AUTOMATIC SPEAKER RECOGNITION USING MSVQ-CODED SPEECH

J Leis[†] , M Phythian[†] and S Sridharan[‡]

[†]University of Southern Queensland
Faculty of Engineering

[‡]Queensland University of Technology
Signal Processing Research Centre

ABSTRACT - Low bitrate speech coding finds application in both telecommunications (bandwidth compression) and archival (filespace compression). Speaker verification is used in telecommunication applications (to gain access to particular services, for example) and implies that either or both of the speech data streams (incoming and reference) may be compressed. In this paper, we investigate the effect of high compression methods on the effectiveness of automatic speaker identification and verification. Lossy compression of the speech (whether transmitted or stored) requires vector quantization of the short-term spectral parameters in order to achieve high compression ratios, and thus implies some loss of accuracy in the representation of these parameters. However, in the situation where the same spectral parameters are utilized in identifying the speaker, the identification accuracy may be compromised by the compression process. We present in this paper our findings on the effect of compression on identification, for one particular family of vector quantization methods.

PROBLEM FORMULATION

In considering the evaluation of the effect of spectrum compression on speaker identification, four possible scenarios arise as shown in Table 1. These are :-

(i) The "benchmark" for all cases, using "raw" speech in the identification process. No compression is performed.

(ii) The speech database is compressed (for example, on CD-ROM) and the incoming speech is available in uncompressed form.

(iii) The incoming speech is compressed, but the reference is not. This arises in telecommunications applications. Note that in this case the speaker identification parameters may be pre-computed and stored (depending on the identification algorithm), allowing the speech database to be compressed.

(iv) Both the existing database and the incoming speech are compressed.

Case (ii) is studied in this paper, and is illustrated in Figure 1. This situation arises in forensic speech processing where the database of suspects has been archived and a new suspect is to be compared.

It is assumed that the distance D_{coded} is available, and the distance $D_{uncoded}$ is *not* available. A Vector Quantization (VQ) scheme is designed for the speech spectral parameters, and two methods of speaker identification are examined: the Mahanolobis distance and the log-probability derived from a Multivariate

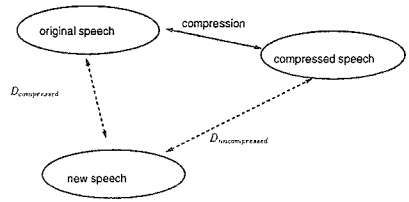

Figure 1: Compressed speaker identification scenario.

Table 1: Compression and Speaker Identification.

Condition	*Speech Database*	*Incoming Speech*
(i)	16-bit PCM	16-bit PCM
(ii)	VQ Compression	PCM
(iii)	PCM	VQ Compression
(iv)	VQ Compression	VQ Compression

Gaussian Mixture Model (GMM). Two families of VQ method which are known to achieve high compression are studied: multistage VQ and split VQ.

VECTOR QUANTIZATION

The coding method examined in this work involves Vector Quantization (VQ) of the Line Spectral Frequency (LSF) parameters obtained from the short-term analysis of the speech every 20 milliseconds. This method produces very large compression of the short-term spectral information, at the expense of a far more complex vector coding operation and increased distortion. The coding of the LSF's is examined in more detail in (Paliwal & Atal 1993). The operation of vector quantization may be divided into two distinct steps. The first of these, the *training phase*, requires a knowledge of the joint statistics of the vector parameter set to be coded. In practice, this is normally done via a training database consisting of a large number of representative codevectors. The second phase, the *coding phase*, may be further subdivided into the encoding operation and the decoding operation. The encoding operation requires a search of the vector codebook for each vector to be encoded to find the minimum error vector. The codebook index of this vector is then transmitted. The decoder on the other hand has a significantly less complex task: to look up the vector index it has received in the local codebook. The codebook design must be sufficiently robust against all possible permutations of the input vector to ensure adequate coverage of the vector space (Collura & Tremain 1993).

Direct vector quantization of the LSF parameter space is known to be unsatisfactory. Before proceeding to the identification phase, we must choose a suitable VQ method that yields accaptable performance in the spectral distortion sense.

Split VQ (SVQ) is illustrated in Figure 2. This method splits the LSF parameters into smaller sub-vectors, each with its' own sub-codebook. Multistage VQ (MSVQ) is illustrated in Figure 3. This method involves several successive VQ codebooks, each encoding the residual of the previous stage(s). MSVQ is utilized as an integral component of the MELP codec (McCree, Truong, George, Barnwell & Viswanathan 1996).

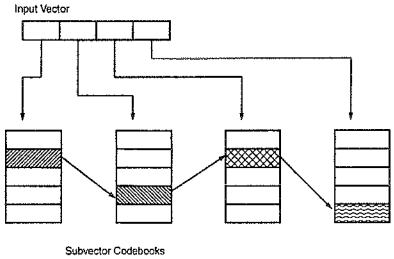

Figure 2: Split Vector Quantization (SVQ)

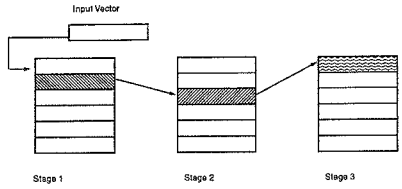

Figure 3: Multistage Vector Quantization (MSVQ)

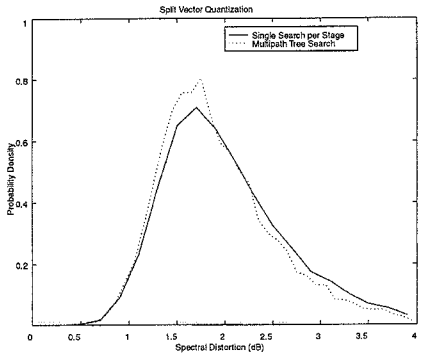

Figure 4: Distribution of spectral distortion in Split Vector Quantization (SVQ) using single and tree-structured multiway search algorithms.

SPEAKER IDENTIFICATION

Speaker identification involves the identification of a speaker from the voice alone (Gish & Schmidt 1994), (Furui 1994). Several measures of distance have been proposed in the literature. In this study, we have utilized two quite different approaches. The first is the Mahanolobis Distance Metric (MDM) $D_m(\vec{x}_t) : t = 1, \ldots, T$, which is easily computed from any m-dimensional vector parameter set $\vec{x}$ (Parsons 1987). Previous work suggests that the line spectral frequencies give superior identification accuracy using the MDM metric, when compared to the LPC coefficients.

The second approach utilized is the Gaussian Mixture Model (GMM). This approach involves the modelling of the sequence of vectors $\vec{x}$ as a mixture of multivariate Gaussian probability density functions. This approach is somewhat more complex than the MDM, but has been shown to provide superior identification results on clean speech (Reynolds & Rose 1995).

Table 2: 24 bits per frame VQ methods studied.

Method	*Parameters*
Multistage VQ	3 stages, 256 codevectors per stage
Tree-Searched Multistage VQ	As above, 2-way branch per codebook
Split VQ	Input vector split 2,2,3,3 with 64-vector codebooks per subvector
Tree-Searched Split VQ	As above, 2-way branch per codebook

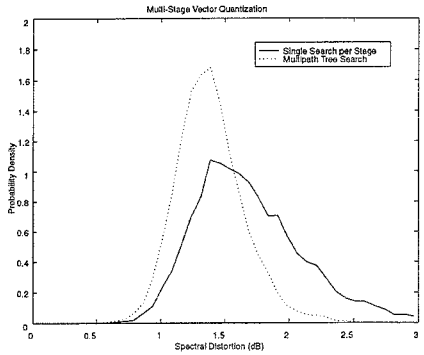

Figure 5: Distribution of spectral distortion in Multistage Vector Quantization (MSVQ) using single and tree-structured multiway search algorithms.

Table 3: Average spectral distortion for VQ methods.

Method	*Spectral Distortion, dB*
Multistage VQ	1.6970
Tree-Searched Multistage VQ	1.3863
Split VQ	2.0378
Tree-Searched Split VQ	1.9451

RESULTS

Results were obtained using Region 2 of the TIMIT speech corpus (Linguistic Data Corporation 1990) and the MSVQ compression algorithm. Figure 6 shows that the effect of compression on the calculated Mahanolobis distances is significant, and that the effect is to reduce the apparent values after compression. This is indicated by the appearance on the scatter plot of the points below the $45°$ line. The relative spread is indicated by the relative width of the ellipses enclosing the points in the vertical and horizontal directions.

Figure 7 indicates that the effect of compression on the log-probability of the set using the Gaussian Mixture Model is negligible. The values are still clustered around the $45°$ line after compression.

The mean of the Mahanolobis values (Table 4) is decreased in approximately the same proportion in each case, moving from 3.06 to 2.98 in the same-speaker case, and 3.98 to 3.81 in the different speaker case. The means of the Gaussian model values (Table 4) are changed slightly after compression for both the same-speaker and different-speaker cases. The change in the same-speaker case is negligible, however the change in the different-speaker case is an increase from 5.12 (uncompressed) to 5.19 (compressed), thus making the speakers "appear" more similar. However, the change is so small as to be negligible.

Table 4: Mean distance metrics for Mahanolobis *left* and Gaussian *right*.

Mahanolobis	Same	Different
Compressed	2.9779	3.8133
Uncompressed	3.0637	3.9777

Gaussian Model	Same	Different
Compressed	7.7955	5.1908
Uncompressed	7.8061	5.1224

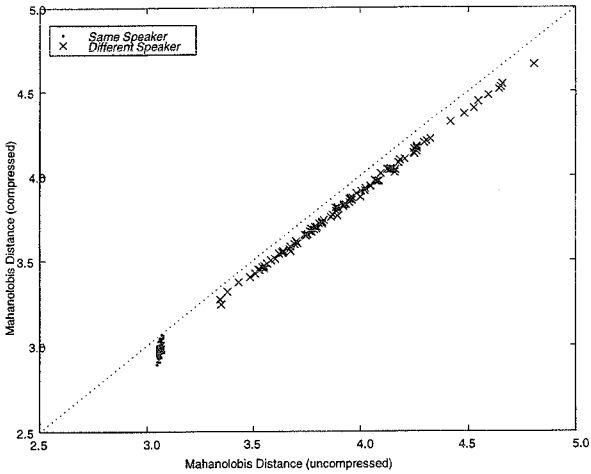

Figure 6: Compressed speaker identification using MSVQ compression and Mahanolobis distance metric.

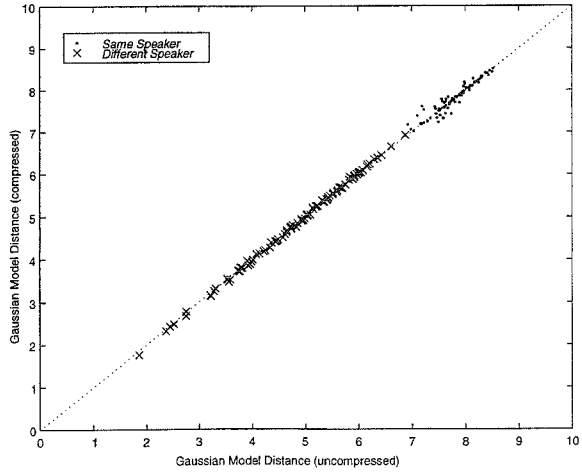

Figure 7: Compressed speaker identification using MSVQ compression and Gaussian Model distance metric.

CONCLUSIONS

We have studied the application of speaker identification/verification methods to compressed speech. It was expected that the process of compression would lead to reduced performance of the identification algorithm. We have demonstrated that this is indeed the case for the low-complexity Mahanolobis distance metric calculation, but that a modelling method using Gaussian mixtures is substantially more robust to the compression process. Further work is needed to determine whether this robustness is dependent upon the number of mixtures used in the modelling process.

REFERENCES

Collura, J. S. & Tremain, T. E. (1993), 'Vector Quantizer Design for the Coding of LSP Parameters', *Proc. ICASSP'93* pp. II29–II32.

Furui, S. (1994), An Overview of Speaker Recognition Technolog, *in* 'Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification', pp. 1–9.

Gish, H. & Schmidt, M. (1994), 'Text-Independent Speaker Identification', *IEEE Signal Processing* **11**(4), 18–31.

Linguistic Data Corporation (1990), *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, National Institute of Standards and Technology.

McCree, A., Truong, K., George, E. B., Barnwell, T. P. & Viswanathan, V. (1996), 'A 2.4 Kbit/s MELP Coder Candidate for the New U.S. Federal Standard', *Proc. ICASSP'96* pp. 200–203.

Paliwal, K. K. & Atal, B. S. (1993), 'Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame', *IEEE Transactions on Speech and Audio Processing* **1**(1), 3–14.

Parsons, T. (1987), *Voice and Speech Processing*, McGraw-Hill, chapter 6 - Recognition: Features and Distances.

Reynolds, D. A. & Rose, R. C. (1995), 'Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models', *IEEE Transactions on Speech and Audio Processing* **3**(1), 72–83.