MULTIPLE FRAME BLOCK QUANTISATION OF LINE SPECTRAL FREQUENCIES USING GAUSSIAN MIXTURE MODELS

Kuldip K. Paliwal and Stephen So

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

ABSTRACT

In this paper, we present a Gaussian mixture model-based block quantiser for coding line spectral frequencies that uses multiple frames and mean squared error as the quantiser selection criterion. The efficiency gained from jointly coding multiple frames permits the use of the mean squared error distortion (MSE) criterion rather than the computationally expensive spectral distortion. The proposed coder encompasses improvements in both distortion performance and complexity with transparency achieved at 23 bits per frame when coding two frames jointly or 21 bits per frame when coding 3 frames.

1. INTRODUCTION

The Code-Excited Linear Predictive (CELP) coder used in low bit rate speech coding requires the quantisation of: (1) the linear predictive coding (LPC) parameters representing the spectral envelope, and (2) the excitation signal representing the fine structure of the speech [3]. With the excitation signal coded using fixed and adaptive codebook vector quantisers (VQ), the focus has been on finding ways to efficiently code the LPC parameters accurately. Direct quantisation of linear prediction (LP) coefficients often leads to problems with unstable synthesis filters due to high sensitivity to errors. In practice, the LP coefficients are converted to line spectral frequencies (LSFs) which have the properties of spectral error localisation and guaranteed stability, and are therefore a better representation of LP coefficients for speech coding [5].

Even though full-search vector quantisers provide optimal performance, the complexity required to accurately represent LSFs, estimated to be in the order of 20 bits, are prohibitively high [6]. Therefore, less complex but suboptimal vector quantisers such as multistage and split VQ have been investigated. It was generally observed that at least 24 bits per frame were required to achieve *transparency* in speech [5]. Matrix quantisation [9] and its derivatives such as split matrix quantisation [10] and multi-mode matrix quantisation [4, 8] perform better by jointly quantising LSF frames in order to exploit interframe correlation.

Recently, a new method of coding LSFs was introduced in [7] which involves the use of a Gaussian mixture model (GMM) to parametrise the probability density function of the source and designing optimised block quantisers. Using this method in its fixed rate mode, transparency was achieved at 24 bits per frame. The main advantages of this method over vector quantisers are: (1) the use of block quantisers which exploit correlation through the use of the Karhunen-Loève transform (KLT) but have lower complexity because of the use of scalar quantisers, (2) bitrate scalability, and (3) the search complexity being independent of the rate of the system. A modified scheme was also proposed that coded the difference between successive frames. 1 dB spectral distortion was achieved at 22 bits per frame [7]. However, the use of spectral distortion (SD) as the criterion for quantiser selection, while giving better performance than mean squared error (MSE), involves more computations.

In this paper, we propose a GMM-based block quantiser that operates on multiple frames and uses the mean squared error (MSE) distortion criterion. We have found this system to perform better than the original single frame, spectral distortion-based coder in terms of both distortion and computational time.

2. 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. For the ith frame, the spectral distortion (in dB), D_i, is defined as:

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

where F_s is the sampling frequency and $P_i(f)$ and $\hat{P}_i(f)$ are the LPC power spectra (in dB) of the coded and uncoded

ith frame, given by:

$$P_i(f) = -20\log_{10}\left|A_i(e^{j2\pi f/F_s})\right| \tag{2}$$

and

$$\hat{P}_i(f) = -20\log_{10}\left|\hat{A}_i(e^{j2\pi f/F_s})\right| \tag{3}$$

where $A_i(z)$ and $\hat{A}_i(z)$ are the original and quantised LPC polynomials of the ith frame respectively [5].

The conditions for transparent speech from LPC parameter quantisation are: (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 [5].

3. GMM-BASED BLOCK QUANTISERS

3.1. PDF Estimation using GMMs and EVD

Gaussian mixture models can be used for modelling any arbitrary distribution using multivariate Gaussians as basis functions [7]. The PDF model, as a mixture of multivariate Gaussians $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu};\boldsymbol{\Sigma})$, can be given by:

$$\begin{aligned} G(\boldsymbol{X}|\boldsymbol{M}) &= \sum_{i=1}^{m} c_i \mathcal{N}_i(\boldsymbol{x};\boldsymbol{\mu}_i;\boldsymbol{\Sigma}_i) && (4) \\ \boldsymbol{M} &= [m, c_i, \ldots, c_m, \mu_1, \ldots, \mu_m, \Sigma_1, \\ &\quad \ldots, \Sigma_m] && (5) \\ \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu};\boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} && (6) \end{aligned}$$

where $\boldsymbol{X}$ are the vectors of transform coefficients, m is the number of mixture components, and n is the dimension of the vectors. $\boldsymbol{M}$ is the set of model parameters consisting of $c_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ which are the weight, mean, and covariance matrix of the ith mixture component respectively. Note the words 'mixture component' and 'cluster' will be used interchangeably in this paper.

The parametric model, represented by parameters, $\boldsymbol{M}$, is initialised by applying the K-means algorithm on the training vectors where m clusters 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. Using the Expectation-Maximisation (EM) algorithm, the maximum-likelihood estimate of the parametric model is computed iteratively until the log likelihood converges, where a final set of means, covariance matrices, and weights are produced.

An eigenvalue decomposition (EVD) is calculated for each of the covariance matrices, producing m sets of eigenvalues, $\{\boldsymbol{\lambda}_i\}_{i=1}^{m}$, and eigenvectors. The eigenvectors form the rows of the orthogonal transformation matrix, $\boldsymbol{K}$, of the KLT.

3.2. Bit Allocation

There are two types of bit allocation that are required: *intercluster bit allocation* and *intracluster bit allocation*. Since the bit allocation is not a computationally expensive procedure, it can be done 'on-the-fly' depending on the chosen bit rate [7].

3.2.1. Intercluster Bit Allocation

With intercluster bit allocation, the number of quantiser levels need to be assigned to each of the m clusters depending upon the covariance and probability of that cluster. If the GMM is viewed a composite Gaussian source where each vector is generated by one of the m clusters, then the cluster weights calculated from the EM algorithm also represent the cluster probabilities [7]. For a fixed-rate quantiser, the total number of quantiser levels is fixed:

$$2^{b_{tot}} = \sum_{i=1}^{m} 2^{b_i} \tag{7}$$

where b_{tot} is the total number of bits in the bit budget, b_i is the number of bits assigned to cluster i, and m is the number of clusters. The average distortion is approximated by [7]:

$$D_{tot} = \sum_{i=1}^{m} c_i D_i(b_i) \tag{8}$$

The high resolution approximation for the distortion of a single Lloyd-Max scalar quantiser from an n-component block quantiser operating on Gaussian sources is given by [7]:

$$\begin{aligned} D_i(b_i) &= nK\Lambda_i 2^{-2\frac{b_i}{n}} && (9) \\ \Lambda_i &= \left(\prod_{j=1}^{n} \lambda_{i,j}\right)^{\frac{1}{n}} && (10) \\ &\text{for } i = 1, 2, \ldots, m \end{aligned}$$

where n is the dimension of the vectors, m is the number of clusters, $\lambda_{i,j}$ is the jth eigenvalue of cluster i, and K is a constant which is approximately equal to $\frac{\pi\sqrt{3}}{2}$ for Gaussian sources [2].

Using Lagrange multipliers, the average distortion can be minimised under the fixed rate constraint of (7), and the following bit allocation formula is derived [7]:

$$\begin{aligned} 2^{b_i} &= 2^{b_{tot}} \frac{(c_i\Lambda_i)^{\frac{n}{n+2}}}{\sum_{i=1}^{m}(c_i\Lambda_i)^{\frac{n}{n+2}}}, && (11) \\ &\text{for } i = 1, 2, \ldots, m \end{aligned}$$

where c_i is the weight of cluster i.

3.2.2. Intracluster Bit Allocation

After the bits are allocated to each cluster, further bit allocation is performed to assign bits to each of the n components. Following the derivation presented in [2], the total number of bits is fixed:

$$b_i = \sum_{j=1}^{n} b_{i,j}, \quad \text{for } i = 1, 2, \ldots, m \tag{12}$$

where $b_{i,j}$ is the number of bits assigned to component j of cluster i. Again, using the high resolution approximation for the distortion of a Lloyd-Max scalar quantiser, the average distortion of cluster i is given by [2]:

$$D_i = \frac{1}{n} \sum_{j=1}^{n} K \lambda_{i,j} 2^{-2b_{i,j}} \tag{13}$$

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

Using Lagrange multipliers, the average distortion is minimised under the fixed rate constraint of (12) and the following bit allocation formula is derived [2]:

$$b_{i,j} = \frac{b_i}{n} + \frac{1}{2} \log_2 \frac{\lambda_{i,j}}{\left(\prod_{j=0}^{n-1} \lambda_{i,j} \right)^{\frac{1}{n}}} \tag{14}$$

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

3.3. Minimum Distortion Block Quantisers

To quantise a vector, $\boldsymbol{x}$, using a particular cluster i, the cluster mean, $\boldsymbol{\mu}_i$, is first subtracted and then a KLT is performed, $\boldsymbol{y}_i = \boldsymbol{K}_i(\boldsymbol{x} - \boldsymbol{\mu}_i)$, where $\boldsymbol{K}_i$ is the transformation matrix of that cluster. The coefficients, $\boldsymbol{y}_i$, are then normalised by the standard deviation, $\boldsymbol{z}_i = \boldsymbol{y}_i / \boldsymbol{\sigma_i}$, and quantised using a Gaussian block quantiser, as described in [2] with its respective bit allocation, $\{b_{i,j}\}_{j=1}^{n}$. The indices, $\boldsymbol{q}_i$, from the quantiser are decoded, multiplied by the standard deviation, $\hat{\boldsymbol{y}}_i = \hat{\boldsymbol{z}}_i \boldsymbol{\sigma}_i$, inverse transformed and the cluster mean is added back, $\hat{\boldsymbol{x}}_i = \boldsymbol{K}_i^T \hat{\boldsymbol{y}}_{\boldsymbol{i}} + \boldsymbol{\mu}_i$. The distortion between this quantised vector and original vector is then calculated, $d(\boldsymbol{x} - \widehat{\boldsymbol{x}}_i)$. The above procedure is performed for all clusters, $i = 1, 2, \ldots m$, in the system. The jth cluster which gives the least distortion, $j = \arg_i \min d(\boldsymbol{x} - \hat{\boldsymbol{x}}_i)$, is chosen and the indices, $\boldsymbol{q}_j$, are transmitted.

4. USING MULTIPLE LSF FRAMES AND MSE DISTORTION CRITERION

A tenth order linear predictive analysis is generally used in CELP coders and thus the dimension of the LSF vectors to be coded is 10. In our multiple frame system, we concatenate p consecutive frames together to form vectors of dimension $n = 10p$. By doing this, the correlation that exists across p consecutive frames can be exploited by the KLT, thus leading to improved performance.

Table 1. Performance of Single Frame Block Quantiser using SD criterion (16 clusters)

		Outliers (in %)	
bits/frame	Avg. SD (dB)	2–4 dB	> 4 dB
23	1.110	1.73	0.00
24	1.043	1.15	0.00
25	0.980	0.79	0.00

The original single frame coder in [7] used spectral distortion as the distortion measure for selecting the appropriate block quantiser. Later, we show that while spectral distortion may be the better distortion measure over MSE for single frames, MSE works just as well for the multiple frame case but with the advantage of lower complexity.

5. EXPERIMENTAL DATABASE

The TIMIT database was used in the training and testing of the multiple frame coder where speech was downsampled to 8 kHz and low pass filtered to 3.4 kHz. Each frame consists of 20 ms of speech with a Hamming window applied. A 10th order linear predictive analysis is performed on each frame using the autocorrelation method [6]. We have also applied high frequency compensation and a bandwidth expansion of 15 Hz[1] to correct the effects of the anti-aliasing filter and formant under-estimation respectively [1, 3]. The training set consists of 707438 vectors while the evaluation set, consisting of speech not contained in the training, has 85353 vectors.

6. RESULTS

Tables 1 and 2 shows the results for the single ($p = 1$) and multiple ($p = 2, 3, 4$) frame block quantisers using 16 clusters respectively. It can be observed that coding two frames jointly, transparency can be achieved using 23 bits per frame. By using more frames (3 and 4), transparency can be achieved using 22 bits per frame. Since MSE is used as the distortion criteria in the multiple frame version, the number of clusters can be increased without dramatically affecting the computation time. Table 3 shows that using 32 clusters, only 21 bits per frame are required for transparent coding.

We have done some informal testing of the computational complexity between the single frame coder using SD

[1] This is the bandwidth expansion used in the US Federal Standard 1016 4.8 kbps CELP coder

Table 2. Performance of Multiple Frame Block Quantiser using MSE criterion (16 Clusters)

			Outliers (in %)	
p	bits/frame	Avg. SD (dB)	2–4 dB	> 4 dB
2	22	1.050	1.79	0.00
	23	0.988	1.73	0.00
	24	0.931	0.80	0.00
3	21	1.063	2.36	0.01
	22	1.001	1.45	0.00
	23	0.943	0.97	0.00
	24	0.887	0.61	0.00
4	21	1.042	1.94	0.00
	22	0.983	1.30	0.00
	23	0.925	0.90	0.00
	24	0.872	0.57	0.00

Table 3. Performance of Multiple Frame Block Quantiser using MSE criterion (32 Clusters)

			Outliers (in %)	
p	bits/frame	Avg. SD (dB)	2–4 dB	> 4 dB
3	20	1.086	2.46	0.01
	21	1.024	1.70	0.00
	22	0.965	1.16	0.00
	23	0.909	0.74	0.00

and multiple frame coder using MSE by measuring the encoding times on a 2.4 GHz Intel Pentium 4 machine. At a bit rate of 24 bits per frame using 16 clusters, the single frame coder took on average 301 seconds to code 85353 frames while the multiple frame coder ($p = 2$) took 10.7 seconds. For $p = 3$ and $p = 4$, the average times were 11.8 seconds and 13.3 seconds respectively. When using 32 clusters and $p = 3$, the average time was 22.5 seconds. From these results, it can be concluded that the proposed multiple frame coder has a very good performance/complexity tradeoff with the number of frames, p, being an additional design parameter.

7. CONCLUSION

The proposed multiple frame GMM-based block quantiser encompasses both lower complexity and good spectral distortion performance. When operating on multiple frames jointly, inter-frame correlation can be exploited by the coder which leads to a significant improvement in the performance of the block quantisers. This improvement allows us to replace the computionally intensive spectral distortion criterion for block quantiser selection in favour of MSE. Depending on the number of frames used, transparency can be achieved using either 23 or 22 bits per frame. With some added complexity (32 clusters, $p = 3$), 21 bits per frame are enough to achieve an SD of 1.024 dB.

8. REFERENCES

[1] B.S. Atal and M.R. Schroeder, “Predictive Coding of Speech Signals and Subjective Error Criteria”, *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-27, No. 3, pp. 247–254, Jun. 1979.

[2] 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.

[3] P. Kroon and W.B. Kleijn, “Linear-Prediction based Analysis-by-Synthesis Coding” in *Speech Coding and Synthesis*, W.B. Kleijn & K.K. Paliwal, Ed. Amsterdam: Elsevier, 1995, pp. 79–119.

[4] J. Nurminen, “Multi-Mode Quantization of Adjacent Speech Parameters Using a Low-Complexity Prediction Scheme”, in *Proc. EuroSpeech ’03*, Sep. 2003, pp. 1073–1076.

[5] 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.

[6] K.K. Paliwal and W.B. Kleijn, “Quantization of LPC Parameters” in *Speech Coding and Synthesis*, W.B. Kleijn & K.K. Paliwal, Ed. Amsterdam: Elsevier, 1995, pp. 443–466.

[7] 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.

[8] U. Sinervo, J. Nurminen, A. Heikkinen, and J. Saarinen, “Multi-Mode Matrix Quantizer for Low Bit Rate LSF Quantization”, in *Proc. EuroSpeech ’03*, Sep. 2003, pp. 1073–1076.

[9] C. Tsao and R.M. Gray, “Matrix Quantizer Design for LPC Speech Using the Generalized Lloyd Algorithm”, in *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-33, No. 3, pp. 537–545, Jun. 1985.

[10] C.S. Xydeas and C. Papanastasiou, “Split Matrix Quantization of LPC Parameters”, *IEEE Trans. Speech Audio Processing*, Vol. 7, No. 2, pp. 113–125, Mar. 1999.