MULTI-FRAME GMM-BASED BLOCK QUANTISATION OF LINE SPECTRAL FREQUENCIES FOR WIDEBAND SPEECH CODING

Stephen So and Kuldip K. Paliwal

School of Microelectronic Engineering,
Griffith University, Brisbane, Australia, 4111.
s.so@griffith.edu.au, k.paliwal@griffith.edu.au

ABSTRACT

In this paper, we explore the use of the multi-frame GMM-based block quantiser for quantising line spectral frequencies for wideband speech coding. Its main advantages over vector quantisers are bitrate scalability and bitrate independent complexity. By concatenating multiple frames together, interframe correlation can be exploited by the KLT, leading to better quantisation. A saving of up to 3 bits/frame can be achieved by switching the quantiser from memoryless mode to jointly quantising two frames, with only a moderate increase in complexity. This quantisation scheme achieves lower spectral distortion than the split-multistage vector quantiser in the AMR-WB speech codec, with transparent coding at 37 bits/frame.

1. INTRODUCTION

The quantisation of linear predictive coding (LPC) parameters in CELP coders for narrowband speech (200–3400 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]. With the introduction of high-speed data services in wireless communication systems, wideband speech (50–7000 Hz) can now be accommodated [4]. Wideband speech has improved naturalness and intelligibility due to the added bandwidth. However, wideband CELP coders typically require 16 LPC parameters for representing the speech spectral envelope, hence vector quantisers will need to operate at higher bitrates and on vectors of larger dimension.

Harborg *et al.* [5] quantised 16 to 18 log-area-ratio coefficients at 60 to 80 bits/frame using non-uniform scalar quantisers. Lefebvre *et al.* [6] and Chen *et al.* [7] used a seven-part split vector quantiser operating at 49 bits/frame to quantise 16 LSF parameters. Transparent results were reported by Biundo *et al.* [8] for a four and five part split vector quantiser at 45 bits/frame. Because successive LSF frames are highly correlated [9], better quantisation can be achieved by exploiting the interframe correlation. Ubale *et al.* [10] used a seven-stage tree-searched multistage vector quantiser [3] with a moving average (MA) predictor at 28 bits/frame, while Biundo *et al.* [8] reported transparent results using an MA predictive split multistage vector quantiser (S-MSVQ) at 42 bits/frame. Guibé *et al.* [9] achieved transparent coding using a safety-net vector quantiser at 38 bits/frame, while the Adaptive Multi-Rate wideband (AMR-WB) speech codec [4, 11] uses an S-MSVQ with MA predictor at 46 bits/frame. Other quantisation schemes recently reported include the predictive Trellis-coded quantiser [12] and the HMM-based recursive quantiser [13] which achieve a spectral distortion of 1 dB at 34 and 40 bits/frame, respectively.

In our previous work on spectral quantisation for narrowband speech coding [14], we explored the multi-frame Gaussian mixture model-based block quantiser, which is a simple extension of the memoryless scheme of [15]. Compared with vector quantisers, this quantiser is *bitrate scalable* [15]. In other words, the bitrate can be instantly changed without any re-training of the codebooks. Also, the computational and memory requirements of this quantisation scheme remain fixed for all bitrates [15].

In this paper, we investigate the application of the multi-frame GMM-based block quantiser [14] for coding wideband speech LSF frames. We show that as more LSF frames are jointly quantised and more clusters are used, the spectral distortion is decreased for a given bitrate, at the expense of increased complexity and delay. This quantisation scheme can achieve transparent coding at 37 bits/frame with a small number of outliers and moderate complexity.

2. MULTI-FRAME GMM-BASED BLOCK QUANTISER

This quantisation scheme is based on the one proposed by Subramaniam and Rao [15] 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). In the AMR-WB speech codec, autocorrelations are calculated frame-wise with a 5 ms overlap between successive frames [11], hence there will be correlation between these frames. In [14], we proposed a quantisation scheme that used vectors formed from p concatenated frames, in order to exploit interframe correlation. Therefore, if the length of the LSF frame is n, then the dimensions of the vectors processed will be np. For more details on this quantisation scheme, the reader is referred to [15].

2.1. Quantiser training

The PDF model 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 (training) and stored at the encoder and decoder. The bit allocations for different bitrates (described in Section 2.2.1) can be calculated 'on-the-fly' based on the PDF model by both encoder and decoder. Hence this scheme is bitrate scalable [15].

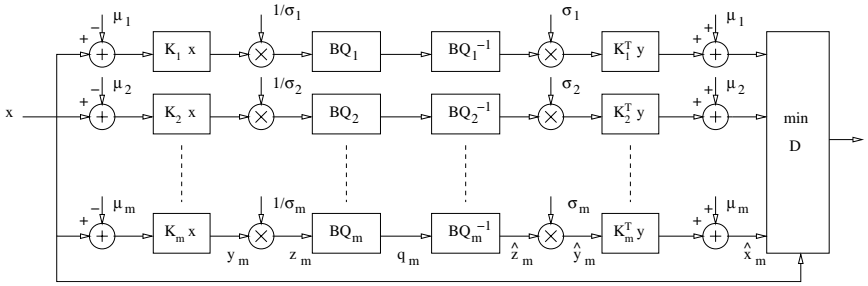

Fig. 1. Block diagram of the GMM-based block quantiser (BQ - block quantiser)

The PDF model, which is in the form of a GMM, is initialised by applying the Linde-Buzo-Gray (LBG) algorithm [16] on the training vectors where m clusters[1] 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 [17] 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 m covariance matrices, producing eigenvalues, $\{\boldsymbol{\lambda}_i\}_{i=1}^{m}$, and eigenvectors. The eigenvectors form the rows of the orthogonal transformation matrix, $\boldsymbol{K}$, of the KLT.

2.2. Encoding process

2.2.1. Bit allocation

Assuming that there are b_{tot} bits available for coding each vector (for an average bitrate of b_{tot}/p bits/frame), these need to allocated to each of the block quantisers for each cluster. The number of bits, b_i, allocated to the block quantiser of cluster i, is given by [15]:

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

where [15]:

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

Then for each block quantiser, the high resolution formula from [18] is used to distribute the b_i bits to each of the vector components:

$$b_{i,j} = \frac{b_i}{np} + \frac{1}{2} \log_2 \frac{\lambda_{i,j}}{\left(\prod_{j=1}^{np} \lambda_{i,j} \right)^{\frac{1}{np}}} \tag{3}$$

$$\text{for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, np$$

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

Table 1. 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	22.03	5120
	2	38.41	18176
	3	54.79	39424
	4	71.17	68864
	5	87.55	106496
32	1	44.06	9984
	2	76.82	36096
	3	109.6	78593
	4	142.3	137472
	5	172.1	212736

2.2.2. Minimum distortion block quantisation

Fig. 1 shows a diagram of minimum distortion block quantisation, which consists of m independent Gaussian block quantisers, BQ_i, each with their own orthogonal matrix, $\boldsymbol{K}_i$, and bit allocations, $\{b_{i,j}\}_{j=1}^{np}$. Each vector, $\boldsymbol{x}$, is coded and decoded by the block quantiser of cluster i to give a reconstructed vector, $\widehat{\boldsymbol{x}}_i$. The distortion between this reconstructed vector and original is then calculated, $d(\boldsymbol{x} - \widehat{\boldsymbol{x}}_i)$. The above procedure is performed for all clusters in the system and the cluster, k, which gives the *minimum distortion* is chosen:

$$k = \underset{i}{\operatorname{argmin}}\, d(\boldsymbol{x} - \widehat{\boldsymbol{x}}_i) \tag{4}$$

We have used the mean-squared-error (MSE) as the distortion criterion due to its low complexity, though a weighted Euclidean distance measure or the high-rate spectral distortion approximation [19] may give better results.

2.2.3. Computational complexity and memory requirements

One of the salient features of the GMM-based block quantiser scheme is the computational complexity and memory requirements being independent of the operating bitrate [15]. Following the analysis given in [15], the complexity (in kflops/frame)[2] and mem-

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

Table 2. Average spectral distortion as a function of bitrate and number of concatenated frames, p, of multi-frame GMM-based block quantiser (16 clusters)

			Outliers (in %)	
p	bits/frame	Avg. SD (dB)	2–4 dB	> 4 dB
1	46	0.844	0.22	0.00
	42	0.985	0.66	0.01
	40	1.064	1.18	0.01
	36	1.240	3.51	0.01
2	46	0.754	0.09	0.00
	42	0.881	0.29	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
	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
	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
	39	0.930	0.28	0.00

ory requirements of the multi-frame GMM-based block quantiser are given in Table 1 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.

3. DISTORTION MEASURES FOR LPC PARAMETERS

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:

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

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 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, and
3. less than 2% of outlier frames are within the range of 2–4 dB.

According to Guibé *et al.* [9], 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.

4. EXPERIMENTAL SETUP

The TIMIT database was used in the training and testing of the multi-frame GMM-based block quantiser where speech is sampled at 16 kHz. We have used the preprocessing and LP analysis of the AMR-WB speech codec (floating point version) [11] to produce linear prediction coefficients which are then converted to line spectral frequency (LSF) representation [1]. The training set consists of 333789 vectors while the evaluation set, consisting of speech not contained in the training, has 85353 vectors.

We have also tested the split-multistage vector quantiser (S-MSVQ) from the AMR-WB speech codec on the database, so that it can be used for comparison. Immittance spectral pairs (ISP) [20] are used in the AMR-WB codec while the quantisation scheme considered in this paper quantises line spectral frequencies (LSF). This presents no problem in our spectral distortion comparisons as (5) requires linear predictive coefficients which can be obtained from ISPs and LSFs.

Table 3. Average spectral distortion as a function of bitrate of split-multistage vector quantiser with MA prediction in AMR-WB speech codec

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

5. RESULTS AND DISCUSSION

Table 3 shows the average spectral distortion of the split-multistage vector quantiser (S-MSVQ) with MA prediction found in the AMR-WB speech codec at 36 and 46 bits/frame. Table 2 shows the average spectral distortion of the 16 cluster, multi-frame GMM-based block quantiser at varying bitrates and number of concatenated frames, p. It can be seen that in its memoryless mode ($p = 1$), this quantisation scheme incurs less spectral distortion and outlier frames than the S-MSVQ at 46 bits/frame and 36 bits/frame, with transparent coding achieved at 42 bits/frame. When quantising 2 frames jointly ($p = 2$), the spectral distortion is roughly 0.1 dB lower, with transparent coding achieved at 39 bits/frame. This saving of 3 bits 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 $p = 3$ scheme has a moderate tradeoff between distortion and complexity, 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 4 shows the average spectral distortion for the 32 cluster, multi-frame GMM-based block quantiser. Comparing with Table 2, 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 1, the computational and memory requirements of the 32 cluster scheme are much higher than those of the 16 cluster one.

6. CONCLUSION

In this paper, we have investigated the use of the multi-frame Gaussian mixture model-based block quantiser for the quantisation of

Table 4. Average spectral distortion as a function of bitrate and number of concatenated frames, p, of multi-frame GMM-based block quantiser (32 clusters)

			Outliers (in %)	
p	bits/frame	Avg. SD (dB)	2–4 dB	$>$ 4 dB
1	46	0.809	0.14	0.00
	42	0.945	0.45	0.00
	41	0.985	0.61	0.00
	40	1.023	0.79	0.00
	36	1.190	2.43	0.00
2	46	0.728	0.07	0.00
	42	0.850	0.21	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
	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
	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

line spectral frequencies for wideband speech coding. By concatenating multiple frames together, correlation between LSFs within each frame and across successive frames can be exploited by the KLT, leading to better quantisation. A saving of up to 3 bits was achieved by switching the quantiser from memoryless mode to jointly quantising two frames. Increasing the number of clusters in the GMM also led to lower spectral distortion and number of outliers. Though these gains come at the expense of an increase in delay and memory requirements, this quantisation scheme is flexible enough for the designer to find the right tradeoff between distortion and complexity.

7. REFERENCES

[1] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals", *J. Acoust. Soc. Amer.*, vol. 57, p. S35, Apr. 1975.

[2] K.K. Paliwal and B.S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame", *IEEE Trans. Speech Audio Processing*, vol. 1, no. 1, pp. 3–14, Jan. 1993.

[3] W.P. LeBlanc, B. Bhattarchaya, S.A. Mahmoud and V. Cuperman, "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding", *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 373–385, Oct. 1993.

[4] B. Bessette, R. Salami, R. Lefebvre, M. Jelínek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Järvinen, "The adaptive multirate wideband speech codec (AMR-WB)", *IEEE Trans. Speech Audio Processing*, vol. 10, no. 8, pp. 620–636, Nov. 2002.

[5] E. Harborg, J.E. Knudsen, A. Fuldseth and 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.

[6] 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.

[7] J.H. Chen and D. Wang, "Transform predictive coding of wideband speech signals", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996, pp. 275–278.

[8] G. Biundo, S. Grassi, M. Ansorge, F. Pellandini and 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, Oct. 2002, pp. 114-119.

[9] G. Guibé, H.T. How and L. Hanzo, "Speech spectral quantizers for wideband speech coding", *European Transactions on Telecommunications*, 12(6), pp. 535–545, 2001.

[10] A. Ubale and A. Gersho, "A multi-band CELP wideband speech coder", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1994, pp. 1367-1370.

[11] "3rd generation partnership project; Technical specification group services and system aspects; Speech codec speech processing functions; AMR wideband speech codec; Transcoding functions", 3GPP TS 26.190.

[12] Y. Shin, S. Kang, T.R. Fischer, C. Son, and 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.

[13] E.R. Duni, A.D. Subramaniam, and 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.

[14] K.K. Paliwal and S. So, "Multiple frame block quantisation of line spectral frequencies using Gaussian mixture models", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2004, pp. 149–152.

[15] A.D. Subramaniam and B.D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies", *IEEE Trans. Speech Audio Processing*, vol. 11, no. 2, pp. 130–142, Mar. 2003.

[16] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design", *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 84–95, Jan. 1980.

[17] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[18] J.J.Y. Huang and P.M. Schultheiss, "Block quantization of correlated Gaussian random variables", *IEEE Trans. Commun. Syst.*, vol. CS-11, pp. 289–296, Sept. 1963.

[19] W.R. Gardner and B.D. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters", *IEEE Trans. Speech Audio Processing*, vol. 3, no. 5, pp. 367–381, Sept. 1995.

[20] Y. Bistritz and S. Pellerm, "Immittance spectral pairs (ISP) for speech encoding", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1993, pp. II-9–II-12.