Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

ELSEVIER

Speech Communication 47 (2005) 265–276

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Multi-frame GMM-based block quantisation of line spectral frequencies

Stephen So, Kuldip K. Paliwal *

School of Microelectronic Engineering, Griffith University, Nathan Campus, Brisbane, QLD 4111, Australia

Received 30 April 2004; received in revised form 21 January 2005; accepted 15 February 2005

Abstract

In this paper, we investigate the use of the Gaussian mixture model-based block quantiser for coding line spectral frequencies that uses multiple frames and mean squared error as the quantiser selection criterion. As a viable alternative to vector quantisers, the GMM-based block quantiser encompasses both low computational and memory requirements as well as bitrate scalability. Jointly quantising multiple frames allows the exploitation of correlation across successive frames which leads to more efficient block quantisation. The efficiency gained from joint quantisation permits the use of the mean squared error distortion criterion for cluster quantiser selection, rather than the computationally expensive spectral distortion. The distortion performance gains come at the cost of an increase in computational complexity and memory. Experiments on narrowband speech from the TIMIT database demonstrate that the multi-frame GMM-based block quantiser can achieve a spectral distortion of 1 dB at 22 bits/frame, or 21 bits/frame with some added complexity.

Keywords: Speech coding; LSF coding; Transform coding; Block quantisation; Gaussian mixture models

1. Introduction

Linear predictive coding (LPC) of speech requires the accurate quantisation of parameters representing the spectral envelope. Speech is windowed into frames and the spectral envelope is parametrically modelled as an all-pole filter, whose coefficients are called linear predictive coding (LPC) parameters. These LPC parameters are generally quantised in terms of line spectral frequencies (LSFs) using a vector quantiser (VQ). Extrapolating from the operating curve of full search VQ suggests that we need about 19 bits/frame to achieve transparent coding of these parameters (Paliwal and Kleijn, 1995), while high

* Corresponding author. Tel.: +61 7 3875 3754/6536; fax: +61 7 3875 5198.

E-mail addresses: s.so@griffith.edu.au (S. So), k.paliwal@griffith.edu.au (K.K. Paliwal).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.02.007

rate analysis predicts a lower bound of 23 bits/frame[1] (Hedelin and Skoglund, 2000). It is not possible to design codebooks at these rates and in addition, the computational cost of the resulting full search vector quantiser is very high.

Less complex but suboptimal vector quantisers such as multistage and split VQ have been investigated in the speech coding literature (LeBlanc et al., 1993; Paliwal and Atal, 1993), where it was generally observed that 22–24 bits/frame were required to achieve *transparent coding*[2] in speech, with varying degrees of complexity. Further gains in performance can be achieved by exploiting temporal correlation between successive frames. Matrix quantisation (Tsao and Gray, 1985) and its derivatives such as split matrix quantisation (Xydeas and Papanastasiou, 1999) and multi-mode matrix quantisation (Nurminen et al., 2003; Sinervo et al., 2003) perform better by jointly quantising LSF frames.

The use of Gaussian mixture models (GMM) for the coding of LSFs has been investigated in (Hedelin and Skoglund, 2000; Shabestary and Hedelin, 2002; Subramaniam and Rao, 2000, 2001, 2003). In (Subramaniam and Rao, 2003), a Gaussian mixture model (GMM) is used to parameterise the probability density function (PDF) of the source and optimised Gaussian block quantisers are designed for each cluster (or, mixture component).[3] Using this quantiser in its fixed rate mode, a spectral distortion of approximately 1 dB was achieved at 24 bits/frame. The main advantages of this scheme over vector quantisers include (Subramaniam and Rao, 2003):

1. lower complexity through the use of block quantisers;
2. bitrate scalability; and
3. search complexity and memory requirements being independent of the rate of the system.

A modified quantiser with memory was also described in (Subramaniam and Rao, 2003) that coded the difference between successive frames, similar to differential pulse code modulation (DPCM) with a one-tap predictor. A spectral distortion of 1 dB was achieved at 22 bits/frame (Subramaniam and Rao, 2003). During the coding process, there is frequent use of the spectral distortion (SD) calculation for cluster quantiser selection. While there are approximate high-rate expressions for the spectral distortion calculation (Gardner and Rao, 1995), the number of computations is still comparatively higher than mean squared error (MSE).

In this paper, we investigate a modified version of the fixed-rate GMM-based block quantiser that operates on multiple frames and uses the mean squared error (MSE) distortion criterion.[4] We have found this system to perform better than the single frame as well as predictive quantiser of (Subramaniam and Rao, 2003) in terms of spectral distortion.

The organisation of this paper is as follows. Section 2 introduces some preliminaries such as the line spectral frequency representation of LPC parameters and distortion measures that are commonly used in speech coding. In Section 3, we describe the operation of the multi-frame GMM-based block quantiser as well as its computational and memory requirements. Section 4 details the LPC analysis method and speech database that we have used to evaluate the performance of the quantiser. Following this is a discussion of the performance of the multi-frame GMM-based block quantiser and how it compares with other quantisation schemes. Finally we conclude in Section 6.

[1] This is the lower bound for full-band spectral distortion (0–4 kHz) while for partial-band (0–3 kHz), the bound is 22 bits/frame (Hedelin and Skoglund, 2000).

[2] Transparent coding means that the coded speech is indistinguishable from the original through listening. An objective measure of quality is the *spectral distortion* (SD), which is defined as the root-mean-square difference between the log-spectra of the coded and original speech. Coded speech is generally accepted as being transparent when the average SD is about 1 dB (Paliwal and Atal, 1993).

[3] Therefore, we refer to this quantisation scheme as a GMM-based block quantiser.

[4] This paper is an extended version of (Paliwal and So, 2004) and contains more comparative results.

2. Preliminaries

2.1. LSF representation of LPC coefficients

In the LPC analysis of speech, a short segment of speech is assumed to be the output of an all-pole filter, $H(z) = \frac{1}{A(z)}$, driven by white Gaussian noise, where $A(z)$ is the inverse filter given by (Paliwal and Atal, 1993):

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n} \tag{1}$$

Here, n is the order of LPC analysis and $\{a_i\}_{i=1}^{n}$ are the LPC coefficients. Because $H(z)$ is used to reconstruct speech in linear predictive speech coders, its stability is of utmost importance and cannot be ensured when LPC coefficients are coded directly. Many representations of LPC coefficients have been proposed in the literature that are more robust, in terms of filter stability. These include the reflection coefficients (RC) or partial autocorrelation coefficients (PARCOR) (Itakura and Saito, 1969), arc-sine reflection coefficients (ASRC) (Gray and Markel, 1976), and log area ratios (LAR) (Viswanathan and Makhoul, 1975). The line spectral frequency (LSF) representation, proposed by Itakura in (Itakura, 1975), has been shown in the literature to be superior to other representations for speech coding (Soong and Juang, 1984; Sugamura and Itakura, 1986; Paliwal and Atal, 1993).

The line spectral frequencies are defined as the roots of the following polynomials:

$$P(z) = A(z) + z^{-(n+1)} A(z^{-1}) \tag{2}$$

and

$$Q(z) = A(z) - z^{-(n+1)} A(z^{-1}) \tag{3}$$

These two polynomials, $P(z)$ and $Q(z)$, are parametric models of the acoustic tube in two extremal states, where the $(n + 1)$th stage (representing the glottis) is either completely closed or completely opened, respectively (Sugamura and Itakura, 1986). Consequently, LSFs have the following properties (Soong and Juang, 1984):

1. All zeros of $P(z)$ and $Q(z)$ lie on the unit circle;
2. zeros of $P(z)$ and $Q(z)$ are interlaced with each other; and
3. the minimum phase property of $A(z)$ is easily preserved after quantisation of the LSFs if the first two properties are satisfied.

Because of property 1, where each zero effectively represents a particular frequency (since it has no bandwidth), clusters of two to three LSFs define the location and bandwidth of formants in the power spectrum (Paliwal and Atal, 1993; Sugamura and Itakura, 1986). Quantisation errors in the LSFs result in localised distortion in the power spectrum (Paliwal and Atal, 1993).

2.2. Distortion measures for LPC parameters

2.2.1. Spectral distortion

In order to objectively measure the distortion between a coded and uncoded LPC parameter vector, the spectral distortion is often used. For the ith frame, the spectral distortion (in dB), $d_{sd}(i)$, is defined as

$$d_{sd}(i) = \sqrt{\frac{1}{F_s} \int_0^{F_s} [P_i(f) - \widehat{P}_i(f)]^2 \, \mathrm{d}f} \tag{4}$$

where F_s is the sampling frequency and $P_i(f)$ and $\widehat{P}_i(f)$ are the LPC power spectra (in dB) of the coded and uncoded ith frame, given by

$$P_i(f) = -20\log_{10}|A_i(\mathrm{e}^{j2\pi f/F_s})| \tag{5}$$

and

$$\widehat{P}_i(f) = -20\log_{10}|\widehat{A}_i(\mathrm{e}^{j2\pi f/F_s})| \tag{6}$$

where $A_i(z)$ and $\widehat{A}_i(z)$ are the original and quantised LPC polynomials of the ith frame respectively (Paliwal and Atal, 1993).

2.2.2. Mean squared error

The mean squared error distortion, $d_{mse}(\boldsymbol{f},\hat{\boldsymbol{f}})$, between the original vector, $\boldsymbol{f}$, and the approximated vector, $\hat{\boldsymbol{f}}$, is given by

$$d_{mse}(\boldsymbol{f},\hat{\boldsymbol{f}}) = \frac{1}{n} \sum_{i=1}^{n} (f_i - \hat{f}_i)^2 \tag{7}$$

where n is the vector dimension, f_i and $\hat{f}_i$ are the ith LSF in the original and approximated vector, respectively.

3. Multi-frame GMM-based block quantisation

The multi-frame GMM-based block quantiser is based on the memoryless version proposed by Subramaniam and Rao (2003) for the coding of speech line spectral frequencies (LSF), where a Gaussian mixture model (GMM) is used to parametrically model the probability density function (PDF) of the source and block quantisers are then designed for each Gaussian mixture component (or, cluster). This modified scheme exploits interframe correlation by concatenating p successive frames into a larger vector. For example, if the length of each individual frame is n, then the extended vectors will have a dimension of $N = np$.

$$\begin{aligned}&\left[x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}\right]^{\mathrm{T}} + \left[x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}\right]^{\mathrm{T}} \\ &\quad + \cdots + \left[x_1^{(p)}, x_2^{(p)}, \ldots, x_n^{(p)}\right]^{\mathrm{T}} \\ &\quad \Rightarrow \left[x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}, x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}, \ldots, x_1^{(p)}, \right. \\ &\qquad \left. x_2^{(p)}, \ldots, x_n^{(p)}\right]^{\mathrm{T}} \end{aligned} \tag{8}$$

These extended vectors are then processed by the GMM-based block quantiser, as per usual.

In the following sections, we provide a description of the training and encoding phase of the multi-frame GMM-based block quantiser. For more details on the GMM-based block quantisation algorithm, the reader should refer to (Subramaniam and Rao, 2003).

3.1. Training phase

The PDF model, which is in the form of a Gaussian mixture model (GMM), is initialised by applying the Linde–Buzo–Gray (LBG) algorithm (Linde et al., 1980) on the training vectors where m clusters[5] are produced, each represented by a mean, $\boldsymbol{\mu}$, a covariance matrix, $\boldsymbol{\Sigma}$, and cluster weight, c. These form the initial parameters for the GMM estimation procedure. The Expectation Maximisation (EM) algorithm (Dempster et al., 1977) is performed, where the maximum likelihood estimate of the parametric model is computed iteratively until the log likelihood converges.

An eigenvalue decomposition is calculated for each of the covariance matrices, producing m sets of eigenvalues, $\{\boldsymbol{\lambda}_i\}_{i=1}^m$, where $\boldsymbol{\lambda}_i = \{\lambda_{i,j}\}_{j=1}^N$, and m sets of eigenvectors, $\{\boldsymbol{v}_i\}_{i=1}^m$, where $\boldsymbol{v}_i = \{\boldsymbol{v}_{i,j}\}_{j=1}^N$. The ith set of eigenvectors form the rows of the orthogonal transformation matrix, $\boldsymbol{P}_i$, which will be used for the Karhunen–Loève transform in the encoding phase.

3.2. Encoding phase

In the encoding phase of the multi-frame GMM-based block quantiser, the bit allocation is initially determined, given the fixed target bitrate, and vectors are then encoded using minimum distortion block quantisation. These two aspects of the encoding phase, which use the GMM parameters and KLT matrices from the training phase, are described in the following subsections.

3.2.1. Bit allocation

If the target bitrate of the p-frame multi-frame GMM-based block quantiser is b bits/frame, the total number of bits, b_{tot}, that are available for coding each extended vector will be equal to pb. These bits need to be divided among the m cluster block quantisers. The number of bits, b_i, allocated to the block quantiser of cluster i, is given by (Subramaniam and Rao, 2003):

$$2^{b_i} = 2^{b_{\mathrm{tot}}} \frac{(c_i \Lambda_i)^{\frac{N}{N+2}}}{\sum_{k=1}^m (c_k \Lambda_k)^{\frac{N}{N+2}}}, \quad \text{for } i = 1, 2, \ldots, m \tag{9}$$

where (Subramaniam and Rao, 2003):

$$\Lambda_i = \left(\prod_{j=1}^N \lambda_{i,j} \right)^{\frac{1}{N}} \quad \text{for } i = 1, 2, \ldots, m \tag{10}$$

and $\lambda_{i,j}$ is the jth eigenvalue of the ith cluster. Then for each block quantiser, the high resolution formula from (Huang and Schultheiss, 1963) is used to distribute the b_i bits to each of the vector components:

[5] The terms 'cluster' and 'mixture component' are used interchangeably in this paper.

$$b_{i,j} = \frac{b_i}{N} + \frac{1}{2}\log_2 \frac{\lambda_{i,j}}{\left(\prod_{j=1}^{N}\lambda_{i,j}\right)^{\frac{1}{N}}}$$
$$\text{for } i = 1,2,\ldots,m \text{ and } j = 1,2,\ldots,N \tag{11}$$

3.2.2. Minimum distortion block quantisation

Fig. 1 shows a diagram of minimum distortion block quantisation, which uses an analysis-by-synthesis approach. At first glance, it can be seen to consist of m independent Gaussian block quantisers, Q_i, operating on N dimensional vectors. Each cluster block quantiser has its own orthogonal matrix, $\boldsymbol{P}_i$, which was calculated from the training phase, and bit allocation, $\{b_{i,j}\}_{j=1}^{N}$. Because the vectors comprise of multiple frames that are concatenated together, the KLT of each block quantiser will allow the exploitation of correlation between components across successive frames, as well as within each frame.

To quantise a vector, $\boldsymbol{x}$, using a particular cluster i, the cluster mean vector, $\boldsymbol{\mu}_i$, is first subtracted and its components decorrelated using the orthogonal matrix, $\boldsymbol{P}_i$, for that cluster. The variance of each component is then normalised to produce a decorrelated, mean-subtracted, and variance-normalised vector, $\boldsymbol{z}_i$:

$$\boldsymbol{z}_i = \frac{\boldsymbol{P}_i(\boldsymbol{x} - \boldsymbol{\mu}_i)}{\boldsymbol{\sigma}_i} \tag{12}$$

where $\boldsymbol{\sigma}_i = \boldsymbol{\lambda}_i^{\frac{1}{2}}$ is the standard deviation vector of the ith cluster. These are then quantised using a set of n Gaussian Lloyd-Max scalar quantisers as described in (Huang and Schultheiss, 1963) with their respective bit allocations producing indices, $\boldsymbol{q}_i$. They are then decoded to give the approximated normalised vector, $\hat{\boldsymbol{z}}_i$, which is multiplied by the standard deviation and correlated again by multiplying with the transpose, $\boldsymbol{P}_i^{\mathrm{T}}$, of the orthogonal matrix. The cluster mean is then added back to give the reconstructed vector, $\hat{\boldsymbol{x}}_i$.

$$\hat{\boldsymbol{x}}_i = \boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{\sigma}_i\hat{\boldsymbol{z}}_i + \boldsymbol{\mu}_i \tag{13}$$

The distortion between this reconstructed vector and original is then calculated, $d(\boldsymbol{x} - \hat{\boldsymbol{x}}_i)$.

The above procedure is performed for all clusters in the system and the cluster, k, which gives the *minimum distortion* is chosen:

$$k = \operatorname*{argmin}_i d(\boldsymbol{x} - \hat{\boldsymbol{x}}_i) \tag{14}$$

An appropriate distortion measure can be chosen for this stage, depending upon the application. In this paper, which is focused on LSF quantisation for speech coding, mean squared error (MSE) and spectral distortion (SD) will be investigated, with emphasis on comparing their relative trade-offs between quantisation performance and complexity.

Because of the use of independent scalar quantisers in the KLT domain, the order of the LSFs within each frame cannot be guaranteed, and this can compromise the stability of the LPC synthesis filter. A simple way of dealing with this is to add a stability check to the minimum distortion criterion. That is, we select the cluster block quantiser that produces a vector which is both stable and has minimum distortion. This procedure, though, is not sufficient as it is possible for all cluster block quantisers to produce unstable frames. However, in our experiments, we have found this situation to be quite rare.

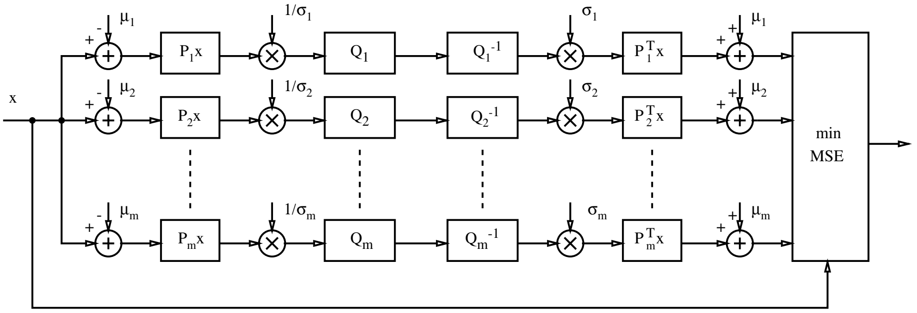

Fig. 1. Minimum distortion block quantisation (Q—block quantiser).

3.2.3. Quantiser index encoding

Each quantised vector has associated with it, a number identifying which cluster was used for coding. As proposed in (Subramaniam and Rao, 2003), this side information can be made inherent in the encoding. For an m cluster system operating at b bits per vector, $\log_2 m$ bits are required to uniquely identify each cluster. Therefore, on average, $b - \log_2 m$ bits are available for quantising each vector which is equivalent to $2^b/m$ quantiser levels. Hence, our range of quantiser indices has been partitioned into m segments.

In effect, this partitioning of the range of quantiser levels allows the cluster number to be found by determining which partition the block code belongs to. An example of this encoding scheme is shown in Fig. 2 where a total of 3 bits are available to encode each block and the system uses 2 clusters. Cluster 1 has been assigned 5 levels while cluster 2 has the remaining 3 levels. If cluster 1 was deemed the most suitable for encoding the current block, the binary codes that are available for quantiser index encoding would be contained within the range of 000–100. Any binary code outside this range, such as 101, would be identified as being a quantiser index belonging to cluster 2. Therefore, the decoder can easily determine which cluster the block belongs to by working out which partition the code falls into. Hence this removes the need for extra side information to be transmitted.

3.3. Bitrate scalability, computational complexity and memory requirements

The GMM parameters and Karhunen–Loève transform (KLT) matrices are the only static and bitrate-independent parameters of the multi-frame GMM-based block quantiser. These only need to be calculated once during training and stored at the encoder and decoder. The bit allocation formulas (described in Section 3.2.1) are closed-form expressions, hence they can be evaluated 'on-the-fly' for different bitrates using the static PDF model, by both encoder and decoder. Therefore, this scheme is bitrate scalable, where the bitrate can be changed without quantiser re-training by the encoder and decoder (Subramaniam and Rao, 2003). Vector quantisers, on the other hand, are not bitrate scalable since the codebook is designed only for a specific bitrate.

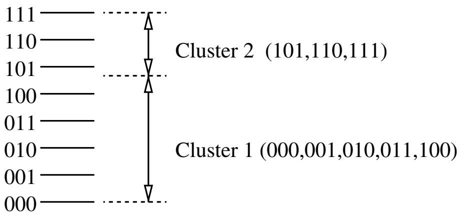

Fig. 2. Example of quantiser level encoding and cluster number partitioning.

As described by Subramaniam and Rao (2003), one of the salient features of the GMM-based block quantiser is the independence of computational complexity and memory requirements to the bitrate. This contrasts with the unconstrained vector quantiser, whose codebook, and therefore storage requirements as well as search complexity, grows exponentially with the number of bits.

Table 1 shows the complexity of each operation of the m cluster, p-frame multi-frame GMM-based block quantiser, based on a similar table in (Subramaniam and Rao, 2003). Rather than using a non-uniform scalar quantiser, where quantiser levels need to be stored, a uniform scalar quantiser with appropriate companding and expanding

Table 1
Bitrate independent computational complexity of the p-frame multi-frame GMM-based block quantiser (after Subramaniam and Rao, 2003)

Operation	Complexity (flops)
Mean subtraction	mN
Decorrelation	$m(2N^2 - N)$
Scaling	mN
Compander + rounding + expander	$mN(2n_{CE} + 2)$
Rescaling	mN
Correlation	$m(2N^2 - N)$
Mean addition	mN
Distortion calculation	mn_{dist}
Final comparison	m
Total	$2mN(2 + 2N + n_{CE}) + mn_{\text{dist}} + m$
Average complexity (flops/frame)	$[2mN(2 + 2N + n_{CE}) + mn_{\text{dist}} + m]/p$

functions is a fast and efficient alternative (Subramaniam and Rao, 2003).

It can be observed from Table 1 that the complexity of the scheme is dependent only on the number of clusters, m, the dimension of the vectors, N, the number of concatenated vectors, p, and the complexity of the distortion measure used, n_{dist}. The difference between this table and Table 1 in (Subramaniam and Rao, 2003) is the complexity of the distortion measure to be used. For the mean squared error criterion, $n_{\text{dist}} = 3N$ flops/frame. For the spectral distortion criterion, $n_{\text{dist}} \approx 15.265$ kflops/frame.[6] The memory requirement of the multi-frame GMM-based block quantiser, as given in (Subramaniam and Rao, 2003), is $2^{n_{CE}+1} + m(N^2 + 3N)$ floats.

The computational complexity (in kflops/frame)[7] and memory requirements of the multi-frame GMM-based block quantiser for LSF coding are given in Table 2 for cluster sizes of 16 and 32. From this table, it can be seen that concatenating more frames to exploit the correlation leads to an increase in computational and memory requirements. This is because as p is increased, the dimension of the extended vectors, N, becomes larger. Also, we can see that the computational complexity of the multi-frame GMM-based block quantiser is considerably higher when using spectral distortion to select the cluster block quantiser than when using mean squared error. We will keep this mind when comparing the quantisation performance between SD and MSE criterion in Section 5.3.

Table 2
Bitrate independent computational complexity (in kflops/frame) and memory requirements (ROM) of the multi-frame GMM-based block quantiser as a function of number of concatenated vectors, p, and number of clusters, m, for MSE and SD criterion

m	p	kflops/frame		ROM (floats)
		MSE	SD	
16	1	10.09	253.9	2336
	2	16.49	276.3	7616
	3	22.89	311.5	16,096
	4	29.28	359.5	27,776
32	1	20.19	507.7	4416
	2	32.98	552.5	14,976
	3	45.77	622.9	31,936
	4	58.57	718.9	55,296

4. Experimental setup

The TIMIT database was used to train and test the various quantisation schemes. It consists of speech down-sampled to 8 kHz with a 3.4 kHz anti-aliasing filter applied. A 20 ms Hamming window is used and a tenth order linear predictive analysis is performed on each frame using the autocorrelation method (Paliwal and Kleijn, 1995). There is no overlap between successive speech frames. High frequency compensation and bandwidth expansion of 15 Hz[8] was used to correct the effects of the anti-aliasing filter (Atal and Schroeder, 1979) as well as formant underestimation, respectively (Kroon and Kleijn, 1995). The training data consists of 707,438 vectors while the evaluation set, which is exclusive of the training, contains 85,353 vectors.

5. Results and discussion

5.1. Spectral distortion performance of the 16 cluster multi-frame GMM-based block quantiser

Table 3 shows the spectral distortion performance of the 16 cluster, multi-frame GMM-based

[6] The spectral distortion involves calculating the root-mean-squared-error between the log power spectrum of both the original and quantised LPC coefficients. Therefore, two 256-point split-radix FFTs are used, each expending 6664 flops, according to Table 6.2 of (Proakis and Manolakis, 1996). The reciprocal of the squared magnitude is calculated for each FFT, requiring an additional 516 flops, which followed by the decibel calculation, brings the total computations to 14.876 kflops. The RMS calculation adds an extra 389 flops. The number of flops required for the square root and logarithm calculation are not known and are assumed to be 1 flop, which brings the total computational complexity of the spectral distortion calculation to approximately 15.27 kflops.

[7] In this study, each addition, multiplication, and comparison is considered one floating point operation (flop).

[8] This is the same high frequency compensation and bandwidth expansion contained in the source code of the US Federal Standard 1016 4.8 kbps CELP coder described in (Campbell et al., 1989).

Table 3
Spectral distortion of the 16 cluster multi-frame GMM-based block quantiser using MSE criterion as a function of bitrate and number of concatenated frames, *p*

p	bits/frame	Avg. SD (dB)	Outliers (in %)	
			2–4 dB	>4 dB
2	21	1.112	2.67	0.01
	22	1.047	1.84	0.00
	23	0.987	1.25	0.00
	24	0.929	0.86	0.00
	25	0.875	0.62	0.00
3	21	1.063	2.14	0.01
	22	1.001	1.42	0.00
	23	0.941	0.98	0.00
	24	0.887	0.66	0.00
	25	0.836	0.48	0.00
4	21	1.042	1.88	0.01
	22	0.982	1.33	0.00
	23	0.926	0.94	0.00
	24	0.871	0.57	0.00
	25	0.821	0.42	0.00

Table 4
Spectral distortion of the 32 cluster multi-frame GMM-based block quantiser using MSE criterion as a function of bitrate and number of concatenated frames, *p*

p	bits/frame	Avg. SD (dB)	Outliers (in %)	
			2–4 dB	>4 dB
2	21	1.069	1.97	0.00
	22	1.008	1.39	0.00
	23	0.950	0.94	0.00
	24	0.895	0.61	0.00
	25	0.842	0.39	0.00
3	20	1.084	2.49	0.01
	21	1.022	1.69	0.00
	22	0.963	1.11	0.00
	23	0.908	0.74	0.00
	24	0.855	0.51	0.00
	25	0.804	0.34	0.00
4	20	1.066	2.12	0.01
	21	1.006	1.40	0.00
	22	0.949	0.94	0.00
	23	0.894	0.64	0.00
	24	0.841	0.43	0.00
	25	0.792	0.25	0.00

block quantiser for varying bitrates and number of concatenated frames, *p*. A spectral distortion of 1 dB has been achieved at 22 bits/frame with $p = 3$. For any given bitrate, the spectral distortion decreases as more frames are concatenated together. This may be attributed to the decorrelation of LSFs within and across frames by the KLT. Because the dimension of the vectors is larger, the block quantiser can operate at a higher bitrate and therefore, the performance is expected to improve. As well as spectral distortion, the percentage of outlier frames has also decreased with an increase in *p*.

5.2. Spectral distortion performance of the 32 cluster multi-frame GMM-based block quantiser

Table 4 shows that when using 32 clusters, only 21 bits/frame are required for a spectral distortion of 1 dB. It can also be seen that the percentage of outliers has decreased as a result of using more clusters. This is to be expected as there are more cluster quantisers to choose from. However, the computational complexity and memory requirements of the 32 cluster scheme (Table 2) are much higher than those of the 16 cluster one. Hence from an implementation standpoint, a saving of 1 bit/frame may not justify the increase in complexity that accompanies the use of more clusters.

5.3. Comparison with the multi-frame GMM-based block quantiser using SD-based quantiser selection

Table 5 shows the spectral distortion performance of the multi-frame GMM-based block quantiser that uses spectral distortion[9] for determining the best cluster quantiser, as is done in the original memoryless GMM-based block quantiser of (Subramaniam and Rao, 2003). For each concatenated frame, the spectral distortion is calculated for each subframe and these are added together to give the final distortion. Because spectral distortion is used as the final objective measure, matching the distortion criteria is expected to be more optimal than a mismatched scheme (in our case, MSE with SD). This is shown when comparing the spectral distortions of Tables 3 and 5. The matched case yields spectral distortions which are

[9] We used the full-band spectral distortion calculation, given by Eq. (4), rather than the approximation from (Gardner and Rao, 1995).

Table 5
Spectral distortion of the 16 cluster multi-frame GMM-based block quantiser using SD criterion as a function of bitrate and number of concatenated frames, *p*

p	bits/frame	Avg. SD (dB)	Outliers (in %)	
			2–4 dB	>4 dB
2	21	1.091	2.13	0.00
	22	1.027	1.41	0.00
	23	0.967	0.92	0.00
	24	0.911	0.63	0.00
	25	0.858	0.42	0.00
3	21	1.046	1.84	0.01
	22	0.985	1.18	0.00
	23	0.925	0.79	0.00
	24	0.872	0.50	0.00
	25	0.823	0.36	0.00
4	21	1.027	1.63	0.00
	22	0.968	1.10	0.00
	23	0.912	0.72	0.00
	24	0.858	0.44	0.00
	25	0.810	0.32	0.00

roughly 0.02 dB lower than the MSE-based scheme. Results for a 32 cluster SD-based scheme, given in Table 6, show the SD-based scheme to be superior to the MSE-based one by approximately 0.02 dB. It would appear that the improvement in quantisation performance, as a result of using an SD-based scheme rather than an MSE-based one, is less than 1 bit/frame.

The advantage of using MSE over SD is the computational simplicity of the former over the latter, as we have seen in Table 2. For example, the 16 cluster multi-frame GMM-based block quantiser with $p = 3$, at 24 bits/frame requires only 22.89 kflops/frame for the MSE criterion, compared with 311.5 kflops/frame for the SD criterion, though the difference in average spectral distortion between the two schemes is only 0.015 dB. In fact, according to Tables 2 and 4, a slightly lower spectral distortion can be achieved using 32 clusters and the MSE criterion at 24 bits/frame, while requiring only 15% of the complexity of the scheme using 16 clusters and the SD criterion (45.77 cf. 311.5 kflops/frame).

Therefore, the significant reduction in computational complexity, as a result of using the MSE criterion, more than outweighs the minor gains in quantisation performance from using the spectral distortion criterion, though one may substitute the spectral distortion calculation with a simpler high-rate approximation from (Gardner and Rao, 1995).

Table 6
Spectral distortion of the 32 cluster multi-frame GMM-based block quantiser using SD criterion as a function of bitrate and number of concatenated frames, *p*

p	bits/frame	Avg. SD (dB)	Outliers (in %)	
			2–4 dB	>4 dB
2	21	1.045	1.52	0.00
	22	0.985	1.05	0.00
	23	0.928	0.66	0.00
	24	0.874	0.43	0.00
	25	0.822	0.26	0.00
3	21	1.003	1.37	0.01
	22	0.945	0.90	0.00
	23	0.891	0.56	0.00
	24	0.838	0.37	0.00
	25	0.788	0.25	0.00
4	21	0.989	1.15	0.00
	22	0.932	0.73	0.00
	23	0.879	0.47	0.00
	24	0.826	0.31	0.00
	25	0.779	0.18	0.00

5.4. Comparison with the memoryless GMM-based block quantiser

Table 7 shows the results for the memoryless ($p = 1$) GMM-based block quantiser using 16 clusters, as used in (Subramaniam and Rao, 2003). The memoryless version uses spectral distortion (SD) to select the appropriate block quantiser.

Table 7
Spectral distortion of the memoryless GMM-based block quantiser using SD criterion (16 clusters) as a function of bitrate

Bits/frame	Avg. SD (dB)	Outliers (in %)	
		2–4 dB	>4 dB
21	1.247	3.65	0.01
22	1.174	2.53	0.01
23	1.107	1.66	0.00
24	1.042	1.12	0.00
25	0.981	0.76	0.00

Comparing with Table 3, it can be observed that by coding two frames jointly, a spectral distortion of approximately 1 dB can be achieved using 23 bits/frame while the memoryless quantiser requires 24 bits/frame. By quantising more frames ($p = 3$ and 4) jointly, only 22 bits/frame are needed to achieve the same level of spectral distortion. In comparison, the 21 bits/frame multi-frame GMM-based block quantiser ($p = 4$, $m = 16$) is equivalent to the 24 bits/frame memoryless GMM-based block quantiser, in terms of spectral distortion. Therefore, the exploitation of memory across four successive frames by the KLT enables a saving of 3 bits/frame.

In addition, by comparing the computational complexity of the two schemes (from Table 2), we can see that the multi-frame GMM-based block quantiser using MSE criterion is also more computationally efficient than the memoryless scheme ($p = 1$) with SD criterion of (Subramaniam and Rao, 2003).

5.5. Comparison with the predictive GMM-based block quantiser

Two configurations of a GMM-based block quantiser with memory were described in (Subramaniam and Rao, 2003), where the differences between successive LSF frames are quantised, similar to differential pulse code modulation (DPCM) and predictive vector quantisation (Subramaniam and Rao, 2003; Gersho and Gray, 1992). In the first configuration, referred to as the 'modified case' (Subramaniam and Rao, 2003), existing codebooks and transformation matrices are used for quantising the difference vectors while the cluster mean is not subtracted. In the second configuration, referred to as the 'trained case' (Subramaniam and Rao, 2003), the codebooks and transformation matrices are trained based on vector differences.

In our experiments, we have implemented the 'trained case' for comparison, where our codebooks are designed based on difference training vectors. The structure is similar to DPCM with a one-tap predictor. The prediction coefficients, $\{a_i\}_{i=1}^{10}$, were calculated using the covariance method and are given in Table 8. The spectral distortion performance of the 'trained case' of the predictive GMM-based block quantiser is shown in Table 9. When comparing this table with the memoryless GMM-based block quantiser in Table 7, we observe lower spectral distortions which are a result of redundancy between successive frames being exploited by the prediction scheme. However, there appears to be a significantly high percentage of outliers at all bitrates.

Table 8
Prediction coefficients for the predictive GMM-based block quantiser (calculated using the covariance method)

i	a_i
1	0.969321
2	0.983092
3	0.988771
4	0.994216
5	0.996263
6	0.997416
7	0.998717
8	0.999198
9	0.999600
10	0.999832

Table 9
Spectral distortion of the predictive GMM-based block quantiser using SD criterion (16 clusters) as a function of bitrate

Bits/frame	Avg. SD (dB)	Outliers (in %)	
		2–4 dB	>4 dB
21	1.154	5.55	0.60
22	1.077	4.26	0.48
23	1.003	3.25	0.38
24	0.939	2.56	0.34
25	0.877	1.97	0.26

Comparing Tables 3 and 9, we can conclude that the multi-frame GMM-based block quantiser achieves lower spectral distortion and less outliers than the predictive GMM-based block quantiser. This may be explained by the effective exploitation of temporal redundancies across multiple frames, rather than just successive ones.

5.6. Comparison with the split vector quantiser

Table 10 shows the spectral distortions, computational complexity, and memory requirements of the two-part split vector quantiser described in

Table 10
Spectral distortion (SD), computational complexity, and memory requirements (ROM) of the two-part split vector quantiser as a function of bitrate

Bits/frame	Avg. SD (in dB)	Outliers (in %)		kflops/ frame	ROM (floats)
		2–4 dB	>4 dB		
23	1.223	3.87	0.00	114.7	28672
24	1.124	2.16	0.00	163.8	40960
25	1.090	1.89	0.00	229.4	57344

(Paliwal and Atal, 1993).[10] In this scheme, LSF vectors are split into two parts (4, 6) and quantised using separate vector quantisers. Bits are allocated uniformly where-ever possible (Paliwal and Atal, 1993). Comparing with Table 3, it can be seen that the multi-frame GMM-based block quantiser performs better than the split vector quantiser in terms of spectral distortion since the latter does not utilise memory across frames. Also, the computational complexity and memory requirements of the split vector quantiser are much higher than the 16 cluster, multi-frame GMM-based block quantiser (Table 2).

6. Conclusion

In this paper, we have investigated the multi-frame GMM-based block quantiser for the coding of line spectral frequencies. By concatenating multiple frames together, correlation between LSFs within each frame and across successive frames can be exploited by the KLT, leading to better coding. The efficiency gained from joint quantisation permits the use of the mean squared error distortion criterion for cluster quantiser selection, rather than the computationally expensive spectral distortion, and this has led to considerable reductions in computational complexity. It has been shown that the multi-frame GMM-based block quantiser can achieve a spectral distortion of 1 dB at 22 bits/frame, or 21 bits/frame with some added complexity. This scheme also performs better than the split vector quantiser at a fraction of the computational complexity and memory.

[10] The full-band spectral distortion (0–4 kHz) results are from a split vector quantiser which uses MSE as the distance measure which we have found to be comparable to one based on the weighted Euclidean distance measure described in (Paliwal and Atal, 1993).

References

Atal, B.S., Schroeder, M.R., 1979. Predictive coding of speech signals and subjective error criteria. IEEE Trans. Acoust., Speech, Signal Process. ASSP-27 (3), 247–254.

Campbell, Jr., J.P., Welch, V.C., Tremain, T.E., 1989. An expandable error-protected 4800 bps CELP Coder (U.S. Federal Standard 4800 bps voice coder). In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Glasgow, Scotland, May 1989, pp. 735–738.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. 39, 1–38.

Gardner, W.R., Rao, B.D., 1995. Theoretical analysis of the high-rate vector quantization of LPC parameters. IEEE Trans. Speech Audio Process. 3 (5), 367–381.

Gersho, A., Gray, R.M., 1992. Vector Quantization and Signal Compression. Kluwer Academic Publishers, Massachusetts.

Gray, A., Markel, J., 1976. Quantization and bit allocation in speech processing. IEEE Trans. Acoust., Speech, Signal Process. ASSP-24, 459–473.

Hedelin, P., Skoglund, J., 2000. Vector quantization based on Gaussian mixture models. IEEE Trans. Speech Audio Process. 8 (4), 385–401.

Huang, J.J.Y., Schultheiss, P.M., 1963. Block quantization of correlated Gaussian random variables. IEEE Trans. Commun. Syst. CS-11, 289–296.

Itakura, F., 1975. Line spectrum representation of linear predictive coefficients of speech signals. J. Acoust. Soc. Am. 57 (Apr.), S35.

Itakura, F., Saito, S., 1969. Speech analysis-synthesis based on the partial autocorrelation coefficient. Proc. JSA, 199–200.

Kroon, P., Kleijn, W.B., 1995. Linear-prediction based analysis-by-synthesis coding. In: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier, Amsterdam, pp. 79–119.

LeBlanc, W.P., Bhattarchaya, B., Mahmoud, S.A., Cuperman, V., 1993. Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding. IEEE Trans. Speech Audio Process. 1 (Oct.), 373–385.

Linde, Y., Buzo, A., Gray, R.M., 1980. An algorithm for vector quantizer design. IEEE Trans. Commun. COM-28 (1), 84–95.

Nurminen, J., 2003. Multi-mode quantization of adjacent speech parameters using a low-complexity prediction scheme. In: Proc. EuroSpeech, September 2003, pp. 1073–1076.

Paliwal, K.K., Atal, B.S., 1993. Efficient vector quantization of LPC parameters at 24 bits/frame. IEEE Trans. Speech Audio Process. 1 (1), 3–14.

Paliwal, K.K., Kleijn, W.B., 1995. Quantization of LPC parameters. In: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier, Amsterdam, pp. 443–466.

Paliwal, K.K., So, S., 2004. Multiple frame block quantisation of line spectral frequencies using Gaussian mixture models. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Montreal, pp. I-149–I-152.

Proakis, J.G., Manolakis, D.G., 1996. Digital Signal Processing: Principles, Algorithms, and Applications, third ed. Prentice-Hall, New Jersey.

Subramaniam, A.D., Rao, B.D., 2000. PDF optimized parametric vector quantization with applications to speech coding. In: 34th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, November 2000.

Subramanian, A.D., Rao, B.D., 2001. Speech LSF quantization with rate independent complexity, bit scalability and learning. In: Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 2, pp. 705–708.

Subramaniam, A.D., Rao, B.D., 2003. PDF optimized parametric vector quantization of speech line spectral frequencies. IEEE Trans. Speech Audio Process. 11 (2), 130–142.

Shabestary, T.Z., Hedelin, P., 2002. Spectral quantization by companding. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, vol. 1, pp. 641–644.

Sinervo, U., Nurminen, J., Heikkinen, A., Saarinen, J., 2003. Multi-mode matrix quantizer for low bit rate LSF quantization. In: Proc. EuroSpeech, September 2003, pp. 1073–1076.

Soong, F.K., Juang, B.H., 1984. Line spectrum pair (LSP) and speech data compression. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, San Diego, California, March 1984, pp. 37–40.

Sugamura, N., Itakura, F., 1986. Speech analysis and synthesis methods developed at ECL in NTT–from LPC to LSP–. Speech Commun. 5 (Jan.), 199–215.

Tsao, C., Gray, R.M., 1985. Matrix quantizer design for LPC speech using the generalized Lloyd algorithm. IEEE Trans. Acoust., Speech, Signal Process. ASSP-33 (3), 537–545.

Viswanathan, R., Makhoul, J., 1975. Quantization properties of transmission parameters in linear predictive systems. IEEE Trans. Acoust., Speech, Signal Process. ASSP-23, 309–321.

Xydeas, C.S., Papanastasiou, C., 1999. Split matrix quantization of LPC parameters. IEEE Trans. Speech Audio Process. 7 (2), 113–125.