Available online at www.sciencedirect.com
ScienceDirect
Digital Signal Processing 17 (2007) 114–137
Digital Signal Processing
www.elsevier.com/locate/dsp

A comparative study of LPC parameter representations and quantisation schemes for wideband speech coding

Stephen So [a,*], Kuldip K. Paliwal [b]

[a] *School of Engineering, Griffith University, PMB 50 Gold Coast Mail Centre, Gold Coast, QLD 9726, Australia*
[b] *School of Microelectronic Engineering, Griffith University, Brisbane, QLD 4111, Australia*

Available online 8 November 2005

Abstract

In this paper, we provide a review of LPC parameter quantisation for wideband speech coding as well as evaluate our contributions, namely the switched split vector quantiser (SSVQ) and multi-frame GMM-based block quantiser. We also compare the performance of various quantisation schemes on the two popular LPC parameter representations: line spectral frequencies (LSFs) and immittance spectral pairs (ISPs). Our experimental results indicate that ISPs are superior to LSFs by 1 bit/frame in independent quantiser schemes, such as scalar quantisers; while LSFs are the superior representation for joint vector quantiser schemes. We also derive informal lower bounds, 35 and 36 bits/frame, for the transparent coding of LSFs and ISPs, respectively, via the extrapolation of the operating distortion-rate curve of the unconstrained vector quantiser. Finally, we report and discuss the results of applying the SSVQ with dynamically-weighted distance measure and the multi-frame GMM-based block quantiser, which achieve transparent coding at 42 and 37 bits/frame, respectively, for LSFs. ISPs were found to be inferior to the LSFs by 1 bit/frame. In our comparative study, other quantisation schemes that were investigated include PDF-optimised scalar quantisers, the memoryless Gaussian mixture model-based block quantiser, the split vector quantiser, and the split-multistage vector quantiser with MA predictor from the AMR-WB (ITU-T G.722.2) speech coder.

1. Introduction

The quantisation of linear predictive coding (LPC) parameters in code-excited linear predictive (CELP) coders for narrowband speech (300–3400 Hz) has been thoroughly investigated in the literature, where product code vector quantisers operating on vectors of 10 line spectral frequency (LSF) parameters [1], generally require 24 bits/frame for transparent quality [2,3].

While narrowband speech has acceptable quality (otherwise known as toll quality) that is similar to telephone speech, it was found to be inadequate for applications that demanded higher quality reconstruction, such as videophones, teleconferencing, multimedia, etc. Problems with narrowband speech include a lack of naturalness and speaker 'presence' (such as that experienced in face-to-face speech communication) as well as difficulty in distinguishing between fricative sounds, such as /*s*/ and /*f*/ [4]. All of these can lead to listener fatigue.

* Corresponding author. Fax: +61 7 5552 8065.
E-mail addresses: s.so@griffith.edu.au (S. So), k.paliwal@griffith.edu.au (K.K. Paliwal).

1051-2004/$ – see front matter
doi:10.1016/j.dsp.2005.10.002

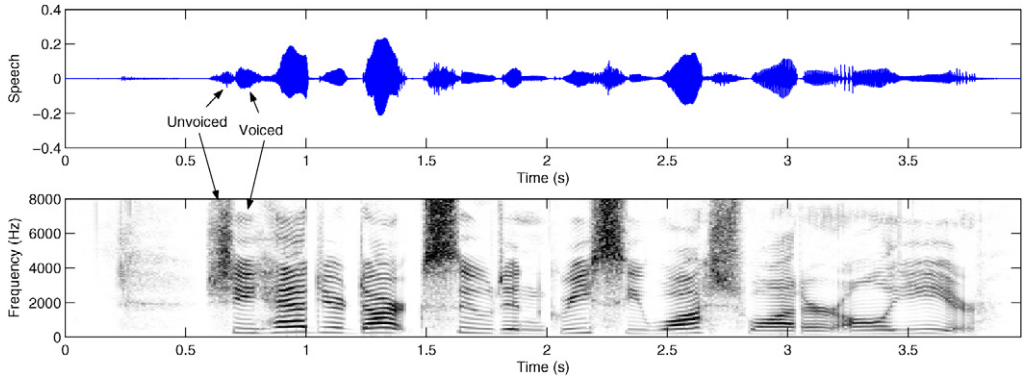

Fig. 1. Waveform and spectrogram of wideband speech. The sentence that is spoken is *'she had your dark suit in greasy wash-water all year,'* and with the unvoiced /*s*/ and voiced /*iy*/ sounds in *she*, highlighted.

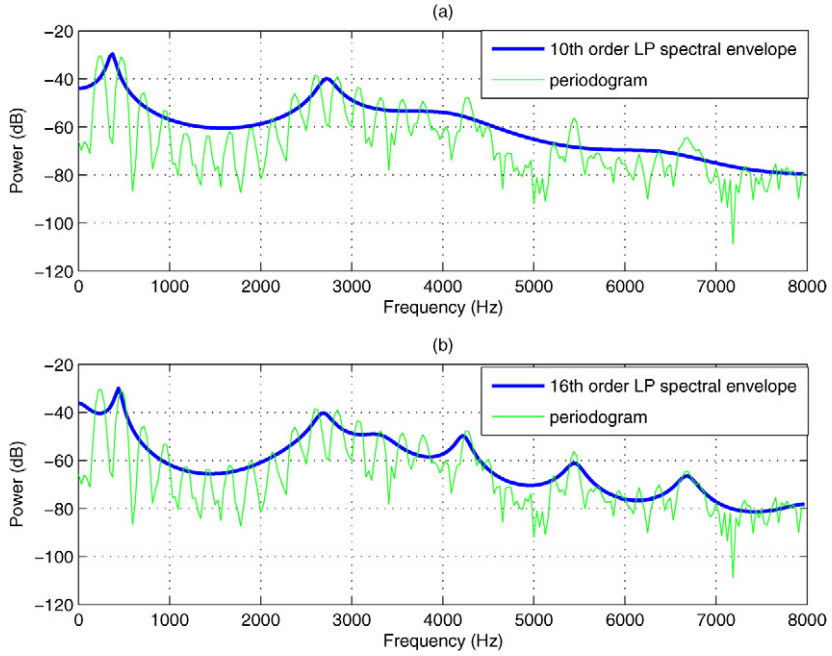

Fig. 2. Spectral envelope estimate of 20 ms of wideband speech (starting from 3.53 s) for different orders of linear prediction analysis: (a) using 10th-order autocorrelation method; (b) using 16th-order autocorrelation method.

By lowering the low frequency cut-off from 300 to 50 Hz, the naturalness and fullness of the speech can be improved, while extending the high frequency cut-off from 3400 to 7000 Hz, improves the distinguishing of fricative sounds [5]. This extended range of 50–7000 Hz of wideband speech roughly corresponds to the bandwidth of speech sampled at 16 kHz. Figure 1 shows the waveform and spectrogram of 16 kHz speech. We can see that most of the spectral information for voiced speech, in the form of formants, occurs below 3.4 kHz. However, for unvoiced speech, there is some spectral information extending beyond 3.4 kHz that may be important for discriminating fricative sounds. For example, the spectrum starting at 1.5 and 2.25 s is stronger at frequencies above 4 kHz (darker areas) than the spectra for other unvoiced sections.

Figure 2 shows the spectral envelope estimates and power spectral densities of wideband speech (50–7000 Hz) using different orders of linear prediction analysis. Using the speech utterance of Fig. 1, a 20 ms frame was extracted starting from 3.53 s. A Hamming window was applied to prevent the masking of high frequency formants by spectral leakage and the autocorrelation method was used in the linear prediction analysis.

We can see that the 10th-order linear prediction analysis has captured the two formants below 4 kHz (Fig. 2a), which corresponds to the frequency range of narrowband speech. However, three formants can be observed above 4 kHz which are not captured by the 10th-order linear prediction analysis. Therefore, higher orders are required for accurate estimation of the short-term correlation information in wideband speech. Figure 2b shows the spectral envelope estimate from a 16th-order linear prediction analysis, where we can see the capturing of the higher order formants.

With the introduction of high-speed data services in wireless communication systems, wideband speech can now be accommodated [6]. However, because wideband CELP coders typically require 16 LPC parameters for representing the speech spectral envelope, vector quantisers need to operate at higher bitrates and on vectors of larger dimension, which both impact on the computational complexity and memory requirement in an exponential fashion. Therefore, structurally-constrained vector quantisers, such as split and multistage vector quantisers, need to be employed for wideband LPC parameter quantisation.

In this paper, we report on our contributions to wideband LPC parameter quantisation, namely the switched split vector quantiser (SSVQ) and multi-frame GMM-based block quantiser, and compare them to other quantisation schemes. In addition, we also compare two LPC parameter representations that are being used in wideband speech coders: line spectral frequencies (LSFs) and immittance spectral pairs (ISPs). The organisation of this paper is as follows. In Section 2, we take a look at LPC parameter representations and the spectral distortion, which is commonly used to objectively measure reconstructed speech quality, as well as review the quantisation schemes that have been reported in the literature. Then in Section 3, we present and discuss the experimental results of various quantisation schemes, such as PDF-optimised scalar quantisers, split-multistage vector quantisers, split and switched split vector quantisers, and memoryless and multi-frame Gaussian mixture model-based block quantisers. We also compare and contrast the relative performance of quantising LSFs and ISPs as well as derive an informal lower bound for achieving transparent coding via extrapolation of the operating distortion-rate curve of an unconstrained vector quantiser. Finally, we offer our conclusions in Section 4.

2. Quantisation of wideband LPC parameters

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, $H(z) = 1/A(z)$, driven by either white Gaussian noise (for unvoiced speech) or a periodic sequence of impulses (for voiced speech), where $A(z)$ is the inverse filter given by [2]

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

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) [7], arc-sine reflection coefficients (ASRC) [8], and log area ratios (LAR) [9]. The line spectral frequency (LSF) representation, proposed by Itakura in [1], has been shown in the literature to be superior to other representations for speech coding [2,10,11].

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 [10,11]. Therefore, n LPC coefficients, $[a_1, a_2, \ldots, a_n]$, can be converted to n line spectral frequencies, $[\omega_1, \omega_2, \ldots, \omega_n]$ [11]. Consequently, LSFs have the following properties [10]:

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;
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 [2,11]. Quantisation errors in the LSFs result in localised distortion in the power spectrum [2].

2.2. ISP representation of LPC coefficients

The immittance spectral pairs (ISP) representation was introduced by Bistritz and Peller [12]. It comprises the poles and zeros of the following immittance[1] function at the glottis [12]:

$$\mathcal{I}_n(z) = \frac{A(z) - z^{-n}A(z^{-1})}{A(z) + z^{-n}A(z^{-1})} \tag{4}$$

as well as the nth reflection coefficient, k_n. Because the coefficients of the immittance function are real, the roots of both the numerator and denominator will occur in complex conjugate pairs. Therefore, there are a total of $n-1$ poles and zeros (excluding the zeros and/or poles at -1 and 1), which lie on the unit circle, and along with a 'constant' gain [12], which is expressed as the nth reflection coefficient, k_n, constitute the n parameter ISP representation, $[\cos\omega_1, \cos\omega_2, \ldots, \cos\omega_{n-1}, k_n]$. Alternatively, the poles and zeros can be transformed to frequencies, $[\omega_1, \omega_2, \ldots, \omega_{n-1}, \frac{1}{2}\cos^{-1} k_n]$, which are sometimes known as immittance spectral frequencies (ISFs) [13].

The ISP representation possesses similar properties to the LSFs, namely that the roots are interlaced and lie on the unit circle. The stability of the LPC filter is guaranteed via the ascending order of the ISPs along with the further condition that the reflection coefficient, $|k_n| < 1$ [12].

The literature on the quantisation performance of ISPs is less than that of LSFs, though Bistritz and Peller [12] report a 1 bit saving, when using ISPs compared with LSFs, in differential scalar quantisation experiments. The use of block and vector quantisation of ISPs may need to be handled in a special way as the ISP representation consists of a concatenation of two different variables: $n-1$ frequencies and a reflection coefficient. Each will have their own unique quantisation characteristics and sensitivity. This is in contrast to the LSFs, which are all of the same type (frequency). Therefore, it is customary to quantise ISFs, where an arc-cosine is applied to the reflection coefficient, which tends to flatten its sensitivity curve.

2.3. Spectral distortion and conditions for transparent coding

In order to objectively measure the distortion between a coded and uncoded LPC parameter vector, the spectral distortion is often used in narrowband speech coding [2]. For the ith frame, the spectral distortion (in dB), D_i, is defined as

$$D_i = \sqrt{\frac{1}{F_s}\int_0^{F_s} \left[10\log_{10} P_i(f) - 10\log_{10} \hat{P}_i(f)\right]^2 df}, \tag{5}$$

where F_s is the sampling frequency, and $P_i(f)$ and $\hat{P}_i(f)$ are the LPC power spectra of the coded and uncoded ith frame, respectively. The conditions for transparent coding of speech from LPC parameter quantisation are [2]:

1. the average spectral distortion (SD) is approximately 1 dB;
2. there is no outlier frame having more than 4 dB of spectral distortion;
3. less than 2% of outlier frames are within the range 2–4 dB.

[1] This is formed from the two words, impedance and admittance [12].

According to Guibé et al. [14], listening tests have shown that these conditions for transparency, which are often quoted in the narrowband speech coding literature, also apply to the wideband case. Therefore, we will make our objective judgement on the quality of the quantisation scheme based on the average spectral distortion.

2.4. Literature review of wideband LPC parameter quantisation

Harborg et al. [15] quantised 16–18 log-area-ratio coefficients at 60–80 bits/frame using non-uniform, PDF-optimised scalar quantisers, in their real-time wideband CELP coder. Lefebvre et al. [16] used a split vector quantiser with seven part splitting $(2, 2, 2, 2, 2, 3, 3)$ to quantise 16 LSFs at 48 bits/frame in their TCX wideband coder. Chen et al. [17] also used a seven-part split vector quantiser, with the same subvector sizes as in [16] operating at 49 bits/frame to quantise 16 LSF parameters in their transform predictive coder. Transparent results were reported by Biundo et al. [18] for a four and five part split vector quantiser at 45 bits/frame.

Because successive LSF frames are highly correlated [14], better quantisation can be achieved by exploiting the interframe correlation. Roy and Kabal [19] quantised 16 LSFs at 50–60 bits/frame using non-uniform differential scalar quantisation. They noted that because the LSFs were related to the formant positions, then the lower frequency LSFs were perceptually more important, hence more bits should be allocated to them. Combescure et al. [20] quantised 12 LSFs from a decimated lower band using a predictive split multistage vector quantiser at 33 bits/frame. Ubale et al. [21] used a seven-stage tree-searched multistage vector quantiser [3] with a moving average (MA) predictor at 28 bits/frame, while Biundo et al. [18] reported transparent results using an MA predictive split multistage vector quantiser (S-MSVQ) at 42 bits/frame. Guibé et al. [14] achieved transparent coding using a safety-net vector quantiser at 38 bits/frame, while the adaptive multi-rate wideband (AMR-WB) speech codec [6,13] uses an S-MSVQ with MA predictor at 46 bits/frame to quantise ISPs.

Other quantisation schemes recently reported include the predictive Trellis-coded quantiser [22] and the hidden Markov model-based recursive quantiser [23] which achieve a spectral distortion of 1 dB at 34 and 40 bits/frame, respectively.

2.5. AMR-WB LPC analysis and quantisation

The AMR-WB wideband speech coder was adopted by the 3GPP in 2000 and standardised by the ITU-T as recommendation G.722.2. It is based on the ACELP coder and can operate at different bitrates (23.85, 23.05, 19.85, 18.25, 15.85, 14.25, 12.65, 8.85, and 6.6 kbps) [6]. In this section, we describe briefly the main points of operation, specifically the pre-processing, LPC analysis, and LPC parameter quantisation. For more details on the AMR-WB speech coder, the reader should consult Refs. [6,13].

2.5.1. Speech pre-processing

The 16 kHz speech is processed in two frequency bands, 50–6400 and 6400–7000 Hz, to decrease complexity and prioritise bit allocation to the subjectively important lower band [6]. In the lower band, speech is down-sampled to 12.8 kHz and pre-processed using a high pass filter, $H_{\mathrm{h}}(z)$, and pre-emphasis filter, $H_{\mathrm{p}}(z)$, which are given by [13]

$$H_{\mathrm{h}}(z) = \frac{0.989502 - 1.979004z^{-1} + 0.989502z^{-2}}{1 - 1.978882z^{-1} + 0.979126z^{-2}}, \tag{6}$$

$$H_{\mathrm{p}}(z) = 1 - 0.68z^{-1}. \tag{7}$$

The role of the high pass filter is to remove undesired low frequency components while the pre-emphasis filter removes the spectral tilt in the speech spectrum and emphasises higher frequency formants for more accurate LPC analysis.

2.5.2. LPC analysis

Speech frames of 20 ms are extracted using a 30 ms asymmetric Hamming window which emphasises the fourth subframe (subframes are 5 ms long each). Autocorrelations are calculated with a 5 ms lookahead. The autocorrelation method is used to perform a 16th-order LPC analysis with a 60 Hz bandwidth expansion (using the lag window method) and a white noise correction factor of 1.0001 [13]. The 16 LPC coefficients are transformed to immittance

spectral pairs (ISPs) [12], $\{q_i\}_{i=1}^{16}$, and then further converted to immittance spectral frequencies[2] (ISFs), $\{f_i\}_{i=1}^{16}$, for quantisation [13]:

$$f_i = \begin{cases} \frac{F_s}{2\pi}\cos^{-1}(q_i) & \text{for } i = 1, 2, \ldots, 15, \\ \frac{F_s}{4\pi}\cos^{-1}(q_i) & \text{for } i = 16, \end{cases} \tag{8}$$

where $F_s = 12.8$ kHz is the sampling frequency.

2.5.3. Quantisation of residual ISF vectors

In order to exploit interframe correlation, a first-order moving average (MA) predictor is applied:

$$\boldsymbol{r}(n) = \boldsymbol{f}(n) - \boldsymbol{p}(n), \tag{9}$$

where $\boldsymbol{r}(n)$ is the prediction residual vector, $\boldsymbol{f}(n)$ is the current frame of ISFs, and $\boldsymbol{p}(n)$ is the predicted ISF that is given by

$$\boldsymbol{p}(n) = \frac{1}{3}\hat{\boldsymbol{r}}(n-1), \tag{10}$$

where $\hat{\boldsymbol{r}}(n-1)$ is the quantised residual vector of the previous frame [13].

The residual ISF vectors are quantised using split-multistage vector quantisation (S-MSVQ) at 36 bits for the lowest bitrate mode (6.6 kbps); and 46 bits for the other higher bitrate modes. S-MSVQ first appeared in the ATCELP wideband speech coder described by Combescure et al. [20] and is essentially a multistage vector quantiser with split vector quantisers in each stage. It combines the low complexity characteristics of both split vector quantisers (SVQ) and multistage vector quantisers (MSVQ), making it practical to quantise 16-dimensional vectors using 46 bits. The S-MSVQ and its performance will be discussed in Section 3.4.

2.6. Switched split vector quantisation

The switched split vector quantiser (SSVQ) was introduced by So and Paliwal [24] for the efficient vector quantisation of LSFs for narrowband speech coding. It is a hybrid scheme, which combines the advantages of the switch vector quantiser and split vector quantiser. The switch vector quantiser classifies the vector space into various clusters and individual SVQs are designed for each cluster. Vectors are switched to the appropriate SVQ and quantised using that particular codebook. The initial switch vector quantiser allows the exploitation of full vector dependencies to compensate for the performance loss caused by vector splitting. Also, the switched vector quantiser is equivalent to a tree structure with m branches. These two characteristics result in a scheme that has better rate-distortion efficiency and lower computational complexity than the SVQ, though the memory requirements are higher. For more details on the operation and advantages of SSVQ, the reader should refer to [24,25].

2.6.1. SSVQ codebook training

Figure 3 shows a block diagram of the SSVQ codebook training. The LBG algorithm [26] is first applied on all vectors to produce m centroids (or means) $\{\boldsymbol{\mu}_i\}_{i=1}^{m}$. In the Euclidean distortion sense, these centroids are the 'best' representation of all the vectors in that Voronoi region. Hence, we can use them to form the switch VQ codebook which will be used for switch-direction selection. All the training vectors are classified based on the nearest-neighbour criterion:

$$j = \underset{i}{\operatorname{argmin}}\, d(\boldsymbol{x}, \boldsymbol{\mu}_i), \tag{11}$$

where $\boldsymbol{x}$ is the vector under consideration, $d(\boldsymbol{x}, \boldsymbol{\mu}_i)$ is the distance measure (usually mean squared error) between the vector and switch VQ code-vector, and j is the cluster (or switching direction) to which the vector is classified. With the training vectors classified to the m clusters, local SVQ codebooks are designed for each cluster (or switching direction) using the corresponding training vectors.

[2] It should be noted that the last ISF corresponds to half the arc-cosine of the 16th reflection coefficient, thus it is a different quantity to the first 15 ISFs.

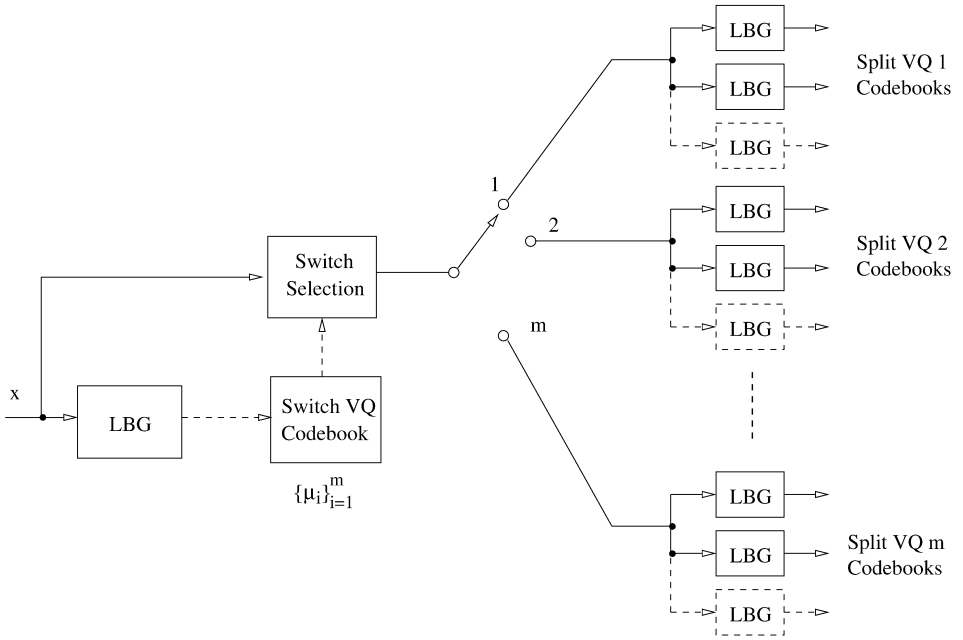

Fig. 3. SSVQ codebook training.

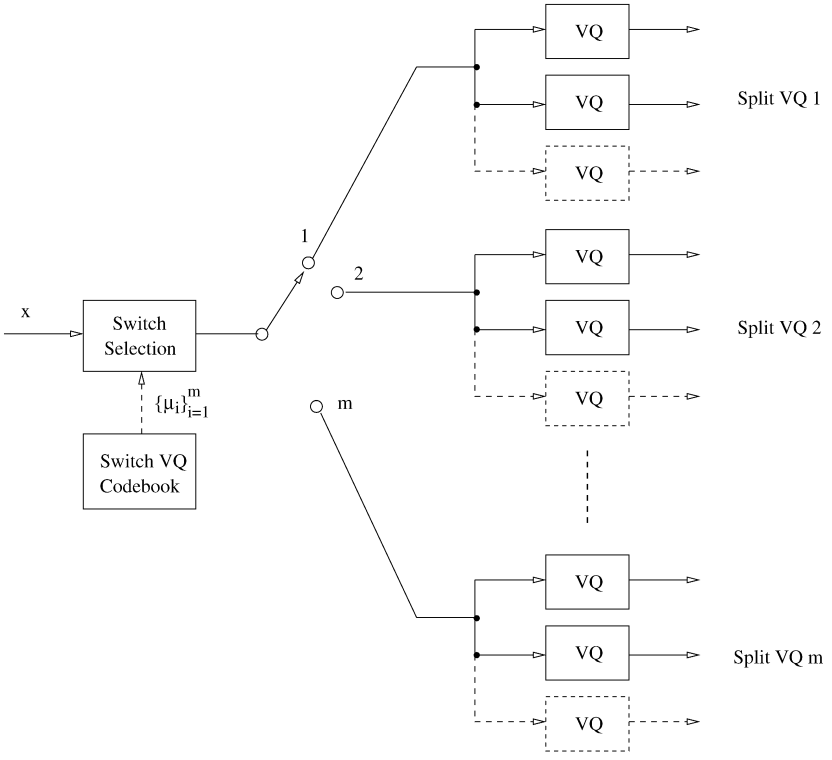

Fig. 4. SSVQ coding.

2.6.2. *SSVQ coding*

Figure 4 shows a block diagram of SSVQ coding. Each vector to be quantised is first switched to one of the m possible directions based on the nearest-neighbour criterion defined by (11), using the switch VQ codebook, $\{\boldsymbol{\mu}_i\}_{i=1}^m$, and then quantised using the corresponding SVQ.

2.7. Multi-frame GMM-based block quantisation

The multi-frame GMM-based block quantiser is a modified version of the scheme described in [27] that exploits correlation across multiple frames. It was introduced by So and Paliwal [28] for narrowband LPC quantisation and recently applied to wideband LPC quantisation [29] as well. p successive frames of n LPC parameters are concatenated to form extended vectors, which have a dimension of $N = np$. These extended vectors are then processed by the memoryless GMM-based block quantiser [27], as per usual.

2.7.1. 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 [30] on the training vectors where m clusters[3] 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 [31] 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.

2.7.2. 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.

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 [27]

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

where [27]

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

and $\lambda_{i,j}$ is the jth eigenvalue of the ith cluster. Then for each block quantiser, the high resolution formula from [32] is used to distribute the b_i bits to each of the vector components:

$$b_{i,j} = \frac{b_i}{N} + \frac{1}{2} \log_2 \frac{\lambda_{i,j}}{\left(\prod_{j=1}^{N} \lambda_{i,j} \right)^{\frac{1}{N}}} \quad \text{for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, N. \tag{12}$$

Figure 5 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,[4] 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, $\boldsymbol{z}_i$:

$$\boldsymbol{z}_i = \frac{\boldsymbol{P}_i (\boldsymbol{x} - \boldsymbol{\mu}_i)}{\boldsymbol{\sigma}_i}, \tag{13}$$

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

[4] A block quantiser, as defined in [32], is a special fixed-rate transform coder.

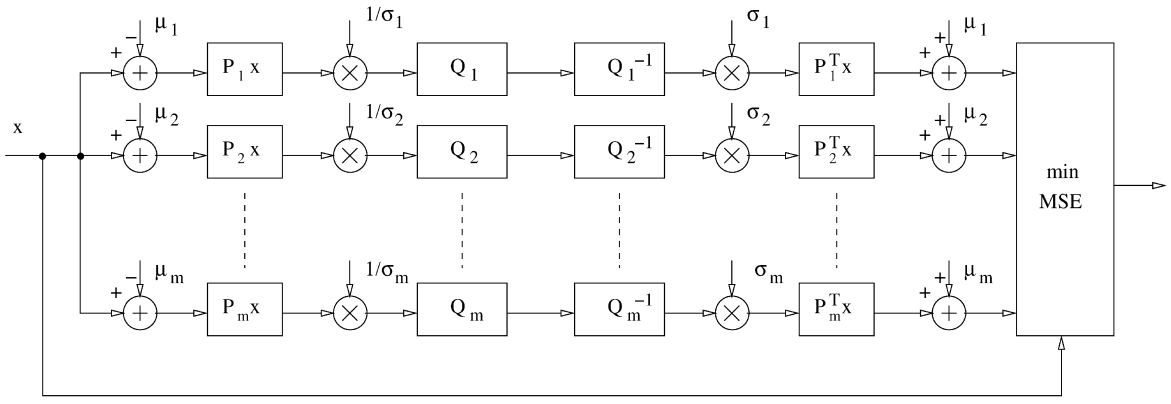

Fig. 5. Minimum distortion block quantisation (Q-block quantiser).

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 [32] 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, $\hat{\boldsymbol{x}}_i$:

$$\hat{\boldsymbol{x}}_i = \boldsymbol{P}_i^{\mathrm{T}} \boldsymbol{\sigma}_i \hat{\boldsymbol{z}}_i + \boldsymbol{\mu}_i. \tag{14}$$

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:

$$k = \underset{i}{\operatorname{argmin}}\, d(\boldsymbol{x} - \hat{\boldsymbol{x}}_i). \tag{15}$$

An appropriate distortion measure can be chosen for this stage, depending upon the application. In LPC parameter quantisation for speech coding, the spectral distortion (SD) is an appropriate distortion measure [27], though we will use the computationally-simpler MSE as the distortion measure for minimum distortion block quantisation in the multi-frame GMM-based block quantiser.

Because of the use of independent scalar quantisers in the KLT domain, the order of the LSFs and ISPs 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.

2.8. *Comments with respect to other reported recursive coding schemes*

The multi-frame GMM-based block quantiser is similar to the GMM-based [33] and HMM-based [23] recursive coders. Recursive coders are memory-based schemes where for each frame to be coded, the codebook is adapted in a way that is dependent on previous frames [34]. For recursive vector quantisers, the calculation of the new codebook needs to be done for each frame, which leads to problems with synchronisation between encoder and decoder as well as a high computational complexity.

The GMM-based recursive coder [33] addresses these two problems by extending the GMM-based block quantisation scheme of [27]. In this scheme, the source model (in the form of a GMM) contains all the information about the vector source itself, such as the PDF and correlation. Hence, because the scalar quantisation codebooks are fixed, then only the source model needs to be adapted via the calculation of the conditional density of the current frame, based on previous quantised frames [33]. The backward prediction resolves the synchronisation issue, as the decoder will have access to previous quantised frames. In a similar fashion to the multi-frame GMM-based block quantiser,

the initial GMM is derived based on concatenated frames [33]. However, the GMM-based recursive coder adaptively adjusts the GMM parameters (the mean and mixture weights) and bit allocation for the current source frame, based on previous quantised frames. This contrasts to the static source model and multiple frame coding that are used in the multi-frame GMM-based block quantiser. Therefore, we would expect the GMM-based recursive coder to achieve lower distortion due to its adaptive source model and bit allocation, as well as lower complexity and delay, since it codes single frames only. However, because backward prediction is used to compute the conditional density, it is expected that bit errors will propagate and lead to a suboptimal GMM. Furthermore, in backward prediction schemes, outlier frames having a spectral distortion of greater than 2 dB tend to be more prominent than for the multi-frame GMM-based block quantiser, which uses forward prediction.

As an informal[5] comparison, based on the cepstrum vector coding results in [33], a saving of approximately 5 bits/frame for transparent coding was observed when going from the fixed-rate memoryless quantiser to the first order recursive quantiser. However, there was a sharp increase in the number of outlier frames (1.3–4.6%) that have a spectral distortion of between 2 and 4 dB [33]. On the other hand, if we compare the fixed-rate memoryless GMM-based block quantiser (given later in Table 7) to the multi-frame GMM-based block quantiser (given later in Table 14) for LSF quantisation, there is a 3 bit/frame saving when quantising two frames ($p = 2$) and only a minor increase in outlier frames (0.88–0.91%).

3. Comparative study between LPC parameter representations and quantisation schemes

3.1. Experimental setup

The TIMIT database was used in the training and testing of the quantisation experiments, where speech is sampled at 16 kHz. We have used the lower band pre-processing and LPC analysis of the 3GPP implementation of the AMR-WB speech codec (floating point version) [13] to produce linear prediction coefficients which are then converted to line spectral frequency (LSF) representation [1] and immittance spectral pairs (ISP) representation [12] (quantisation is performed on the frequency form of ISPs, i.e., ISFs).

The training set consists of 333,789 vectors while the evaluation set, consisting of speech not contained in the training, has 85,353 vectors. Unless specified otherwise, unweighted mean squared error is used as the distance measure for quantiser design and spectral distortion is used for evaluating quantisation performance.

3.2. PDF-optimised scalar quantisers with non-uniform bit allocation

Figure 6 shows the histogram of ISFs and LSFs from the test set of the TIMIT database. We can see that the distribution of the low to medium frequency parameters tend to be multimodal while high frequency ones are unimodal. Therefore, the best strategy is to quantise each LPC parameter using a scalar quantiser that is optimised for each specific probability density function (PDF). In this experiment, we have designed non-uniform scalar quantisers using the generalised Lloyd algorithm for each LPC parameter. Bit allocation was performed using a greedy algorithm that is similar to that described by Soong and Juang [35], where each bit is given to the scalar quantiser which results in the largest distortion improvement. This simple algorithm results in a locally optimal allocation of bits. To reduce the computational complexity of the bit allocation procedure, we have used mean squared error as the distortion measure rather than spectral distortion.

Table 1 shows the spectral distortion performance of non-uniform, PDF-optimised scalar quantisers on immittance spectral frequency vectors from the TIMIT database. We can see that a spectral distortion of 1 dB can be achieved at a bitrate of 58 bits/frame. In Table 2, which shows the spectral distortion performance of non-uniform scalar quantisation of line spectral frequency vectors, we can see that 1 dB spectral distortion is achieved at 59 bits/frame. We note that these results reflect the narrowband speech coding results given by Bistritz and Peller [12], where ISPs were observed to also perform better than LSFs by 1 bit/frame, in differential scalar quantisation.

[5] The comparison is informal since the results given for GMM-based recursive coding in [33] are for cepstral parameters, which tend to have different quantisation properties to LSF parameters.

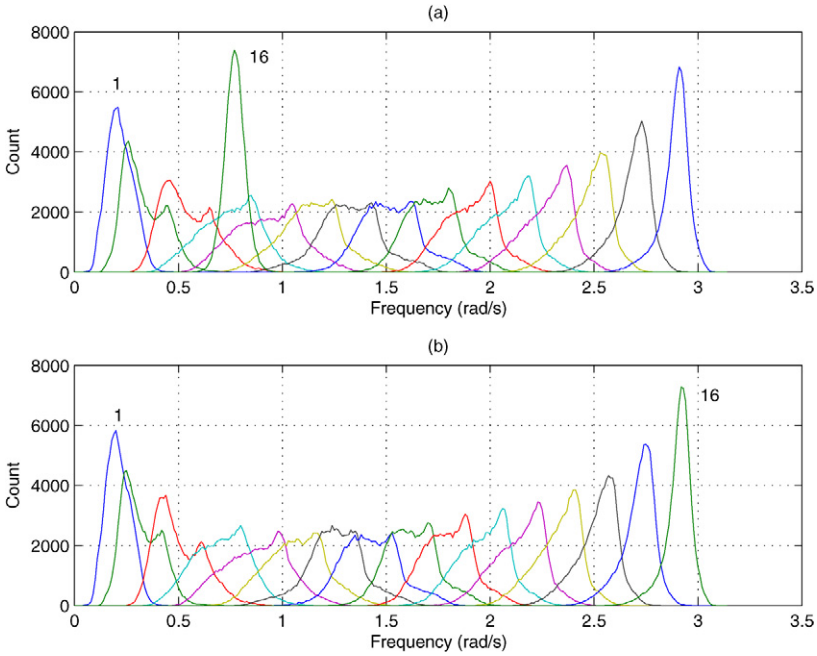

Fig. 6. Histograms of LPC parameters from the TIMIT database (the first and last components are labelled): (a) immittance spectral frequencies (ISFs); (b) line spectral frequencies (LSFs).

Table 1
Average spectral distortion of the PDF-optimised scalar quantisers as a function of bitrate on wideband ISF vectors from the TIMIT database

Bits/frame	Avg. SD (dB)	Outliers (%)	
		2–4 dB	>4 dB
61	0.911	0.85	0.00
60	0.947	0.97	0.00
59	0.961	1.00	0.01
58	1.016	1.22	0.01
57	1.060	1.51	0.01
56	1.137	2.19	0.01
55	1.181	2.63	0.01

Table 2
Average spectral distortion of the PDF-optimised scalar quantisers as a function of bitrate on wideband LSF vectors from the TIMIT database

Bits/frame	Avg. SD (dB)	Outliers (%)	
		2–4 dB	>4 dB
61	0.918	0.82	0.00
60	0.970	0.95	0.00
59	1.011	1.18	0.01
58	1.080	1.64	0.01
57	1.120	1.88	0.01
56	1.162	2.26	0.01
55	1.219	2.98	0.03

3.3. Unconstrained vector quantisers and an informal lower bound for transparent coding

In theory, unconstrained vector quantisers can achieve the lowest distortion of any quantisation scheme at a given bitrate and dimension. However, the exponentially growing complexity and storage, with respect to increasing bitrate

Table 3
Average spectral distortion of the unconstrained vector quantiser as a function of bitrate on wideband ISF vectors from the TIMIT database

Bits/frame	Avg. SD (dB)	Outliers (%)	
		2–4 dB	>4 dB
16	2.445	70.36	1.86
15	2.521	72.09	2.69
14	2.609	73.16	4.01
13	2.705	74.07	5.53
12	2.814	73.96	7.96
11	2.934	72.60	11.30
10	3.067	70.50	15.56
9	3.218	66.79	21.20
8	3.385	61.96	27.80
7	3.577	55.96	35.50

Table 4
Average spectral distortion of the unconstrained vector quantiser as a function of bitrate on wideband LSF vectors from the TIMIT database

Bits/frame	Avg. SD (dB)	Outliers (%)	
		2–4 dB	>4 dB
16	2.420	70.22	1.54
15	2.498	72.09	2.29
14	2.585	73.58	3.28
13	2.686	74.46	4.92
12	2.794	74.31	7.31
11	2.916	73.28	10.63
10	3.051	71.06	14.96
9	3.203	67.20	20.67
8	3.370	62.23	27.41
7	3.563	56.44	35.02

and dimensionality, inhibits their use in practical schemes that require high bitrates. However, they can be used to provide an informal lower bound on the spectral distortion via extrapolation to higher bitrates, similar to that reported in [36] for narrowband LSF quantisation.

We have applied unconstrained vector quantisers to the task of quantising ISF and LSF vectors from the TIMIT database and their spectral distortion performance is shown in Tables 3 and 4, respectively. Unweighted mean squared error is used in the design and encoding phases of the vector quantiser. Due to computational constraints, we have only been able to train VQ codebooks up to 16 bits. For higher bitrates (18 bits/frame), we have used a VQ codebook that consists of randomly selected vectors from the training set. The resulting codebook will perform suboptimally, hence the spectral distortion will serve as an upper bound.

We notice from Tables 3 and 4 that for vector quantisation, LSFs produce slightly lower spectral distortion than ISFs, which is contrary to the results for scalar quantisation in Section 3.2. This is probably due to the inclusion of the 16th immittance spectral frequency (which is related to the reflection coefficient), in the joint quantisation of the vector. Because this parameter is not really a 'frequency,' unlike the first 15 ISFs, and possesses different quantisation properties and sensitivity characteristics, then using the unweighted mean squared error (which assumes all vector components are similar) as a distance measure may not be optimal. The VQ code-points for the 16th ISF become dependent on (and their locations are more or less constrained by) the other 15 ISFs. Therefore, independent quantisation of the 16th ISF, as is the case with the PDF-optimised scalar quantisers, will generally result in lower spectral distortion.

In order to highlight the different quantisation characteristics of the 16th ISF, Figs. 7a and 7b show the original and reconstructed spectral envelope estimates for the LSF and ISF representation, where the last (16th) parameter has been shifted by 142 Hz. We note the spectral localisation of the distortion caused by a shift of the 16th LSF. On the

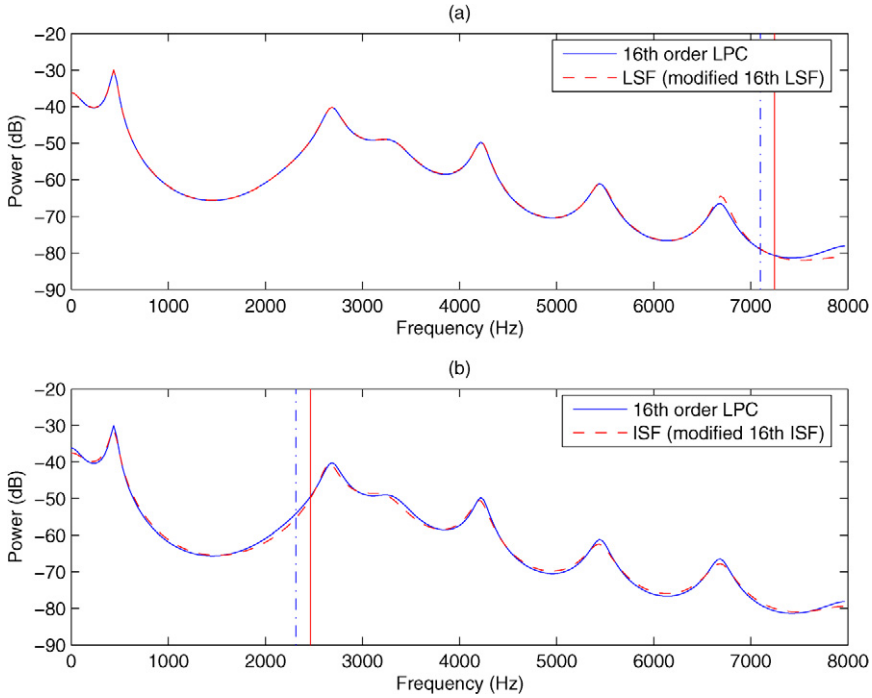

Fig. 7. Original and reconstructed power spectral envelope estimates for 16th-order LPC analysis: (a) Shifting the 16th LSF by 142 Hz (SD = 0.583 dB); (b) Shifting the 16th ISF by 142 Hz (SD = 0.6838 dB). The solid and dashed vertical lines show the original and shifted parameters (LSF and ISF), respectively.

other hand, because the 16th ISF is not really a 'frequency,' but is essentially related to the 16th reflection coefficient, a shift of 142 Hz results in distortion appearing throughout the entire spectrum. The average spectral distortion of the modified LSF power spectral density is also less than that of the modified ISF power spectral density (0.583 cf. 0.684 dB).

Figures 8a and 8b shows the operating distortion–rate (D–R) curves of the unconstrained vector quantisation of LSFs and ISFs, respectively. The squares represent the performance of the vector quantiser that is properly trained using the LBG algorithm while the triangles represent the performance of the vector quantiser with a codebook formed from randomly picked training vectors. We can see that at low bitrates, the D–R curve is exponential-like, while at high bitrates, the curve is more linear. If we make the loose assumption of a linear D–R curve from 13 bits/frame and onwards, then a least-squares linear regression (the line in Fig. 8) shows that single frame, vector quantisers need at least 35 and 36 bits/frame to achieve a spectral distortion of 1 dB, for LSFs and ISFs, respectively. We should note that this is by no means a tight lower bound for several reasons. First, the high rate linearity assumption of the D–R curve is rather loose as there are not enough operating points to determine such a trend. Also, due to the finite and limited number of training vectors, the vector quantiser becomes more and more 'over-trained.' Furthermore, the LBG algorithm generally produces codebooks that are locally optimal, depending on the initialisation used. Finally, the distortion performance of a VQ codebook consisting of randomly selected training vectors, should generally be worse than an LBG trained one. This is demonstrated in Fig. 8a at 15 bits/frame, where there is a 0.5 dB spectral distortion difference between the randomly selected codebook and LBG trained codebook. Therefore, 35 and 36 bits/frame can be said to be informal lower bounds only, for the transparent coding of LSFs and ISFs, respectively.

3.4. Split-multistage vector quantisers with MA predictor

The split-multistage vector quantiser (S-MSVQ), which is shown in Fig. 9, is used in the AMR-WB speech coder for coding immittance spectral frequencies. It can be seen that it is essentially a modified multistage vector quantiser with split vector quantiser stages. This combination of product code vector quantisers allows for a large reduction in

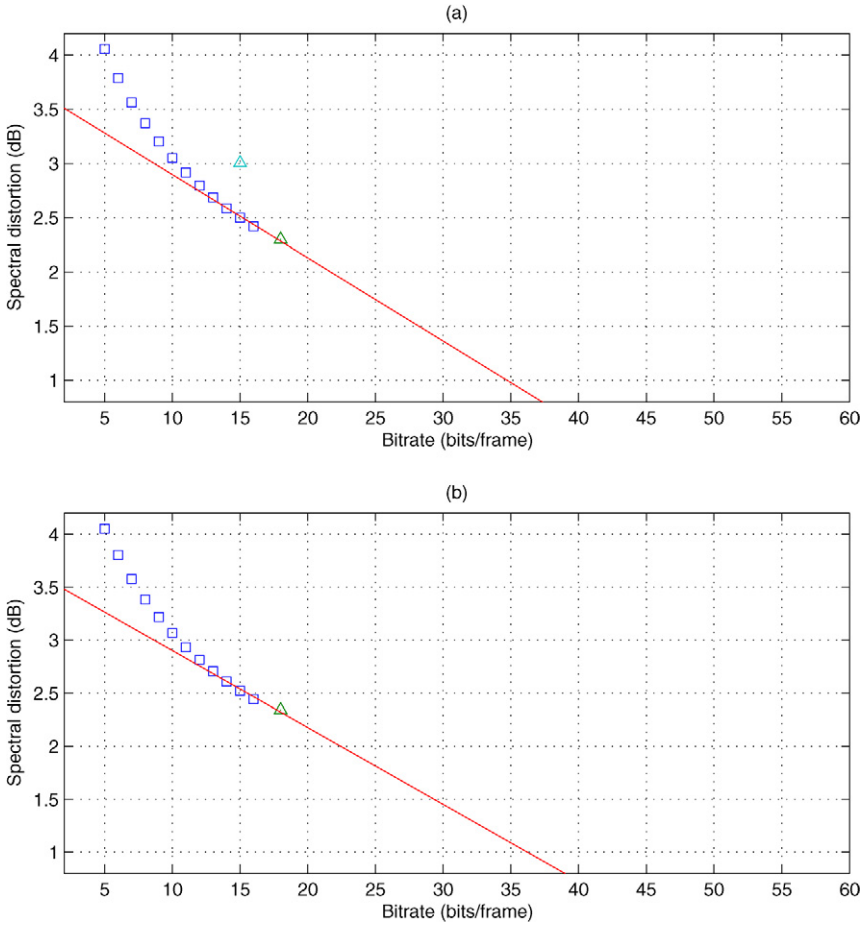

Fig. 8. Extrapolating from the operating distortion-rate curve of the unconstrained vector quantisers to approximate a lower bound for transparent coding, when using: (a) line spectral frequencies (wideband); (b) immittance spectral frequencies (wideband). □'s indicate the performance of a VQ codebook trained using LBG, while △'s indicate the performance of a VQ codebook consisting of random vectors from the training set.

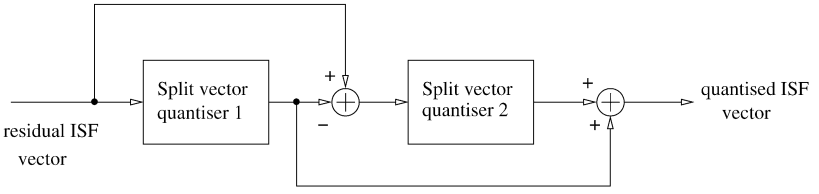

Fig. 9. Block diagram of the split-multistage vector quantiser (after [18]).

computational complexity. During residual ISF encoding, the S-MSVQ is searched using the efficient M–L tree search algorithm [3], where M surviving paths are determined and the path which minimises the distortion is selected.

For the 46 bits/frame S-MSVQ in the AMR-WB speech coder, residual ISF vectors are split into two subvectors of dimension 9 and 7, respectively. Each subvector is then quantised using a two-stage multistage vector quantiser, where in the first stage, each subvector is quantised using 8 bits, and in the second stage, the two residual subvectors are further split into 3 and 2 subvectors, respectively. The bit allocation for each subvector in the second stage are $(6, 7, 7)$ bits and $(5, 5)$ bits [13].

Table 5 shows the spectral distortion performance of the S-MSVQ with MA predictor on ISF vectors from the TIMIT database. The S-MSVQ was adapted from the 3GPP implementation of the AMR-WB speech coder [13] and operates at 36 and 46 bits/frame. We can see that the S-MSVQ at 46 bits/frame, which is used in all bitrate modes

Table 5
Average spectral distortion as a function of bitrate of the AMR-WB (ITU-T G.722.2) split-multistage vector quantiser with MA prediction on wideband ISF vectors from the TIMIT database

Bits/frame	Avg. SD (dB)	Outliers (%)	
		2–4 dB	>4 dB
46	0.894	0.76	0.01
36	1.304	5.94	0.03

Table 6
Average spectral distortion performance of the 16 cluster memoryless, fixed rate GMM-based block quantiser using spectral distortion criterion at different bitrates on wideband ISF vectors from the TIMIT database

Bits/frame	Avg. SD (dB)	Outliers (%)	
		2–4 dB	>4 dB
46	0.879	0.43	0.00
42	1.028	1.27	0.00
41	1.068	1.55	0.01
40	1.109	2.00	0.01
36	1.288	5.37	0.02

Table 7
Average spectral distortion performance of the 16 cluster memoryless, fixed rate GMM-based block quantiser using spectral distortion criterion at different bitrates on wideband LSF vectors from the TIMIT database

Bits/frame	Avg. SD (dB)	Outliers (%)	
		2–4 dB	>4 dB
46	0.844	0.22	0.00
42	0.985	0.66	0.01
41	1.025	0.88	0.01
40	1.064	1.18	0.01
36	1.240	3.51	0.01

except the lowest, achieves a spectral distortion that is lower than 1 dB, though there is a small percentage of outlier frames having a spectral distortion of greater than 4 dB. At 36 bits/frame, which is used in the lowest bitrate mode, we observe a higher spectral distortion and percentage of outlier frames, hence the coding is speech is less likely to be transparent.

The computational complexity and memory requirements of the S-MSVQ at 46 bits/frame are 28.03 kflops/frame and 5280 floats, respectively [37].

3.5. Memoryless GMM-based block quantisers

Subramaniam and Rao [27] applied the GMM-based block quantiser to quantise narrowband speech LSFs. The distortion measure used for the minimum distortion block quantisation was spectral distortion. We can see in Fig. 6 that higher frequency ISFs and LSFs tend to be unimodal, while the lower frequency ones have a multimodal distribution. Therefore, the GMM-based block quantiser is expected to perform well because of the ability to estimate and quantise the multimodal PDFs of the lower frequency LSFs, which are perceptually more important. Also, because of the use of decorrelating transforms, the GMM-based block quantiser is expected to achieve lower spectral distortions than the independent PDF-optimised scalar quantisers at a given bitrate.

Tables 6 and 7 show the spectral distortion performance of the fixed-rate GMM-based block quantiser on ISF and LSF vectors, respectively, from the TIMIT database. The GMM consists of 16 clusters and during the training process, we have used 20 iterations of the EM algorithm. We can observe that a spectral distortion of 1 dB is achieved at 41 bits/frame when using the LSF representation and 42 bits/frame when using the ISF representation. This difference

Table 8
Bitrate independent computational complexity (in kflops/frame) and memory requirements (ROM) of the GMM-based block quantiser using spectral distortion-based quantiser selection as a function of the number of clusters for wideband speech coding

m	kflops/frame	ROM (floats)
4	66.4	1472
8	132.9	2688
16	265.8	5120
32	531.5	9984

in performance is similar to that seen in the vector quantiser from Section 3.4, and is contrary to the scalar quantiser result. Like the vector quantiser, the GMM-based block quantiser exploits correlation within each vector and tends to lose performance if some of the vector components are rather lightly correlated. Because the 16th ISF is a different type of parameter from the first 15 ISFs, this fact may explain the performance loss of using the ISF representation in joint vector and block quantisation schemes.

Comparing Tables 7 and 2, we can see that the 41 bits/frame GMM-based block quantiser is comparable in spectral distortion performance to a 59 bits/frame PDF optimised scalar quantiser. Both schemes utilise scalar codebooks that are optimised for the PDF of each vector component. Therefore, it is apparent that the saving of up to 18 bits/frame by using the GMM-based block quantiser is mostly due to the exploitation of correlation within LSF frame by the multiple Karhunen–Loève transforms.

Comparing Tables 6 and 5, we can see that the GMM-based block quantiser achieves slightly less spectral distortion and outlier frames than the S-MSVQ with MA predictor at 46 and 36 bits/frame. Because the GMM-based block quantiser is a single frame scheme and does not exploit interframe dependencies, then we can conclude that it is more efficient, in the rate distortion sense, than the S-MSVQ (without MA predictor).

We determined an informal lower bound of 35 bits/frame for the transparent coding of LSFs in Section 3.3. Therefore, we can say that the GMM-based block quantiser performs about 6 bits/frame worse than the unconstrained vector quantiser, when transparently coding LSFs, though it must be stressed that the lower bound is by no means a tight one.

Table 8 shows the bitrate independent computational complexity and memory requirements of the GMM-based block quantiser for different numbers of clusters. The spectral distortion calculation accounts for a large part of the complexity, requiring at least 15.3 kflops/frame (assuming a 256 point FFT). However, the memory requirements of the GMM-based block quantiser are relativity low.

3.6. Split vector quantisers

Tables 9 and 10 show the average spectral distortion, computational complexity, and memory requirements of a five-part split vector quantiser (SVQ) on ISF and LSF vectors, respectively. Split vector quantisers were first applied to narrowband LPC parameter quantisation by Paliwal and Atal [2]. Vectors of dimension 16 were split in the following way: $(3, 3, 3, 3, 4)$. We can see that the five-part SVQ requires 46 bits/frame to achieve transparent coding for LSFs and 47 bits/frame for ISFs. Again, we observe better performance when quantising LSFs, amounting to a 1 bit/frame difference.

Comparing these results with those of the PDF-optimised scalar quantisers, we observe that SVQ achieves 1 dB of spectral distortion at a lower bitrate. This may be attributed to the advantages of vector quantising each of the five parts, which exploits properties such as memory as well as optimal PDF and quantiser cell shape [38]. On the other hand, scalar quantisation is equivalent to a 16-part SVQ which, due to the high degree of structural constraints, has poorer rate-distortion efficiency.

Finally, by comparing the SVQ with the memoryless GMM-based block quantiser, we see that the latter achieves lower spectral distortion for the same bitrate. This may be attributed to the ability of the GMM-based block quantiser to exploit full vector correlation via the Karhunen–Loève transform. While on the other hand due to the vector splitting, the SVQ cannot exploit memory that exists between vector components from different parts, and this impacts on its rate-distortion performance. SVQ may be improved in this respect by splitting vectors into less parts but this increases the computational burden.

Table 9
Average spectral distortion (SD), computational complexity, and memory requirements (ROM) of the five-part split vector quantiser as a function of bitrate on wideband ISF vectors from the TIMIT database

Bits/frame	Avg. SD (dB)	Outliers (%)		kflops/frame	ROM (floats)
		2–4 dB	>4 dB		
47	0.997	0.70	0.00	47.10	11,776
46	1.030	0.88	0.00	40.96	10,240
45	1.070	1.21	0.00	32.76	8192
44	1.106	1.32	0.00	29.69	7424
43	1.168	2.13	0.00	26.62	6656

Table 10
Average spectral distortion (SD), computational complexity, and memory requirements (ROM) of the five-part split vector quantiser as a function of bitrate on wideband LSF vectors from the TIMIT database

Bits/frame	Avg. SD (dB)	Outliers (%)		kflops/frame	ROM (floats)
		2–4 dB	>4 dB		
46	1.012	0.68	0.00	40.96	10,240
45	1.061	0.99	0.00	32.76	8192
44	1.092	1.10	0.00	29.69	7424
43	1.151	1.70	0.00	26.62	6656
42	1.200	2.31	0.00	23.55	5888

Table 11
Average spectral distortion (SD), computational complexity, and memory requirements (ROM) of the five-part switched split vector quantiser using unweighted MSE as a function of bitrate and number of switch directions of wideband LSF vectors from the TIMIT database

m	Total bits/frame $(b_1 + b_2 + b_3 + b_4 + b_5 + b_m)$	Avg. SD (dB)	Outliers (%)		kflops/frame	ROM (floats)
			2–4 dB	>4 dB		
8	46 $(8+8+9+9+9+3)$	0.919	0.54	0.00	27.13	53,376
	45 $(8+8+8+9+9+3)$	0.953	0.64	0.00	24.06	47,232
	44 $(8+8+8+8+9+3)$	0.984	0.79	0.00	20.99	41,088
	43 $(7+8+8+8+9+3)$	1.018	0.90	0.00	19.45	38,016
	42 $(7+7+8+8+9+3)$	1.066	1.37	0.00	15.35	34,944
16	46 $(8+8+8+9+9+3)$	0.903	0.48	0.00	24.57	94,464
	45 $(8+8+8+8+9+3)$	0.932	0.60	0.00	21.50	82,176
	44 $(7+8+8+8+9+3)$	0.964	0.73	0.00	19.96	76,032
	43 $(7+7+8+8+9+3)$	1.007	0.97	0.00	18.43	69,888
	42 $(6+7+8+8+9+3)$	1.050	1.19	0.01	17.66	66,816

3.7. Switched split vector quantisers

3.7.1. Using the unweighted mean squared error as the distance measure

Table 11 shows the average spectral distortion, computational complexity, and memory requirements of the five-part switched split vector quantiser (SSVQ) on wideband LSF vectors from the TIMIT database. Also shown are the bit allocations used, which were determined experimentally to result in minimal spectral distortion, while maintaining moderate complexity, with b_m being the number of bits that were given to the switch vector quantiser. The 16-dimensional vectors are split into 5 subvectors of sizes $(3, 3, 3, 3, 4)$. An unweighted mean squared error was used as the distance measure for codebook training and searching. We can see that the 8 switch, five-part SSVQ can achieve transparent coding at 43 bits/frame with a moderate computational complexity (19.45 kflops/frame). By using more switches (16 switches), the SSVQ achieves slightly lower spectral distortions and complexity, which is offset by a large increase in memory requirements.

Comparing the performance of the SSVQ with that of SVQ in Tables 9 and 10, we can see that the benefits of using the initial switch vector quantiser are two-fold. First, SSVQ achieves better rate-distortion efficiency since the

Table 12
Average spectral distortion (SD), computational complexity, and memory requirements (ROM) of the five-part switched split vector quantiser using unweighted MSE as a function of bitrate and number of switch directions of wideband ISF vectors from the TIMIT database

m	Total bits/frame $(b_1+b_2+b_3+b_4+b_5+b_m)$	Avg. SD (dB)	Outliers (%)		kflops/frame	ROM (floats)
			2–4 dB	>4 dB		
8	46 (8+8+9+9+9+3)	0.931	0.53	0.00	27.13	53,376
	45 (8+8+8+9+9+3)	0.968	0.86	0.00	24.06	47,232
	44 (8+8+8+8+9+3)	0.999	1.07	0.00	20.99	41,088
	43 (7+8+8+8+9+3)	1.037	1.21	0.00	19.45	38,016
	42 (7+7+8+8+9+3)	1.080	1.68	0.00	15.35	34,944
16	46 (8+8+8+9+9+3)	0.920	0.63	0.00	24.57	94,464
	45 (8+8+8+8+9+3)	0.948	0.77	0.00	21.50	82,176
	44 (7+8+8+8+9+3)	0.983	0.87	0.00	19.96	76,032
	43 (7+7+8+8+9+3)	1.032	1.30	0.00	18.43	69,888
	42 (6+7+8+8+9+3)	1.078	1.57	0.01	17.66	66,816

switch VQ allows the exploitation of full vector dependencies. And second, the tree-structured nature of the switch VQ provides for lower search complexity.

Comparing the performance of the SSVQ with that of the S-MSVQ with MA predictor in Table 5, we can see that the SSVQ achieves slightly higher spectral distortion (0.919 cf. 0.894 dB) at 46 bits/frame, though the number of outlier frames having a spectral distortion of 2 and 4 dB is less (0.54 cf. 0.76%) than that of the S-MSVQ. These results are to be expected since the S-MSVQ scheme is an interframe scheme which uses an MA predictor, while the SSVQ operates on single frames only and does not exploit interframe correlation. Prediction-based schemes tend to achieve lower distortion at a given bitrate with a higher percentage of outlier frames, due to the inability of predictors to capture rapid changes [14]. In terms of computational complexity at 46 bits/frame, the 8-switch SSVQ is similar to the S-MSVQ in the AMR-WB (27.13 cf. 28.03 kflops/frame). However, the SSVQ requires a larger amount of memory.

Comparing Tables 11 and 7, we can see that the GMM-based block quantiser achieves lower spectral distortion than the SSVQ, which is in contrast to narrowband LSF quantisation, where the SSVQ was found to be better [24]. This is because of the larger dimension and bitrates in wideband LSF quantisation. That is, in order to keep the complexity manageable, the SSVQ splits vectors into five subvectors. Splitting vectors into more and more parts, while dramatically reducing the computational complexity and memory requirements, also reduces all three vector quantiser advantages (space filling, shape, and memory) [39], which penalises the rate-distortion performance of the vector quantisation scheme. However, comparing the computational complexity of the SSVQ with that of the GMM-based block quantiser (in Table 8), we can see that the former requires only 7.3% of the complexity of the latter, at 43 bits/frame. Therefore, the SSVQ is more computationally efficient than the GMM-based block quantiser.

Table 12 shows the average spectral distortion of the five-part switched split vector quantiser on wideband ISF vectors from the TIMIT database. The same vector splitting (3, 3, 3, 3, 4) was applied and an unweighted mean squared error was used as the distance measure. We can see that the 8 switch, five-part SSVQ achieves transparent coding at 44 bits/frame for wideband ISFs. Comparing the spectral distortions with those in Table 11, we notice that, as observed previously with the GMM-based block quantiser and unconstrained vector quantiser, ISFs perform slightly worser than LSFs, which can amount to a 1 bit/frame difference.

3.7.2. Using the weighted mean squared error as the distance measure

It is well known that each quantised LSF[6] influences the reconstructed power spectral envelope in different ways, depending on its location. Figure 10 shows the original and reconstructed spectral envelopes when two different LSFs have been shifted. The 15th LSF falls on top of a formant while the 4th LSF is located in a spectral valley. We can see from the figure that a shift in the 15th LSF results in a shifting of the formant that is more dramatic than the distortion

[6] We have observed the first 15 ISFs to possess the same properties, hence a similar weighted Euclidean distance measure can be applied. However, the 16th ISF does not possess spectral error localisation properties, but rather, shifting this 'frequency' affects the entire spectrum. Therefore, direct application of a weighted Euclidean distance measure is not as straightforward.

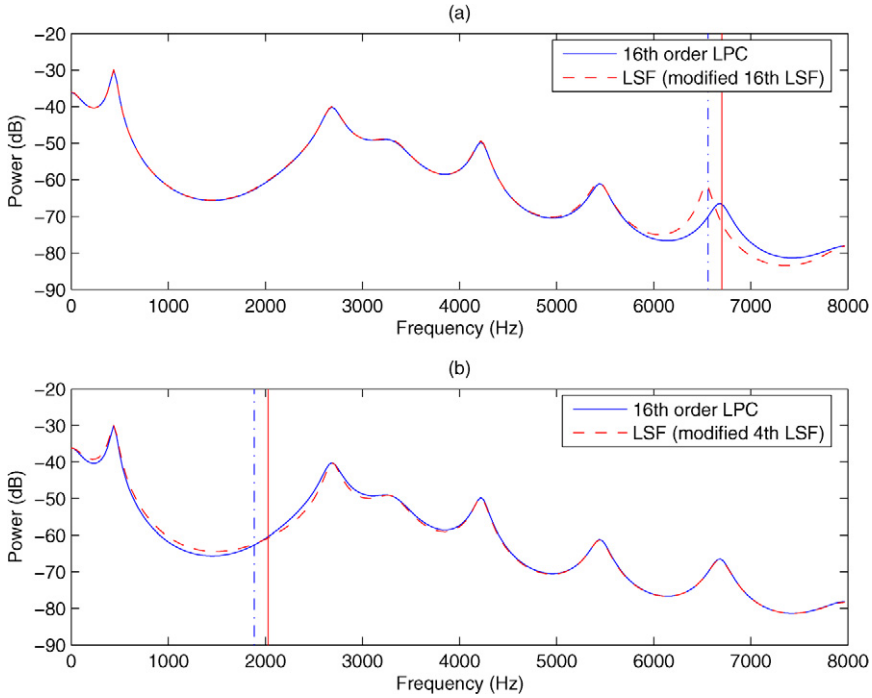

Fig. 10. Original and reconstructed spectral envelope for a 16th-order LPC analysis: (a) shifting the 15th LSF (SD = 1.7475 dB); (b) shifting the 4th LSF (SD = 0.5441 dB). The solid and dashed vertical lines show the original and shifted LSFs, respectively.

caused by a similar shift to the 4th LSF. Therefore, we can say that 'not all LSFs are considered equal,' and a weighted distance measure that finely quantises the LSFs that are located in the vicinity of a spectral peak, should result in less spectral distortion, as shown in Fig. 10.

We have adopted a partial implementation of the weighted mean squared error measure introduced by Paliwal and Atal [2] in narrowband LSF quantisation, which applies a dynamic weighting (that emphasises LSFs in strong regions of the power spectral density) as well as a fixed weighting (which accounts for the difference in sensitivity of the human ear). In this work, our weighted distance measure implements only the dynamic weighting and is used in both the training and searching of the VQ codebooks.

The weighted distance measure, $d_{\mathrm{w}}(\boldsymbol{f}, \hat{\boldsymbol{f}})$, between the original vector, $\boldsymbol{f}$, and the approximated vector, $\hat{\boldsymbol{f}}$, is defined as [2]

$$d_{\mathrm{w}}(\boldsymbol{f}, \hat{\boldsymbol{f}}) = \sum_{i=1}^{16} \left[w_i \left(f_i - \hat{f}_i \right) \right]^2, \tag{16}$$

where f_i and $\hat{f}_i$ are the ith LSF in the original and approximated vector respectively. The dynamic weights, $\{w_i\}$ are given by [2]

$$w_i = \left[P(f_i) \right]^r, \tag{17}$$

where $P(f)$ is the LPC power spectral density and r is a constant (typical value used is 0.15).

Table 13 shows the average spectral distortion performance of the five-part SSVQ using a weighted MSE. By comparing these results with the unweighted MSE-based scheme (Table 11), we can see that by using a weighted distance measure, which emphasises specific LSFs that are located near the formant peaks, the SSVQ results in lower spectral distortions and percentages of outlier frames. The SSVQ with weighted MSE can achieve transparent coding at 42 bits/frame. In comparison, the memoryless, five-part split vector quantiser reported by Biundo et al. [18] required 45 bits/frame for transparent coding.

Comparing Tables 13 and 5, we can see that the SSVQ with weighted MSE, which is a memoryless scheme, achieves comparable spectral distortion performance to the S-MSVQ with MA predictor scheme from the AMR-WB

Table 13
Average spectral distortion (SD) of the five-part switched split vector quantiser using weighted MSE as a function of bitrate and number of switch directions of wideband LSF vectors from the TIMIT database

m	Total bits/frame $(b_1 + b_2 + b_3 + b_4 + b_5 + b_m)$	Avg. SD (dB)	Outliers (%)	
			2–4 dB	>4 dB
8	46 (8+8+9+9+9+3)	0.889	0.33	0.00
	45 (8+8+8+9+9+3)	0.922	0.43	0.00
	44 (8+8+8+8+9+3)	0.953	0.57	0.00
	43 (7+8+8+8+9+3)	0.986	0.66	0.00
	42 (7+7+8+8+9+3)	1.037	1.05	0.00
16	46 (8+8+8+9+9+3)	0.878	0.34	0.00
	45 (8+8+8+8+9+3)	0.906	0.44	0.00
	44 (7+8+8+8+9+3)	0.936	0.50	0.00
	43 (7+7+8+8+9+3)	0.975	0.64	0.00
	42 (6+7+8+8+9+3)	1.018	0.83	0.00

Table 14
Average spectral distortion as a function of bitrate and number of concatenated frames, p, of the 16 cluster multi-frame GMM-based block quantiser on wideband LSF vectors from the TIMIT database

p	Bits/frame	Avg. SD (dB)	Outliers (%)	
			2–4 dB	>4 dB
2	46	0.754	0.09	0.00
	42	0.881	0.29	0.00
	40	0.951	0.52	0.00
	39	0.991	0.69	0.00
	38	1.028	0.91	0.00
	37	1.067	1.23	0.00
3	46	0.725	0.07	0.00
	42	0.845	0.18	0.00
	40	0.911	0.37	0.00
	39	0.946	0.49	0.00
	38	0.983	0.65	0.00
	37	1.021	0.84	0.00
	36	1.060	1.15	0.00
4	46	0.713	0.05	0.00
	42	0.831	0.14	0.00
	40	0.897	0.26	0.00
	39	0.931	0.36	0.00
	38	0.967	0.47	0.00
	37	1.004	0.61	0.00
	36	1.042	0.86	0.00
5	46	0.711	0.02	0.00
	42	0.830	0.10	0.00
	40	0.895	0.18	0.00
	39	0.930	0.28	0.00

speech coder, at 46 bits/frame. In addition, the SSVQ with weighted MSE has produced only half the number of outlier frames than the S-MSVQ, which is to be expected, since the latter has a predictive component.

3.8. Multi-frame GMM-based block quantisers

3.8.1. Spectral distortion performance when using 16 clusters

Table 14 shows the average spectral distortion of the 16 cluster, multi-frame GMM-based block quantiser at varying bitrates and number of concatenated frames, p. Unweighted mean squared error is used for the cluster quantiser selection. Table 16 shows the bitrate independent computational complexity and memory requirements of the multi-

Table 15
Average spectral distortion as a function of bitrate and number of concatenated frames, p, of the 16 cluster multi-frame GMM-based block quantiser on wideband ISF vectors from the TIMIT database

p	Bits/frame	Avg. SD (dB)	Outliers (%)	
			2–4 dB	>4 dB
2	46	0.781	0.16	0.00
	42	0.910	0.45	0.00
	40	0.983	0.80	0.00
	39	1.021	1.02	0.00
	38	1.060	1.30	0.00
	37	1.100	1.77	0.00
3	46	0.753	0.11	0.00
	42	0.879	0.32	0.00
	40	0.946	0.56	0.00
	39	0.982	0.78	0.00
	38	1.019	1.01	0.00
	37	1.058	1.35	0.00
	36	1.097	1.74	0.00
4	46	0.743	0.07	0.00
	42	0.868	0.24	0.00
	40	0.935	0.46	0.00
	39	0.972	0.56	0.00
	38	1.010	0.81	0.00
	37	1.049	1.08	0.00
	36	1.087	1.40	0.00
5	46	0.744	0.06	0.00
	42	0.867	0.20	0.00
	40	0.934	0.34	0.00
	39	0.970	0.51	0.00

Table 16
Bitrate independent computational complexity (in kflops/frame) and memory requirements (ROM) of multi-frame GMM-based block quantiser as a function of the number of concatenated frames, p and number of clusters, m

m	p	kflops/frame	ROM (floats)
16	1	22.29	5120
	2	38.66	18,176
	3	55.05	39,424
	4	71.43	68,864
	5	87.81	106,496
32	1	44.58	9984
	2	77.33	36,096
	3	110.10	78,593
	4	142.90	137,472
	5	175.60	212,736

frame GMM-based block quantiser. Comparing this with Table 8, we can see that despite the larger dimensionality, the multi-frame GMM-based block quantiser is still computationally more efficient than the single frame GMM-based block quantiser of [27]. This is due to the replacement of the spectral distortion calculation in the cluster quantiser selection with the mean squared error, which is computationally less complex. However, a large delay is introduced because of the multi-frame nature of this scheme, hence it is not suitable for two-way communications.

When quantising two frames jointly ($p = 2$), transparent coding is achieved at 39 bits/frame. By comparing this with the memoryless GMM-based block quantiser in Table 7, where transparent coding was achieved at 41 bits/frame, the saving of 2 bits/frame by the former may be attributed to the exploitation of correlation between successive pairs of frames. Also, there is a drop in the percentage of outlier frames having spectral distortion between 2 and 4 dB. The

Table 17
Average spectral distortion as a function of bitrate and number of concatenated frames, p, of the 32 cluster multi-frame GMM-based block quantiser on wideband LSF vectors from the TIMIT database

p	Bits/frame	Avg. SD (dB)	Outliers (%)	
			2–4 dB	>4 dB
2	46	0.728	0.07	0.00
	42	0.850	0.21	0.00
	40	0.919	0.34	0.00
	39	0.955	0.44	0.00
	38	0.992	0.62	0.00
	36	1.069	1.13	0.00
3	46	0.700	0.02	0.00
	42	0.817	0.12	0.00
	40	0.882	0.22	0.00
	39	0.916	0.31	0.00
	38	0.951	0.40	0.00
	37	0.987	0.54	0.00
	36	1.026	0.77	0.00
4	46	0.693	0.02	0.00
	42	0.810	0.09	0.00
	40	0.873	0.18	0.00
	39	0.910	0.28	0.00
	38	0.942	0.35	0.00
	37	0.979	0.48	0.00
	36	1.015	0.62	0.00

$p = 3$ scheme has a moderate tradeoff between distortion and complexity, as shown in Table 16, where transparent coding is achieved at 37 bits/frame. As more frames are concatenated, the average spectral distortions and number of outliers decrease, though the benefit of joint quantisation starts to diminish for $p > 4$.

Table 15 shows the average spectral distortion performance of the multi-frame GMM-based block quantiser on wideband ISF vectors. It can be observed that the spectral distortions are slightly higher than those in the LSF quantiser. For $p = 2$ and 3, 40 and 38 bits/frame are required for transparent coding of ISFs, which is 1 bit/frame more than for LSFs.

Comparing Tables 15 and 5, we can see that the multi-frame GMM-based block quantiser outperforms the S-MSVQ from the AMR-WB speech coder in all bitrates. It is particularly interesting to compare the $p = 2$ multi-frame GMM-based block quantiser with S-MSVQ with MA predictor since both these schemes exploit memory across two consecutive frames, where we can see that the former achieves a spectral distortion that is 0.11 dB lower than the latter at 46 bits/frame. In terms of computational complexity and memory requirements, the two-frame GMM-based block quantiser requires approximately 11 kflops/frame more than the S-MSVQ and three times more memory.

3.8.2. Spectral distortion performance when using 32 clusters

Table 17 shows the average spectral distortion for the 32 cluster, multi-frame GMM-based block quantiser on wideband LSF vectors. Comparing with Table 14, we note that the spectral distortion and percentage of outliers are lower. This may be attributed to more accurate modelling of the PDF by using more clusters in the GMM. As we can see from Table 16, the computational and memory requirements of the 32 cluster scheme are much higher than those of the 16 cluster one.

4. Conclusions

In this paper, we have reviewed wideband speech coding in general and evaluated several LPC parameter quantisation schemes used in CELP-based wideband speech coders. We have also compared the performance of these schemes on the two LPC parameter representations: line spectral frequencies (LSFs) and immittance spectral pairs (ISPs). We have found that in schemes where each vector component is independently quantised, using the ISP representation incurs less spectral distortion than the LSF representation, for the same bitrate, with up to a 1 bit/frame difference at the

point of transparent coding. However, for schemes that quantise vector components jointly, such as vector quantisers and transform coders, we have found the LSF representation to be superior. This may be attributed to the reflection coefficient in the ISP representation, which has different quantisation sensitivities than the other frequencies, and is therefore not suitable for joint quantisation. Through linear extrapolation of the operating distortion-rate curve of an unconstrained vector quantiser, we derived an informal lower bound of 35 and 36 bits/frame for achieving transparent coding using LSFs and ISPs, respectively. Two new quantisation schemes were also presented in this paper. The switched split vector quantiser with dynamically-weighted distance measure and multi-frame GMM-based block quantiser require 42 and 37 bits/frame for the transparent coding of LSFs, respectively.

Acknowledgment

We thank the two anonymous reviewers for their detailed comments, which have been extremely helpful in improving the clarity and quality of this paper.

References

[1] F. Itakura, Line spectrum representation of linear predictive coefficients of speech signals, J. Acoust. Soc. Amer. 57 (1975) S35.
[2] K.K. Paliwal, B.S. Atal, Efficient vector quantization of LPC parameters at 24 bits/frame, IEEE Trans. Speech Audio Process. 1 (1) (1993) 3–14.
[3] W.P. LeBlanc, B. Bhattacharya, S.A. Mahmoud, V. Cuperman, Efficient search and design procedures for robust multi-stage VQ for LPC parameters for 4 kb/s speech coding, IEEE Trans. Speech Audio Process. 1 (4) (1993) 373–385.
[4] J.-P. Adoul, R. Lefebvre, Wideband speech coding, in: W.B. Kleijn, K.K. Paliwal (Eds.), Speech Coding and Synthesis, Elsevier, Amsterdam, 1995, pp. 289–309.
[5] J.W. Paulus, J. Schnitzler, 16 kbits/s wideband speech coding based on unequal subbands, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 1996, pp. 255–258.
[6] B. Bessette, R. Salami, R. Lefebvre, M. Jelínek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, K. Järvinen, The adaptive multirate wideband speech codec (AMR-WB), IEEE Trans. Speech Audio Process. 10 (8) (2002) 620–636.
[7] F. Itakura, S. Saito, Speech analysis-synthesis based on the partial autocorrelation coefficient, in: Proc. JSA, 1969, pp. 199–200.
[8] A. Gray, J. Markel, Quantization and bit allocation in speech processing, IEEE Trans. Acoust. Speech Signal Process. ASSP-24 (1976) 459–473.
[9] R. Viswanathan, J. Makhoul, Quantization properties of transmission parameters in linear predictive systems, IEEE Trans. Acoust. Speech Signal Process. ASSP-23 (1975) 309–321.
[10] F.K. Soong, B.H. Juang, Line spectrum pair (LSP) and speech data compression, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, San Diego, CA, March 1984, pp. 37–40.
[11] N. Sugamura, F. Itakura, Speech analysis and synthesis methods developed at ECL in NTT—from LPC to LSP, Speech Commun. 5 (1986) 199–215.
[12] Y. Bistritz, S. Pellerm, Immittance spectral pairs (ISP) for speech encoding, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 1993, pp. II-9–II-12.
[13] 3rd generation partnership project; Technical specification group services and system aspects; speech codec speech processing functions; AMR wideband speech codec; Transcoding functions (Release 5), Technical Specification TS 26.190, 3rd Generation Partnership Project (3GPP), December 2001.
[14] G. Guibé, H.T. How, L. Hanzo, Speech spectral quantizers for wideband speech coding, European Trans. Telecommun. 12 (6) (2001) 535–545.
[15] E. Harborg, J.E. Knudsen, A. Fuldseth, F.T. Johansen, A real-time wideband CELP coder for a videophone application, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 1994, pp. 121–124.
[16] R. Lefebvre, R. Salami, C. Laflamme, J.-P. Adoul, High quality coding of wideband audio signals using transform coded excitation (TCX), in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 1994, pp. 193–196.
[17] J.H. Chen, D. Wang, Transform predictive coding of wideband speech signals, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 1996, pp. 275–278.
[18] G. Biundo, S. Grassi, M. Ansorge, F. Pellandini, P.A. Farine, Design techniques for spectral quantization in wideband speech coding, in: Proc. of 3rd COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication, Budapest, 2002, pp. 114–119.
[19] G. Roy, P. Kabal, Wideband CELP speech coding at 16 kbits/s, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 1991, pp. 17–20.
[20] P. Combescure, J. Schnitzler, K. Fischer, R. Kirchherr, C. Lamblin, A. le Guyader, D. Massaloux, C. Quinquis, J. Stegmann, P. Vary, A 16, 24, 32 kbits/s wideband speech codec based on ATCELP, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 1999, pp. 5–8.
[21] A. Ubale, A. Gersho, A multi-band CELP wideband speech coder, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 1994, pp. 1367–1370.
[22] Y. Shin, S. Kang, T.R. Fischer, C. Son, Y. Lee, Low-complexity predictive trellis coded quantization of wideband speech LSF parameters, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 2004, pp. 145–148.
[23] E.R. Duni, A.D. Subramaniam, B.D. Rao, Improved quantization structures using generalised HMM modelling with application to wideband speech coding, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, May 2004, pp. 161–164.

[24] S. So, K.K. Paliwal, Efficient vector quantisation of line spectral frequencies using the switched split vector quantiser, in: Proc. Int. Conf. Spoken Language Processing, October 2004.
[25] S. So, K.K. Paliwal, Efficient product code vector quantisation using the switched split vector quantiser, Digital Signal Process. 17 (2007) 138–171.
[26] A. Gersho, R.M. Gray, Vector Quantization and Signal Compression, Kluwer Academic, MA, 1992.
[27] A.D. Subramaniam, B.D. Rao, PDF optimized parametric vector quantization of speech line spectral frequencies, IEEE Trans. Speech Audio Process. 11 (2) (2003) 130–142.
[28] S. So, K.K. Paliwal, Multi-frame GMM-based block quantisation of line spectral frequencies, Speech Commun. (2005), in press.
[29] S. So, K.K. Paliwal, Multi-frame GMM-based block quantisation of line spectral frequencies for wideband speech coding, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, Philadelphia, March 2005, pp. 121–124.
[30] Y. Linde, A. Buzo, R.M. Gray, An algorithm for vector quantizer design, IEEE Trans. Commun. COM-28 (1) (1980) 84–95.
[31] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Statist. Soc. 39 (1977) 1–38.
[32] J.J.Y. Huang, P.M. Schultheiss, Block quantization of correlated Gaussian random variables, IEEE Trans. Commun. Syst. CS-11 (1963) 289–296.
[33] A.D. Subramaniam, W.R. Gardner, B.D. Rao, Low complexity recursive coding of spectrum parameters, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 2002, pp. I-637–I-640.
[34] J. Samuelsson, P. Hedelin, Recursive coding of spectrum parameters, IEEE Trans. Speech Audio Process. 9 (5) (2001) 492–502.
[35] F.K. Soong, B.H. Juang, Optimal quantization of LSP parameters, in: Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, New York, April 1988, pp. 394–397.
[36] K.K. Paliwal, W.B. Kleijn, Quantization of LPC parameters, in: W.B. Kleijn, K.K. Paliwal (Eds.), Speech Coding and Synthesis, Elsevier, Amsterdam, 1995, pp. 443–466.
[37] S. Kang, Y. Shin, T.R. Fischer, Low-complexity predictive Trellis-coded quantization of speech line spectral frequencies, IEEE Trans. Signal Process. 52 (7) (2004) 2070–2079.
[38] J. Makhoul, S. Roucos, H. Gish, Vector quantization in speech coding, Proc. IEEE 73 (1985) 1551–1588.
[39] T.D. Lookabaugh, R.M. Gray, High-resolution quantization theory and the vector quantizer advantage, IEEE Trans. Inform. Theory 35 (5) (1989) 1020–1033.

Stephen So was born in Hong Kong in 1977. He received the BEng (Hons) in 1999 and the PhD degree in 2005, both from the School of Microelectronic Engineering, Griffith University, Australia. He is currently working as an Associate Lecturer in electronic engineering in the School of Microelectronic Engineering, Griffith University, Australia. His research interests include block and vector quantisation schemes, image coding, speech coding, audio coding, distributed speech recognition, and face recognition.

Kuldip K. Paliwal was born in Aligarh, India, in 1952. He received the BS degree from Agra University, Agra, India, in 1969, the MS degree from Aligarh Muslim University, Aligarh, India, in 1971, and the PhD degree from Bombay University, Bombay, India, in 1978.

He has been carrying out research in the area of speech processing since 1972. He has worked at a number of organizations including Tata Institute of Fundamental Research, Bombay, India; Norwegian Institute of Technology, Trondheim, Norway; University of Keele, UK; AT&T Bell Laboratories, Murray Hill, NJ, USA; AT&T Shannon Laboratories, Florham Park, NJ, USA; and Advanced Telecommunication Research Laboratories, Kyoto, Japan. Since July 1993, he has been a Professor at Griffith University, Brisbane, Australia, in the School of Microelectronic Engineering. His current research interests include speech recognition, speech coding, speaker recognition, speech enhancement, face recognition, image coding, pattern recognition, and artificial neural networks. He has published more than 250 papers in these research areas.

Dr. Paliwal is a Fellow of Acoustical Society of India. He has served the IEEE Signal Processing Society's Neural Networks Technical Committee as a Founding Member from 1991 to 1995 and the Speech Processing Technical Committee from 1999 to 2003. He was an Associate Editor of the IEEE Transactions on Speech and Audio Processing during the periods 1994–1997 and 2003–2004. He also served as Associate Editor of the IEEE Signal Processing Letters from 1997 to 2000. He was the General Co-Chair of the Tenth IEEE Workshop on Neural Networks for Signal Processing (NNSP 2000). He has co-edited two books: "Speech Coding and Synthesis" (published by Elsevier), and "Speech and Speaker Recognition: Advanced Topics" (published by Kluwer). He has received IEEE Signal Processing Society's best (senior) paper award in 1995 for his paper on LPC quantization. He is currently serving the Speech Communication Journal (published by Elsevier) as its Editor-in-Chief.