QUANTIZATION TO MAXIMIZE SNR IN NON-ORTHOGONAL SUBBAND CODERS

Rajeev Gandhi *

Motorola Broadband Comm. Sector
6450 Sequence Drive,
San Diego, CA 92121
rgandhi@gi.com

Sanjit K. Mitra †

Dept. of Elec. and Comp. Engg.
University of California,
Santa Barbara, CA 93106
mitra@ece.ucsb.edu

ABSTRACT

Recent research in the design of filter banks has shown that non-orthogonal filter banks can potentially provide higher coding gains over orthogonal filter banks. The use of non-orthogonal filter banks, however, poses a difficulty in the quantization of subband signals. The conventional nearest-neighbor (NN) encoding rule for the quantization of subband signals is no longer optimal. In this paper, we propose two schemes for the quantization of subband signals in non-orthogonal subband coders. An optimal scheme for quantization of subband signals is proposed first. The complexity of the optimal quantization scheme is shown to grow exponentially with the length of the synthesis filters, which motivates the development of low-complexity quantization schemes when the length of the filters in the synthesis filter bank is large. The second quantization technique uses an iterative method to quantize the subband signals such that the mean square error between the input and the reconstructed output signals is minimized.

1. INTRODUCTION

Many image compression techniques involve the subband decomposition of the image, followed by the quantization, and the entropy coding, of the subband signals. One of the concerns regarding the quantization of subband signals is the orthogonality of the analysis and the synthesis filter banks. Specifically, if the filter banks are orthogonal, then, the mean square error (MSE) between the actual and the quantized subband coefficients is preserved at the output. This allows the use of an efficient nearest-neighbor (NN) encoding rule to quantize the subband signals. On the other hand, recent results in filter bank design [1, 2, 3] have shown that non-orthogonal filter banks can achieve larger coding gains than orthogonal filter banks. The potential of non-orthogonal filter banks to provide larger coding gains, as well as their inherent flexibility (fewer constraints), makes them attractive for use in subband coding.

However, the use of non-orthogonal filter banks complicates the quantization of subband signals. The aim of quantization is to minimize the MSE between the input and the reconstructed output sequence. For non-orthogonal filter banks, this error may not be equal to the error between the original and the quantized subband signals. The quantization schemes for encoding subband signals in non-orthogonal subband coders may be much more complex than those for orthogonal subband coders. One technique to avoid the problem related to quantization of subband signals, is to design the filter banks to be close to orthogonal [4] while simultaneously satisfying some desirable property. The “quasi-orthogonality” of such filter banks makes the nearest-neighbor (NN) encoding of the subband signals approximately optimal. Another technique to quantize the subband signals in non-orthogonal filter banks is to choose the quantized subband signals such that a weighted sum [5] of the errors in the different channels is minimized. The main drawback of this technique is that certain high-rate assumptions are made to model the quantization noise as white noise. The use of such a technique is not valid if the assumptions related to the model are not satisfied. A relaxation-based approach [6] for the quantization of the subband signals in non-orthogonal subband coders can also be used to minimize the MSE between the input and the output signals. While this relaxation-based technique does not assume any model for the quantization noise, the quantization procedure is not optimal.

In this paper, we propose two schemes for the quantization of subband signals in non-orthogonal subband coders. No assumptions are made regarding the model for the quantization noise. We first propose an optimal scheme for the quantization of subband signals in non-orthogonal subband coders. The optimal quantization scheme involves a trellis-based search for the sequence of quantized subband coefficients that results in the lowest MSE between the input and the output signals. The main drawback of the optimal quantization scheme is its enormous computational complexity when the length of the synthesis filters is large. The second scheme is an iterative algorithm that tries to minimize the MSE between the input and the output. The key advantage of this algorithm is its low computational complexity. Typically, the algorithm converges to a solution within a few iterations, though the solution to which it converges may not be optimal. Our simulation results show that appreciable improvement in SNR can be obtained over the NN encoding approach and minimizing the weighted sum of errors approach [5], for encoding the subband signals in strongly non-orthogonal subband coders. Furthermore, the simulation results indicate that the performance of the sub-optimal

---

* The first author performed this work while at the ECE Department, University of California, Santa Barbara

† This work was supported by a University of California MICRO grant with matching supports from Lucent Technologies, National Semiconductors, Raytheon Missile Systems, Tektronix Corporation and Xerox Corporation.

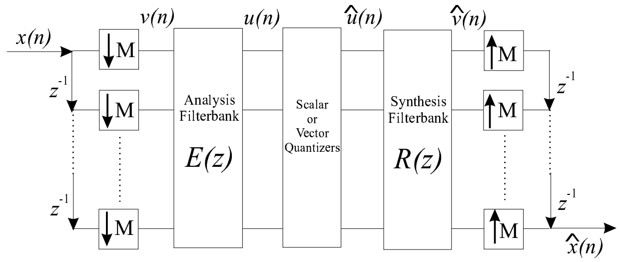

Figure 1: Polyphase Decomposition of an M-channel subband coder

iterative quantization scheme is quite close to that of the optimal trellis-based quantization scheme.

2. OPTIMAL QUANTIZATION OF SUBBAND SIGNALS IN NON-ORTHOGONAL CODERS

Consider the polyphase representation of a typical M-channel subband coder, as shown in Figure 1. In this figure, $\mathbf{E}(z)$ and $\mathbf{R}(z)$ represent the polyphase matrices associated with the analysis and the synthesis filter banks respectively. Throughout the paper, we assume that the analysis and the synthesis filter banks consist of FIR filters. The input $x(n)$ is first converted to M-dimensional vectors, $\mathbf{v}(n) = [x(Mn), x(Mn-1), \ldots, x(Mn-M+1)]^T$, which are filtered by the block filter $\mathbf{E}(z)$. The subband vectors $\mathbf{u}(n)$ are then quantized either by using a scalar quantizer in each of the M-channels, or by using a vector quantizer (VQ) of dimension M. It is also possible to use inter-band VQ; however such schemes are not considered in this paper. The quantized subband vectors, $\hat{\mathbf{u}}(n)$, are filtered through the synthesis block filter $\mathbf{R}(z)$ to produce the output vectors $\hat{\mathbf{v}}(n)$. The mean square error (MSE) between the input and the output sequence is equal to the MSE between the vectors $\mathbf{v}(n)$ and $\hat{\mathbf{v}}(n)$. However, unless $\mathbf{R}(z)$ is orthogonal, the MSE between $\mathbf{u}(n)$ and $\hat{\mathbf{u}}(n)$ may not be equal to the MSE between $\mathbf{v}(n)$ and $\hat{\mathbf{v}}(n)$.

For the sake of simplicity, we assume that the length of the input signal, $x(n)$, is a multiple of M, i.e., $P = KM$, where P is the length of the input signal and that the length of the synthesis block filter is L i.e.

$$\mathbf{R}(z) = \sum_{k=0}^{L-1} \mathbf{R}_k z^{-k}, \tag{1}$$

where $\mathbf{R}_k$ are $M \times M$ matrices. Without loss of generality, we only consider the case when the vector $\mathbf{u}(n)$ is vector-quantized. The number of code vectors in the VQ codebook are assumed to be N. The assumption of using VQ to quantize the subband signals does not exclude the case where the subband signals are scalar-quantized. In such a case, if the scalar quantizers in each of the M-channels are known, we can construct an equivalent vector quantizer whose codebook is the cross-product of the scalar codebooks for the M-channels. The assumption of the subband signals being vector-quantized is made only for the ease of presentation of the algorithms.

With reference to Figure 1, the problem of quantizing $\mathbf{u}(n)$ can be formulated as that of choosing the reconstructed subband vectors $\hat{\mathbf{u}}(n)$ such that the total distortion D is minimized, where

$$\begin{aligned} D &= \sum_{n=0}^{K-1} \| \mathbf{v}(n) - \hat{\mathbf{v}}(n) \|^2, \\ &= \sum_{n=0}^{K-1} \| \mathbf{v}(n) - \sum_{k=0}^{L-1} \mathbf{R_k} \hat{\mathbf{u}}(n-k) \|^2 . \end{aligned} \tag{2}$$

Since there are N code vectors in the codebook, then, the input to the synthesis filter bank at any instant can consist of only one of the N possible vectors, regardless of the true output of the analysis filter bank. A conceptually simple technique to quantize the subband vectors is as follows – given that there are K vectors in the sequence $\mathbf{u}(n)$, construct all N^K possible sequences corresponding to the quantized subband vectors, $\hat{\mathbf{u}}_r(n)$, where $r = 0, 1, \ldots N^K - 1$ is used to index the possible N^K quantized sequences. Next, each one of the N^K sequences, $\hat{\mathbf{u}}_r(n)$, is filtered through the synthesis block filter $\mathbf{R}(z)$ to produce the corresponding output sequence, $\hat{\mathbf{v}}_r(n)$, of the synthesis filter bank. The particular sequence $\hat{\mathbf{u}}_l(n)$ that produces the minimum MSE is then transmitted to the decoder. While this exhaustive-search technique is optimal, its computational complexity grows exponentially with the length, K, of the input sequence.

The complexity of the exhaustive-search quantization scheme can be reduced by using the fact that the output of the synthesis filter bank at any instant depends only upon the choice of the current reconstruction vector and the past $L-1$ reconstruction vectors. Therefore, it might be possible to restrict the search to sequences of length much smaller than K. By suitably defining the "state" of the synthesis filter bank, it is possible to draw a parallel between the synthesis filter bank and a finite-state machine. This similarity can be exploited to reduce the complexity of the optimal quantization scheme. To illustrate the similarity between the synthesis filter bank and a finite-state machine, we define the state of the filter bank at any instant of time to be the indices of the past $L-1$ reconstruction vectors i.e. the indices of $\hat{\mathbf{u}}(n-1), \hat{\mathbf{u}}(n-2), \ldots, \hat{\mathbf{u}}(n-L+1)$. If the number of codevectors in the VQ codebook is N, then, there are N^{L-1} distinct states that can be associated with the synthesis filter bank. The output of the synthesis filter bank, $\hat{\mathbf{v}}(n) = \sum_{k=0}^{L-1} \mathbf{R}_k \hat{\mathbf{u}}(n-k)$, can be rewritten as

$$\hat{\mathbf{v}}(n) = \hat{\mathbf{v}}_s(n) + \hat{\mathbf{v}}_c(n), \tag{3}$$

where

$$\hat{\mathbf{v}}_c(n) = \mathbf{R}_0 \hat{\mathbf{u}}(n) \quad \text{and} \quad \hat{\mathbf{v}}_s(n) = \sum_{i=1}^{L-1} \mathbf{R}_i \hat{\mathbf{u}}(n-i). \tag{4}$$

Eq.(4) reveals that the output of the synthesis filter bank is composed of two terms : $\hat{\mathbf{v}}_s(n)$, which depends entirely upon the state of the filter bank, and $\hat{\mathbf{v}}_c(n)$, which depends only upon the choice of the current reconstruction vector. The choice of a particular reconstruction vector, $\hat{\mathbf{u}}(n)$, determines the state of the synthesis filter bank at time $n+1$. There are N possible state transitions from each state at time n to the states at time $n+1$. These state transitions can be represented by means of a trellis.

There is a one-to-one correspondence between between the state of the filter bank and the past $L-1$ outputs of the filter

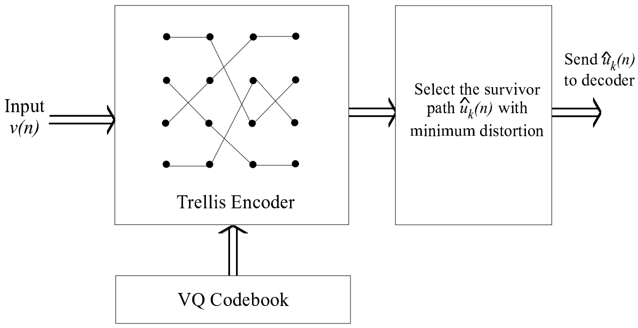

Figure 2: Block diagram of the encoder in a subband coder that uses the optimal quantization scheme

bank. Thus, the problem of finding the optimal sequence of reconstruction vectors can be formulated in terms of finding the optimal sequence of states in the trellis associated with the synthesis filter bank. The latter problem can be efficiently solved using the Viterbi Algorithm [8]. The output of the synthesis filter bank, and hence the MSE, after time n depends only upon the state of the filter bank at time n, and not upon the path used to reach the state at time n. Thus, of all the N paths that reach a particular state, the path with the lowest distortion is retained and the remaining paths can be discarded.

The block diagram for the encoder in a subband coder using the optimal quantization scheme is as shown in Figure 2. A key feature of a subband coder involving optimal quantization is that once the codebook for the VQ is fixed, the output is independent of the analysis filters used to carry out the subband decomposition. This is evident since the quantization scheme does not make use of the subband vectors $\mathbf{u}(n)$, and works directly with the input sequence of vectors, $\mathbf{v}(n)$. The filtering and quantization operations are done jointly by a non-linear trellis-based encoder. Such trellis-based encoding schemes have been proposed [7] earlier to encode sources with memory. The current derivation of such a scheme was, however, motivated in a different context – the problem of optimal quantization in non-orthogonal subband coders.

The complexity of the trellis-based search depends upon the length of the synthesis block filter and the number of codevectors in the codebook. The complexity is proportional to $K \times N^L$ (filtering, addition and comparison) operations. Thus, unlike the exhaustive-search approach, the complexity of the trellis-based search approach increases linearly with the length of the signal. However, the encoding complexity grows exponentially with the length of the synthesis filter bank. The complexity of the trellis-based quantization procedure can be reduced, at the cost of extra storage, by pre-computing the values of $\hat{\mathbf{v}}_s(n)$ corresponding to each of the N^{L-1} states,and those of $\hat{\mathbf{v}}_c(n)$ corresponding to each of the N codevectors. In this case, the filtering operation involves just adding pre-computed vectors. It is also possible to reduce the complexity of the trellis-based approach, at the cost of optimality, by using reduced state-search [7] techniques.

3. ITERATIVE QUANTIZATION OF SUBBAND SIGNALS IN NON-ORTHOGONAL CODERS

In the previous section, we developed an optimal scheme for quantization of subband signals in non-orthogonal subband coders. However, the complexity of the optimal scheme increases exponentially with the length of the synthesis filters. Thus, when the length of the filters used in the synthesis filter bank is large, trellis-based quantization approach becomes impractical due to its enormous computational complexity. The large complexity of the optimal quantization scheme provides the motivation behind the design of low-complexity quantization schemes when the length of the filters used in the synthesis filter bank is large.

The expression for the MSE between the input and the output, as given in Eq. (2), indicates that the choice of a particular reconstruction vector effects a number of terms in the first summation. Consider the case where we have selected the reconstruction vectors up to the time instant $r-1$. In order to choose, $\hat{\mathbf{u}}(r)$, we consider all the terms in Eq. (2) that involve $\hat{\mathbf{u}}(r)$, namely,

$$D_r = \sum_{n=r}^{r+L-1} \| \mathbf{v}(n) - \sum_{k=0}^{L-1} \mathbf{R_k} \hat{\mathbf{u}}(n-k) \|^2 . \tag{5}$$

The reconstruction vector $\hat{\mathbf{u}}(r)$ should be chosen to minimize the distortion given by Eq. (5). The problem involved in choosing $\hat{\mathbf{u}}(r)$ in this manner is that, at the current instant, only the past reconstruction vectors, $\hat{\mathbf{u}}(r-k)$ for $k = 1, 2, \ldots, r$, are known. However, Eq. (5) also involves some of the reconstruction vectors $\hat{\mathbf{u}}(r+k)$ for $k = 1, 2, \ldots, P$ which are as yet unknown, and depend upon the choice of $\hat{\mathbf{u}}(r)$.

This problem can be eliminated by using an estimate of $\hat{\mathbf{u}}(r+k)$ rather than its true value. To this end, we propose an iterative technique to estimate $\hat{\mathbf{u}}(r+k)$. The actual value of the subband vectors $\mathbf{u}(r+k)$ is used as an initial estimate of $\hat{\mathbf{u}}(r+k)$. Based on the estimate of $\hat{\mathbf{u}}(r+k)$, a value of $\hat{\mathbf{u}}(r)$ is chosen to minimize Eq. (5). This process is carried out till all the subband vectors are quantized. At the end of the first iteration, we revisit the problem of quantizing $\hat{\mathbf{u}}(r)$ under the assumption that the values of $\hat{\mathbf{u}}(r+k)$ are fixed to the values obtained in the previous iteration. Since the values of $\hat{\mathbf{u}}(r+k)$ are not necessarily identical to their estimated values used to determine $\hat{\mathbf{u}}(r)$ in the previous iteration, it is possible to refine the value of $\hat{\mathbf{u}}(r)$ to minimize Eq. (5). This procedure is carried out for all the subband vectors, untill the change in total distortion between two successive iterations is less than a specified value. The proposed algorithm can be thought of as sample-by-sample quantization of the subband vectors with the values of the future samples ($\hat{\mathbf{u}}(r+k)$, when quantizing $\mathbf{u}(r)$) estimated by their quantized value in the last iteration. The iterative quantization scheme is not guaranteed to be optimal, however the complexity of the iterative scheme is much lower than that of the trellis search-based approach. The iterative quantization scheme is guaranteed to converge to a local minima since each $\hat{\mathbf{u}}(r)$ is chosen to minimize the total distortion while keeping the rest of the vectors ($\hat{\mathbf{u}}(n)$, $n \neq r$) fixed. Thus after each iteration the value of the total distortion, D, can only decrease or remain constant thereby guaranting the convergence of the iterative algorithm to a local minima.

4. RESULTS

The performance of our proposed quantization schemes was compared to that of quantization schemes that minimize the weighted sum of errors in the different subband channels [5] (hereafter, referred to as the weighted MMSE scheme). For this purpose, 2048 samples of an AR(1) source, with the parmeter $\rho = 0.95$, were encoded at 1.0 and 2.0 bits/sample by a two-channel subband coder. The analysis and synthesis filters that were used for comparisons included the 7-9 filter [9], the 3-9 filter [6], the 5-11 filter [6] and the 4-4 filter [10]. The reconstruction codevectors were generated by using the Generalized Lloyd Algorithm (GLA) over a set of training vectors. The training vectors were generated by filtering an AR(1) signal (different from the test signal) through the analysis filter bank. The codevectors used in the proposed schemes and the weighted MMSE subband coding scheme were the same. An M-best search technique [7] was used to reduce the number of paths retained in the trellis-based quantization scheme. At each stage, in the trellis-based search, the number of survivor paths were limited to 256. The results shown in Table 1 compare the SNR at the output of the subband coder with our quantization schemes to the SNR obtained with the weighted MMSE encoding scheme. These results indicate that the trellis-based quantization scheme out-performs the sub-optimal weighted MMSE scheme by upto $0.5 - 1.1$ dB for strongly non-orthogonal filter banks while the increase in SNR is very small (about $0.05 - 0.1$ dB) for filter banks that are close to orthogonal (the 7-9 and the 3-9 filter bank).

In the trellis-based search, the state from which the decoder should start filtering needs to be transmitted to the decoder. This causes an overhead of $(L-1)log(N)$ bits. Typically, the overhead involved is small, and the percentage increase in bit rate is marginal when the input sequence is of large length. Alternatively, the overhead can be avoided by constraining the trellis search to always begin from a particular pre-determined state. In the results presented in Table 1, the trellis search was constrained to begin from an all zero state.

Filter Bank used	rate = 1.0 bits/sample SNR(dB)	SNR(dB)	SNR(dB)
7-9	7.42	7.43	7.45
3-9	7.91	7.96	7.98
5-11	8.52	8.93	8.98
4-4	5.56	6.02	6.15
	rate = 2.0 bits/sample		
7-9	12.8	12.84	12.86
3-9	12.7	12.8	12.81
5-11	13.10	14.17	14.27
4-4	9.36	10.37	10.47

Table 1: SNR in dB at the output of a non-orthogonal subband coder using the weighted MMSE approach (first column), the iterative approach (second column) and the trellis-based approach (third column).

5. CONCLUSIONS

In this paper, we presented two schemes for the quantization of the subband signals in subband coders involving non-orthogonal filter banks. The first technique optimally quantizes the subband signals in non-orthogonal subband coders. This scheme involves the use of a trellis-based encoder to search the optimal sequence of the reconstruction vectors. However, the complexity of the trellis-based quantization scheme increases exponentially with the length of the sythesis filter banks. The second quantization scheme is an iterative technique to minimize the MSE between the input and the output signals. The advantage of the iterative scheme is its low computational complexity; its drawback is that it may not necessarily result in the optimal choice of the reconstructed subband vectors. Simulation results indicate that appreciable gains in SNR can be obtained using the proposed quantization schemes over the weighted MMSE scheme of [5] in the case when the filter bank used to perform the subband decomposition is strongly non-orthogonal. The simulation results also indicate that even though the iterative quantization scheme is sub-optimal, the loss in SNR values (when compared with the optimal quantization scheme) is very small.

6. REFERENCES

[1] K. C. Aas and C. T. Mullis, "Minimum mean-squared error transform coding and subband coding," *IEEE Trans. Info. Theory*, pp.1179-1192, July 1996.

[2] S. O. Aase and T. A. Ramstad,"On the optimality of nonunitary filter banks in subband coding," *IEEE Trans. Image Proc.*, vol. 4, pp. 1585-1592, December 1995.

[3] P. P. Vaidyanthan and A. Kirac, "Results on optimal biorthogonal filter banks," *IEEE Trans. Circuits and Sys. II: Analog and Digital Sig. Process.*, vol. 45, pp. 932-947, August 1998.

[4] F. M. de Saint-Martin, P. Siohan and A. Cohen, "Biorthogonal filterbanks and energy preservation property in image compression," *IEEE Trans. Image Proc.*, vol. 8, pp. 168-178, February 1999.

[5] K. Gosse and P. Duhammel, "Perfect reconstruction versus MMSE filter banks in source coding," *IEEE Trans. Signal Processing* , vol. 45, pp. 2188-2202, September 1997.

[6] P. Moulin, "A multiscale relaxation algorithm for SNR maximization in 2-D nonorthogonal subband coding," *IEEE Trans. Image Proc.*, vol. 4, pp. 1269-1281, September 1995.

[7] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer, 1992.

[8] G. D. Forney, "The Viterbi Algorithm," *Proc. IEEE*, vol. 61(3), pp. 268-278, March 1973.

[9] M. Antonini, M. Barlaud, P. Mathieu and I. Daubechies, " Image Coding using wavelet transform," *IEEE Trans. Image Proc.*, vol. 1, pp. 205-220, April 1992.

[10] D. LeGall and A. Tabatabai, "Sub-band coding of images using symmetric short kernel filters and arithmetic coding techniques," *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, pp. 761-764, 1988.