Such a drop in performance is attributable to lack of training samples for the codebook size. A codebook size of 50 was fixed upon for all following experiments. A similar experiment was conducted in which codebooks were

constructed for individual words for each speaker. A similar relationship between verification performance and codebook size was found, though verification results were found to vary significantly between individual words.

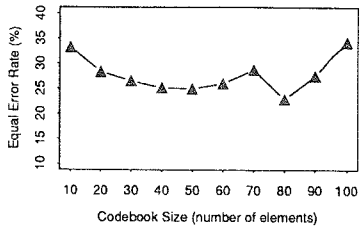

Figure 2: VQ speaker verification performance as a function of codebook size.

STATISTICAL ANALYSIS OF PARAMETERS

The conventional DTW and VQ-distortion algorithms produce a single scalar output as the result of the comparison between a reference and input template. Recently Barlow (1991) has shown that additional parameters of the DTW algorithms may be extracted such that total speaker recognition performance is improved if the additional parameters are utilised.

The following section reports on variants to both the DTW and VQ algorithms in which a "vector" of values, rather than a single scalar, is produced as the result of each comparison between reference and test templates. Discriminant analysis (Hays & Winkler, 1971) is applied to the resulting vectors to determine a weighting to apply to each vector element so as to yield optimal separation between inter-speaker and intra-speaker vectors. A separate set of weightings are derived for each word from each speaker in the customer set. In this study the weights are derived based on a priori knowledge of the identity of the speaker and applied to the data on which the statistical analysis was performed. Hence the results indicate the additional speaker specific information encoded in the data, as insufficient data made statistically significant speaker verification trials impossible.

Dynamic Time Warping

As part of the dynamic time warping algorithm a warp path, the path of best fit between the two templates being compared, is computed. Figure 3 is one representation of the DTW algorithm showing the warp path as computed between two templates. Barlow (1991) has shown that the warp path encodes time-varying information regarding the identity of a speaker and may be used to enhance the performance of a conventional DTW algorithm.

An experiment was conducted in which a total of 13 parameters of the DTW comparison:- the DTW distance and 12 additional parameters related to the warp path, were extracted. A linear weighting derived from the application of discriminant analysis was applied to the vector result of each DTW comparison. The resulting value was then thresholded via the normal method. Table 1 presents the results for the conventional DTW algorithm and the enhanced scheme of 13 parameters with linear weighting. These results clearly suggest that significant improvement in speaker verification performance is possible with the addition of statistical linear weighting.

Vector Quantisation

Recognising that different codebook elements will encode different levels of speaker specific information a scheme was derived to apply weightings to the distortion score for individual codebook elements. Rather than obtain a single mean VQ distortion value for all codebook elements a mean distortion value was kept

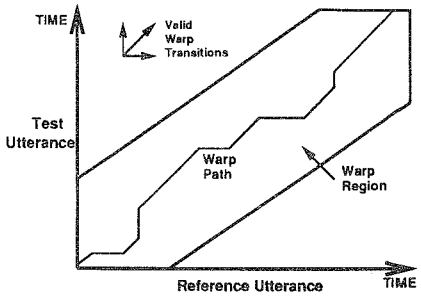

Figure 3: Schematic of Dynamic Time Warping Process showing warp path calculated between reference and test utterance to obtain minimum distance.

Type	*Number of Words*			
	1	3	5	10
Standard	21.3	14.4	12.0	7.2
Weighted	13.4	6.5	4.3	2.4

Table 1: Dynamic Time Warping speaker verification mean error rates contrasting standard DTW against a linearly weighted scheme.

for each individual codebook element, the mean for a codebook element being updated only when that codebook element was found to be the closest match to the input frame. Hence, rather than a single scalar value being obtained as the result of a VQ comparison a vector of 50 (the number of codebook elements) was obtained.

A linear weighting derived from the application of discriminant analysis was applied to the vector resulting from each VQ comparison. Table 2 shows the results following discriminant analysis of the additional VQ parameters. These result clearly suggest that major performance improvements may be possible using a statistical weighting of codebook elements.

Type	*Number of Words*			
	1	3	5	10
Standard	17.3	10.0	6.6	3.3
Weighted	3.2	0.6	0.1	0.0

Table 2: Vector Quantisation speaker verification mean error rates contrasting standard VQ-distortion against a linearly weighted scheme.

DISTANCE NORMALISATION

A recent proposal to improve speaker verification performance is to apply likelihood or distance normalisation (Higgins et. al., 1991). Rather than solely make the comparison between input speech and references templates for the claimed identity; the input speech is also compared with the reference templates for a number of other speakers. Instead of applying a "hard threshold" the score obtained for the claimed identity is contrasted with the score obtained for the other members of the speaker population.

For the following VQ and DTW results; distance normalisation was applied by determining the three smallest distances from the customer set besides the claimed identity. The mean of these three distances was then subtracted from the distance for the claimed identity and the result thresholded. Equation 1 expresses the derivation of the normalised distance for a claimed identity c.

$$d_c' = d_c - \frac{d_i + d_j + d_k}{3}$$
$$\text{s.t.} \quad d_a \leq d_b, a \in \{i,j,k\}, b \ni \{i,j,k,c\} \tag{1}$$

Dynamic Time Warping

Figure 4 plots DTW based speaker verification performance for both un-normalised and normalised distance schemes. As can be seen from the figure a normalised distance DTW scheme significantly improves the speaker verification performance as compared to a baseline system.

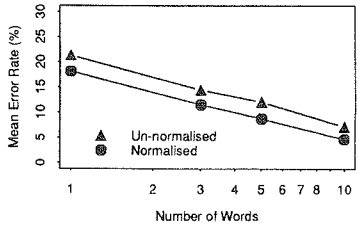

Figure 4: DTW speaker verification performance for standard un-normalised DTW distance and normalised against the 3 closest from the remaining customer set.

Vector Quantisation

Figure 5 plots VQ based speaker performance for both un-normalised and normalised distance schemes. As can been seen from the figure the normalised scheme is significantly superior to un-normalised, particularly when fewer words are used as input.

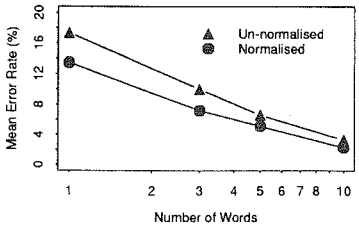

Figure 5: VQ speaker verification performance for standard un-normalised VQ distortion and normalised against the 3 closest from the remaining customer set.

A further series of experiments were conducted in which the number of additional speakers used to normalise the score obtained for the claimed identity was varied. Speaker verification performance improved with increasing normalisation set size though above 3 speakers the improvements were marginal.

CONCLUSION

Baseline DTW and VQ-distortion speaker verification systems were implemented for a speaker population of 42 uttering the isolated digits. Investigation of the reference template strategy for DTW found that using

five separate templates was significantly superior to using a single composite template and that choosing the five reference templates to span the period of sampling was superior to using the first five utterances. VQ codebooks of varying sizes trained using all ten digits were examined for speaker verification performance. It was found that verification performance improved with increasing codebook size, to a point, before again degrading. A codebook size of 50 was found to be the best compromise between codebook size and verification performance.

Two novel enhancements to the DTW and VQ techniques were examined. For DTW properties of the warp path were extracted in addition to the DTW distance. For VQ, rather than a single mean distortion figure, a set of 50 mean distortion values, one per codebook element, was produced. Statistical discriminant analysis was applied to the DTW and VQ derived vectors. In both cases it was found that the additional parameters encoded further speaker specific information and hence could be used to improve speaker verification performance.

A distance normalisation technique, in which the output of the decision algorithm for the claimed identity was compared against the closest three outputs for the remaining speaker population, was also examined. For both DTW and VQ algorithms the method of distance normalisation was found to significantly improve speaker verification performance.

It may be seen from the results that a number of significant improvements are possible to both the DTW and VQ techniques as applied to speaker verification. Additional areas for improvement are currently being examined, in particular using Artificial Neural Networks (ANNs) to perform non-linear weightings for the vector outputs (in place of the statistical technique currently used) and further refinements to the method of distance normalisation. Techniques to reduce the amount of training data for both techniques are also being investigated.

REFERENCES

Barlow, M. (1991) *Prosodic Acoustic Correlates of Speaker Characteristics*, PhD Thesis, The University of NSW.

Barlow, M., Booth, I., & Parr, A. (1992) *The Collection of Two Speaker Recognition Targeted Speech Databases*, Proc. Aust. Int. Conf. Speech Science and Technology.

Doddington, G.R. (1971) *A Method of Speaker verification*, PhD Thesis, The University of Wisconsin.

Doddington, G.R. (1985) *Speaker Recognition- Identifying People by their Voices*, Proc. IEEE, 73(11), 1651-1664.

Furui, S. (1981) *Cepstral Analysis Technique for Automatic Speaker Verification*, IEEE Trans. ASSP, 29(2), 254-272.

Furui, S. (1986) *Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectra*, IEEE Trans. ASSP, vol 34, 52-59.

Gersho, A., & Gray, R.M. (1992) *Vector Quantisation and Signal Compression*, Kluwer Academic Publishers, Boston.

Hays, W.L., & Winkler R.L. (1971) *Statistics: Probability, Inference and Decision*, Holt, Rinehart and Winston Inc., New York.

Higgins, A., Bahler, L., & Porter, J. (1991) *Speaker Verification Using Randomized Phrase Prompting*, Digital Signal Processing, 1, 89-106.

Matsui, T., & Furui, S. (1992) *Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMMs*, Proc. ICASSP 92, Volume 2, 157-160.

Smith, J.E.K. (1962) *Decision-Theoretic Speaker Recognizer*, J. Acoust. Soc. Am., 34, 1968.

Soong, F., Rosenberg, A., Rabiner, L., & Juang, B. (1985) *A Vector Quantisation Approach to Speaker Recognition*, Proc. ICASSP 85, 387-390.