Subband Coding Systems Incorporating Quantizer Models

Jelena Kovačević, *Member, IEEE*

***Abstract*—A new method for dealing with the effects of quantization in a subband system is proposed. It uses the "gain plus additive noise" linear model for the Lloyd–Max quantizer. Based on this, it is demonstrated how, by an appropriate choice of synthesis filters, one can cancel all signal-dependent errors at the output of the system. The only remaining error is random in nature and not correlated with the input signal. We therefore have a tradeoff between the error being only random or having signal-dependent components as well (since the error variances in both cases are comparable). As a result of having only a random error, it is possible to reduce this error using, for example, a noise removal technique. The result is then extended to the case where the input is a multidimensional signal, and arbitrary sampling lattices are used, as well as to the QMF (alias cancellation) case. To demonstrate the validity of the proposed approach, two types of experiments on images are carried out: In a toy example, it is shown that using noise removal could be beneficial. For a more realistic coding scheme, however, it is demonstrated that even in the case when the model is no longer valid (when some of the subbands are discarded), the output error is still much less correlated with the input signal as opposed to the commonly used subband system, while visually, the reconstructed images look very similar.**

I. Introduction

Subband coding systems have been used for signal compression for more than a decade and the corresponding theory has progressed from initial, alias-cancellation QMF solutions [1] to perfect reconstruction systems [2]–[5]. However, all of these solutions were developed assuming that there is no coding loss. In reality, the system will possess a quantizer in the middle, and hence, information loss will occur. A typical approach in designing a subband coding system has been to find a perfect reconstruction (or an alias-cancellation) filter bank and then design appropriate subband quantizers. The problem of analyzing the system as a whole, although of significant theoretical and practical importance, has not been addressed by many authors.

Among the works on the subject are [6], where Kronander proposes several criteria to be used when designing filters to be used in a subband coding system with quantization. A variety of analysis/synthesis systems have been explored in [7], for use in speech and image coding. These include DFT's, QMF banks, as well as pseudo-QMF filter banks. In [8], the authors use a statistical model for the optimal design of analysis/synthesis systems that include quantization. Based on this, optimal (in the mean squared error (MSE) sense) synthesis filters are designed given analysis ones (for particular quantizers, such as fine quantization modeled by the additive noise model).

In [9], however, the authors have incorporated a "gain plus additive noise" model for the Lloyd–Max quantizer into a QMF system (which achieves alias cancellation only but not perfect reconstruction). Using such a model, the error at the output of the system can be broken down into different types of errors, such as the aliasing, signal, random, and QMF errors. This, in turn, allows one to investigate the nature and the impact of these various types of errors on the output signal.

By using the same model for the Lloyd–Max quantizer, the aim of this paper is to demonstrate how, by an appropriate choice of synthesis filters, all signal-dependent errors at the output of the system can be cancelled. In other words, the difference between the input and the output is random in nature and not correlated with the input. Note, however, that what we will achieve is a tradeoff between different types of errors. In other words, the total error in the first case will be comparable in energy to the total, but only random, error in the second case. The second case has the potential benefit of having to deal only with a random, noise-like error at the output, which can then be eliminated (or, at least reduced) with an appropriate noise removal technique. Two sets of experiments will be performed. The first experiment is a "toy-example," exaggerated so as to be able to investigate the effects of the unconventional synthesis filters. We use the Haar filters, which provide the most aliasing, and show that even in that case, all signal-dependent parts are cancelled. In the second experiment, we employ a more sophisticated, realistic coding scheme, with a logarithmic frequency split. There, we investigate the performance of the conventional versus unconventional system. It is shown that in the case of practical interest, where some of the subbands are discarded, even if in theory the output has a signal-dependent component, it has still a much smaller correlation coefficient than its unscaled counterpart, while visually the images look very similar.

Note that this is a continuation of previous work [10]. Note also that in the course of revising this paper, we have become aware that similar theoretical results have been independently derived in [11] and [12].

The outline of the paper is as follows: Section II reviews perfect reconstruction filter banks and gives a short account of the work in [9] with particular emphasis on the model used for the Lloyd–Max quantizer. Section III presents original

Manuscript received December 9, 1992; revised April 26, 1994. The associate editor coordinating the review of this paper and approving it for publication was Dr.-Ing. Bernd Girod.

The author is with the Signal Processing Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974 USA.

IEEE Log Number 9410213.

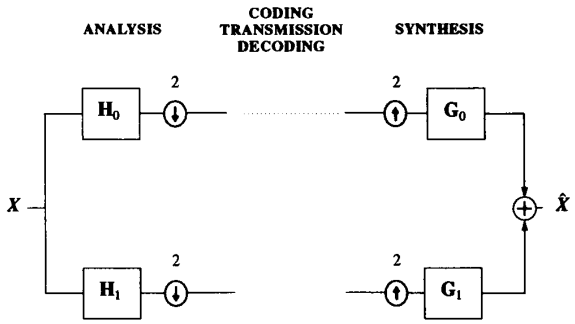


Fig. 1. Two-channel filter bank.

work and introduces the main new concept in the paper, i.e., by careful choice of synthesis filters one can eliminate all signal-dependent errors at the output. This results in the output errors that are not correlated with the input signal. Section IV extends this result to the multidimensional case using arbitrary sampling lattices, as well as to the QMF (alias cancellation) case. Finally, Section V presents experimental results on images, whereupon the conclusions are drawn.

II. Review Material

A. A Glimpse at Perfect Reconstruction Filter Banks

Here, we briefly recall some of the concepts from the theory of perfect reconstruction filter banks that are going to be used in the remainder of this paper. For a more extensive treatment of the subject, refer to [3]–[5] and [14].

An analysis filter bank is a signal processing device that splits the input signal into M channel signals by means of filtering and downsampling by N (where $N \leq M$). In what follows, we will assume that $N = M$, i.e., the filter bank is critically sampled. The synthesis filter bank performs the inverse task. Throughout the remainder of this paper, we will, without any loss of generality, concentrate on the case $N = 2$ (see Fig. 1). The output of the system (in the absence of quantization) is

$$\hat{X}(z) = \frac{1}{2}[G_0(z)H_0(z) + G_1(z)H_1(z)]\, X(z) + \frac{1}{2}[G_0(z)H_0(-z) + G_1(z)H_1(-z)]\, X(-z). \quad (1)$$

The component $X(-z)$ is the aliased version of the signal, and systems designed to remove this part of the signal are termed "alias cancellation." A well-known solution cancelling aliasing is a so-called quadrature mirror filter solution (QMF) [1], with the following choice of filters:

$$\begin{aligned} H_0(z) = \quad G_0(z) &= H(z), \\ H_1(z) = -G_1(z) &= H(-z). \end{aligned} \quad (2)$$

It can be shown that once the filters are chosen as above, it is not possible to obtain perfect reconstruction of the signal,[1] i.e., $\hat{X}(z) = X(z)$. Note, however, that by numerically approximating perfect reconstruction, filters of extremely high quality can be designed (see, for example, [15]).

To achieve both alias cancellation and perfect reconstruction, the filters have to satisfy the following:

$$G_0(z)H_0(z) + G_1(z)H_1(z) = 2 \quad (3)$$
$$G_0(z)H_0(-z) + G_1(z)H_1(-z) = 0. \quad (4)$$

Note that the choice of the filters as above would achieve perfect reconstruction only in the absence of quantization.

B. Gain Plus Additive Noise Model

In the last section, the conditions for perfect reconstruction were given, and it was stressed that they were valid only in the absence of quantization and coding. However, all real systems include a quantizer in the middle, resulting in the loss of information. Consequently, perfect reconstruction property of the system is lost. Although of vast theoretical and practical importance, surprisingly few authors have addressed the problem of joint design of filters and quantizers.

One of the few works on the topic is due to Westerink et al. [9]. The authors use the optimal scalar quantizer to quantize the subbands—Lloyd-Max. For that particular quantizer, it can be shown that (see, for example, [16])

$$\sigma_y^2 = \sigma_x^2 - \sigma_q^2 \quad (5)$$

where $\sigma_q^2, \sigma_x^2, \sigma_y^2$ are the variances of the quantizer, its input and its output, respectively. Consider now a so-called "gain plus additive noise" linear model for this quantizer. Its input/output relationship is given by

$$\mathbf{y} = \alpha \mathbf{x} + \mathbf{r} \quad (6)$$

where $\mathbf{x}, \mathbf{y}$ are the input/output of the quantizer,[2] $\mathbf{r}$ is the additive noise term, and α is the gain factor ($\alpha \leq 1$). The main advantage of this model is that, by choosing

$$\alpha = 1 - \frac{\sigma_q^2}{\sigma_x^2} \quad (7)$$

the additive noise will not be correlated with the signal, and (5) will hold. In other words, to fit the model to our given quantizer, (7) must be satisfied. Note also that the additive noise term is not correlated with the output signal. Even when the input is not zero-mean unity-variance, (7) still holds (although the derivation is slightly different—see the Appendix).

The authors in [9] then incorporate this model into a QMF system (where the filters are designed to cancel aliasing, as given in (2)). Consequently, the error at the output of the system can be written as

$$E(z) = E_Q(z) + E_S(z) + E_A(z) + E_R(z) \quad (8)$$

[1] Except for trivial, two-tap filters in the FIR case.

[2] Bold values denote random variables.

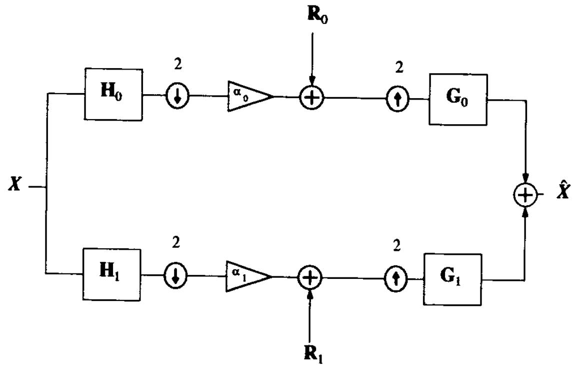


Fig. 2. Subband coding system with the "gain plus additive noise" linear model for the Lloyd–Max quantizer.

where

$$E_Q(z) = \frac{1}{2}[H^2(z) - H^2(-z) - 2]\, X(z) \tag{9}$$

$$E_S(z) = \frac{1}{2}[(\alpha_0 - 1)H^2(z) - (\alpha_1 - 1)H^2(-z)]\, X(z) \tag{10}$$

$$E_A(z) = \frac{1}{2}(\alpha_0 - \alpha_1)\, H(z)\, H(-z)\, X(-z) \tag{11}$$

$$E_R(z) = H(z)R_0(z^2) - H(-z)R_1(z^2). \tag{12}$$

Note that here, z^2 in $R_i(z^2)$ appears since the noise component passes through the upsampler. Also note that in the above, $X(z)$ denotes the signal, while in (6), it would denote a particular subband signal. This breakdown into different types of errors allows one to investigate their influence and severity. Here, E_Q denotes the QMF (lack of perfect reconstruction) error, E_S is the signal error (term with $X(z)$), E_A is the aliasing error (term with $X(-z)$), and E_R is the random error. Note that only the random error E_R is signal independent.

III. Changing Synthesis According To Quantization

As we pointed out earlier, once quantization is used in the system, even if filters are designed properly, the system loses its perfect reconstruction property. The question is then, why use the synthesis part of the system as the exact inverse of the analysis if we know in advance that the signal is not going to be reconstructed perfectly.

Our aim now is to use a general subband system where the synthesis bank is not the inverse of the analysis one, incorporate the linear model for the Lloyd–Max quantizer from Section II-B and see whether anything can be done to eliminate certain types of errors (see Fig. 2). Note that here, no assumptions are made about the filters, that is, filters (H_0, H_1) and (G_0, G_1) do not constitute a perfect reconstruction pair. Assume, however, that given a suitable analysis filter pair (H_0, H_1), we find (T_0, T_1) such that the system is perfect reconstruction. Then, filters H_i and T_i satisfy the conditions (3), (4). Let us now find the expression for the output of the system

$$\begin{aligned}\hat{X}(z) &= \frac{1}{2}[\alpha_0 G_0(z)H_0(z) + \alpha_1 G_1(z)H_1(z)]\, X(z)\\ &+ \frac{1}{2}[\alpha_0 G_0(z)H_0(-z) + \alpha_1 G_1(z)H_1(-z)]\, X(-z)\\ &+ G_0(z)R_0(z^2) + G_1(z)R_1(z^2).\end{aligned} \tag{13}$$

The error is then

$$\begin{aligned}E(z) &= \hat{X}(z) - X(z)\\ &= \frac{1}{2}[\alpha_0 G_0(z)H_0(z) + \alpha_1 G_1(z)H_1(z) - 2]\, X(z)\\ &+ \frac{1}{2}[\alpha_0 G_0(z)H_0(-z) + \alpha_1 G_1(z)H_1(-z)]\, X(-z)\\ &+ G_0(z)R_0(z^2) + G_1(z)R_1(z^2),\\ &= E_S(z) + E_A(z) + E_R(z)\end{aligned} \tag{14}$$

where the signal error $E_S(z)$ is the term with $X(z)$

$$E_S(z) = \frac{1}{2}[\alpha_0 G_0(z)H_0(z) + \alpha_1 G_1(z)H_1(z) - 2]\, X(z) \tag{15}$$

aliasing error $E_A(z)$ is the term with $X(-z)$

$$E_A(z) = \frac{1}{2}[\alpha_0 G_0(z)H_0(-z) + \alpha_1 G_1(z)H_1(-z)]\, X(-z) \tag{16}$$

and random error $E_R(z)$ is

$$E_R(z) = G_0(z)R_0(z^2) + G_1(z)R_1(z^2). \tag{17}$$

Now comes the crucial step. Since after designing quantizers we know α_0 and α_1, choose the synthesis filters (G_0, G_1) as follows:

$$G_0(z) = \frac{1}{\alpha_0}T_0(z), \quad G_1(z) = \frac{1}{\alpha_1}T_1(z). \tag{18}$$

Substituting this into (15)–(17) and taking into account (3) and (4), one can see that the errors become

$$E_S(z) = \frac{1}{2}[T_0(z)H_0(z) + T_1(z)H_1(z) - 2]\, X(z) = 0 \tag{19}$$

$$E_A(z) = \frac{1}{2}[T_0(z)H_0(-z) + T_1(z)H_1(-z)]\, X(-z) = 0 \tag{20}$$

$$E_R(z) = \frac{1}{\alpha_0}T_0(z)R_0(z^2) + \frac{1}{\alpha_1}T_1(z)R_1(z^2) \tag{21}$$

that is, by an appropriate choice of synthesis filters, all signal-dependent errors have been cancelled, and the only remaining error is the random error $E_R(z)$, not correlated with the signal.

What we have done until now might seem a cumbersome way to state an obvious fact: If in Fig. 2 we perform the inverse of the scaling by α_i (that is, $1/\alpha_i$) before upsampling, the effect of the model will propagate only through the noise components $\mathbf{r}_i$. Although deceptively simple, this allows us to preserve perfect reconstruction, except for the noise terms.

The advantage of this approach is that one has to deal with only one type of an error (signal independent). Thus, one could use any noise removal technique to try to eliminate $E_R(z)$ (see Section V). Note, however, that the random error in (21) has been boosted by dividing the terms by $\alpha_i \leq 1$. Note also that the above approach introduces a new concept, that is, after analysis section, use a synthesis filter tuned to a particular quantizer.

Finally, let us discuss the two important border-line cases: $\alpha = 0.0$ and $\alpha = 1.0$. When $\alpha = 1.0$, it means that there are no coding errors, that is, $\sigma_q^2 = 0.0$. In that case, the random noise variance is zero as well (since it can be shown that $\sigma_r^2 = \alpha(1-\alpha)\sigma_x^2$). When $\alpha = 0.0$, on the other hand, we code the signal with its mean value (that is, in practice, we discard that subband). In this case, $\sigma_r^2 = 0.0$ again. Suppose that $\alpha_1 = 0.0$. Then, in (15) and (16), the terms with α_1 go to zero. This further means that we will not be able to recover the part of the signal going though that branch. However, experimental results show that it is still beneficial to perform the scaling on the subbands with $\alpha \neq 0$ since the the output error in that case will be still much less correlated with the input than in the usual case (without scaling). For more details, see Section V.

IV. Generalizations

A. Multidimensional Case

The approach presented earlier is completely general, and can be easily extended to the case where the input is a multidimensional signal. Choosing the synthesis filters as

$$G_i(\mathbf{z}) = \frac{1}{\alpha_i} T_i(\mathbf{z}), i = 0, \ldots, N-1 \tag{22}$$

where N is the number of channels and $\mathbf{z}$ is the n-dimensional z-transform vector, will cancel all signal-dependent errors. The only remaining, random error will be

$$E_R(\mathbf{z}) = \sum_{i=0}^{N-1} \frac{1}{\alpha_i} T_i(\mathbf{z}) R_i(\mathbf{z}^{\mathbf{D}}) \tag{23}$$

where $\mathbf{D}$ is the sampling matrix representing the sampling lattice, and $\mathbf{z}^{\mathbf{D}}$ denotes multidimensional upsampling (for the details of the notation, see [17]).

B. QMF Case

One can see that the above idea works in the perfect reconstruction case. Now, we will show how it can be modified so as to include the QMF case as well. In other words, suppose we have the exact scheme used in [9] and see whether we can eliminate certain errors there. The QMF choice of filters in [9] is as given in (2). Suppose, however, that we apply the same idea as before, that is

$$G_0(z) = \frac{1}{\alpha_0} H(z), \quad G_1(z) = -\frac{1}{\alpha_1} H(-z). \tag{24}$$

The resulting system is then

$$\begin{aligned}\hat{X}(z) &= \frac{1}{2}[\alpha_0 G_0(z) H_0(z) + \alpha_1 G_1(z) H_1(z)]\, X(z) \\ &+ \frac{1}{2}[\alpha_0 G_0(z) H_0(-z) + \alpha_1 G_1(z) H_1(-z)]\, X(-z) \\ &+ G_0(z) R_0(z^2) + G_1(z) R_1(z^2)\end{aligned} \tag{25}$$

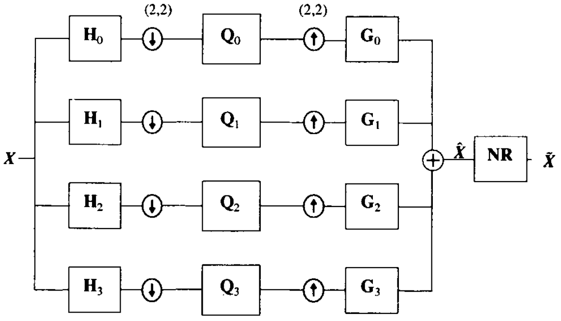


Fig. 3. Scheme used to carry out experiments. **NR** stands for noise removal and is optional.

or, after substituting $H_1(z) = H(-z)$ and (24)

$$\begin{aligned}\hat{X}(z) &= \frac{1}{2}[H^2(z) - H^2(-z)]\, X(z) \\ &+ \frac{1}{2}[H(z)H(-z) - H(-z)H(z)]\, X(-z) \\ &+ \frac{1}{\alpha_0} H(z) R_0(z^2) - \frac{1}{\alpha_1} H(-z) R_1(z^2).\end{aligned} \tag{26}$$

This further means that the overall error is

$$\begin{aligned}E(z) &= \frac{1}{2}[H^2(z) - H^2(-z) - 1]\, X(z) \\ &+ \frac{1}{\alpha_0} H(z) R_0(z^2) - \frac{1}{\alpha_1} H(-z) R_1(z^2), \qquad (27) \\ &= E_Q(z) + E_R(z) \qquad (28)\end{aligned}$$

i.e., it consists only of the QMF error (lack of perfect reconstruction, term with $X(z)$) and the random error. According to the conclusions from [9], the QMF error is almost negligible, meaning that we could use the same technique as proposed for the perfect reconstruction case (the output signal followed by noise removal). This can also be seen as an extension of the scheme in [9].

V. Experimental Results

Our aim in this section is to evaluate the validity of the proposed approach. To this end, experiments on images were performed. The first experiment is a "toy-example," exaggerated so as to be able to investigate the effects of the unconventional synthesis filters. We use the Haar filters, which provide the most aliasing, and show that even in that case, all signal-dependent parts are cancelled. In the second experiment, we employ a more sophisticated coding scheme, with a logarithmic frequency split. There, we investigate the performance of the conventional versus unconventional system. As a test image, "Lena" of size 256 × 256 was used.

A. Toy Example

We will use the simplest meaningful subband splitting scheme, as given in Fig. 3. Each one of the four bands will correspond to one possible combination lowpass/highpass in the horizontal/vertical directions. Then, for example, subband 0 will contain the input signal lowpassed in both directions. The sampling is separable by 2 in each direction.

TABLE I
OPTIMAL QUANTIZER FOR THE ZERO-MEAN UNITY VARIANCE INPUT WITH $c = 0.5$

Level	Probability	Code word
−0.548	0.5	0
0.548	0.5	1

The analysis filters $H_0, \cdots, H_3$ are obtained from the simplest FIR perfect reconstruction filters—Haar, by applying them separately in both directions, i.e.

$$\begin{aligned} H_0(z_1, z_2) &= \frac{1}{2}(1 + z_1^{-1})(1 + z_2^{-1}), \\ H_1(z_1, z_2) &= \frac{1}{2}(1 + z_1^{-1})(1 - z_2^{-1}), \\ H_2(z_1, z_2) &= \frac{1}{2}(1 - z_1^{-1})(1 + z_2^{-1}), \\ H_3(z_1, z_2) &= \frac{1}{2}(1 - z_1^{-1})(1 - z_2^{-1}). \end{aligned} \quad (29)$$

Their perfect reconstruction synthesis filters will be denoted by $T_0, \cdots, T_3$.

For the quantizer design part, we will follow closely work in [9]. Looking at subband histograms, the authors conclude that using the generalized Gaussian probability density function (pdf) will match subband data more closely than some other commonly used distributions, such as the Laplacian pdf. The generalized Gaussian pdf is given by

$$p(x) = ae^{-|bx|^c} \quad (30)$$

where

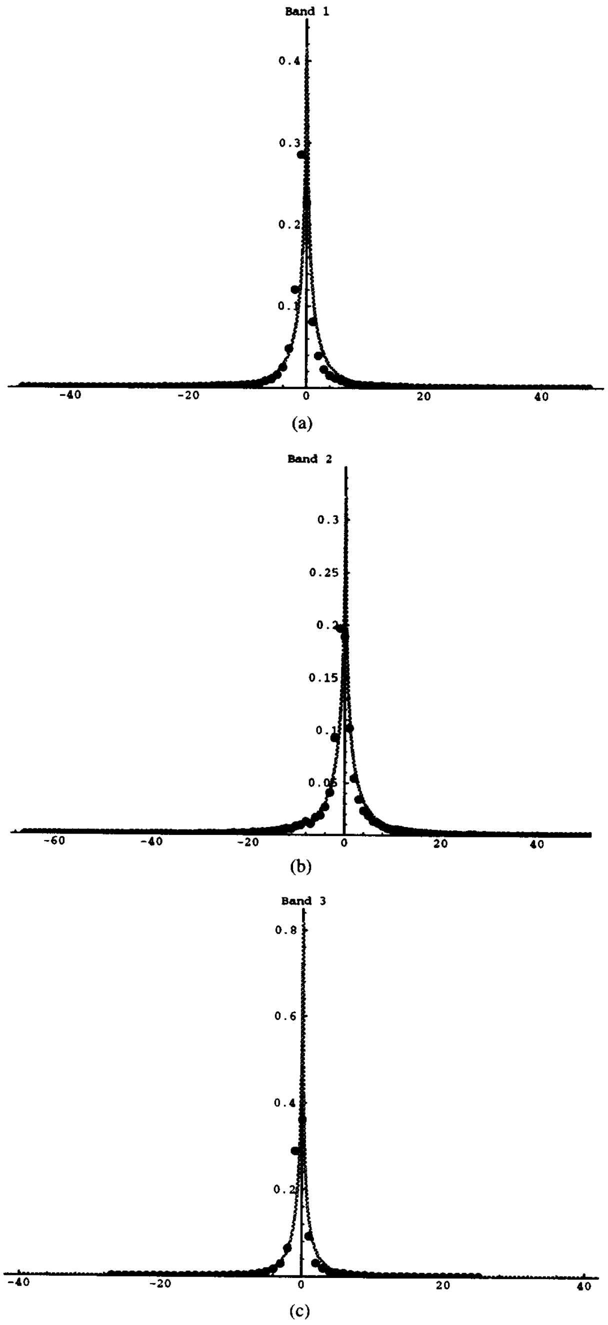
$$a = \frac{bc}{2\Gamma(\frac{1}{c})}, \quad b = \frac{1}{\sigma_x}\sqrt{\frac{\Gamma(\frac{3}{c})}{\Gamma(\frac{1}{c})}} \quad (31)$$

and $\Gamma(.)$ is the Gamma function. The parameter c determines the shape of the function (for example, for $c = 1.0$, one obtains the Laplacian pdf, while $c = 2.0$, results in the Gaussian pdf). For the subbands 1, 2, and 3, the value of $c = 0.5$ yields a good match to the subband data (see Fig. 4). For subband 0, one could fit a Gaussian to the subband data, but it is not a very close match. Therefore, in this example, to avoid having artifacts due to the quantizer mismatch, we will not quantize the low band (which is equivalent to $\alpha_0 = 1.0$, according to (7)).[3] Also, to make the effects of quantization pronounced, we will coarsely quantize the higher three subbands, using only two representation levels. Such a quantizer was designed in [18] and is given in Table I.

Once we have designed quantizers, we can compute the gain factors α_i.Using (7), the gain factors obtained for the test image "Lena" are

$$\alpha_1 = 0.299398, \; \alpha_2 = 0.300149, \; \alpha_3 = 0.296744. \; (32)$$

It should be noted that, even for other images we tried (with the same splitting scheme), the gain factor for higher bands was always around $\alpha \approx 0.3$. In the next section, where a different splitting scheme is used, one will see different gain factors.



(a)

(b)

(c)

Fig. 4. Histogram fitting of the highest three subbands. The true histograms are given by large dots, overlaying the corresponding generalized Gaussian pdf with $c = 0.5$.

1) Conventional System (CS): By conventional system we will denote the system in which the synthesis bank is the exact inverse (or alias cancellation pair) of the analysis bank. The filters used in this case would be the same as given in (29), except time-reversed.

2) Unconventional System (US): By unconventional system we will denote the system in which the synthesis filters

[3] This means that the only artifacts will be due to the quantization of the higher three bands.

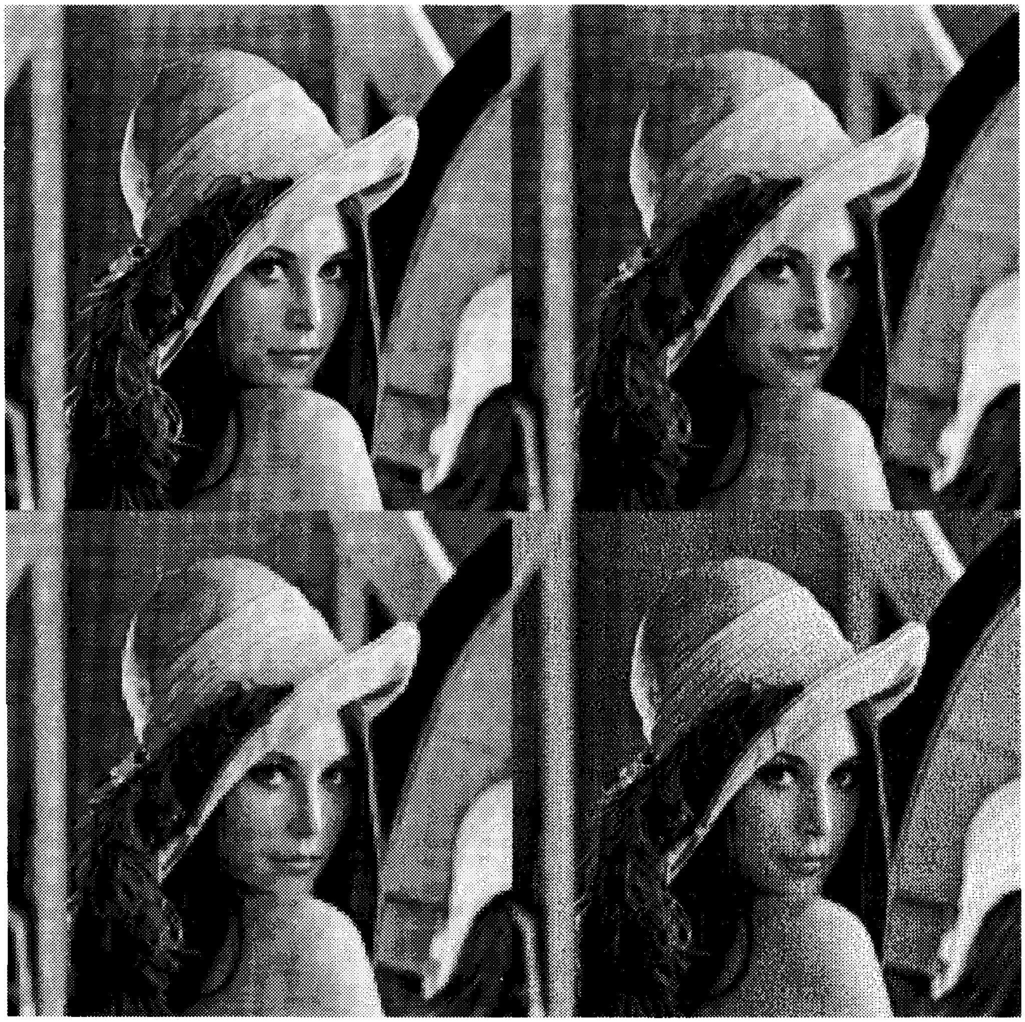


Fig. 5. Results on the test image "Lena" with unquantized subband 0 and 1-bit quantization of subbands 1, 2, and 3: Upper left-hand corner: Original image; lower left-hand corner: The output image obtained with the conventional system; lower right-hand corner: The output image obtained with with the unconventional system; upper right-hand corner: The output image obtained with the unconventional system followed by noise removal.

have been scaled according to (18). Thus, one would use the time-reversed versions of (29) and scale them using the gain factors given in (32).

To observe the output of such a system, the previously described filtering and quantization were applied. Fig. 5 shows four images: original "Lena," the output of the CS, the output of the US, and an example for what could be done after the US, that is US followed by noise removal (to be explained later). Fig. 6 shows the same, except that difference images between the original and the outputs are given. Let us concentrate now on the output of the CS versus US. Note first, how in Fig. 6, the difference between the original and the output of the US is clearly random in nature (compare to the difference between the original and the output of the CS). To quantify this statement the correlation coefficients[4] between the input and the outputs of the CS and US were computed

$$r_{CS} = 0.16097660, \quad r_{US} = 0.00603774. \tag{33}$$

As can be seen, the correlation coefficient for the unconventional system is significantly lower (close to 0) than the one for the conventional system. In the difference image, one can still detect some correlation to the input image (as given by a nonzero r_{US}), which can be attributed to a slight quantizer mismatch.

3) Example: Unconventional System Followed by Noise Removal (US+NR): The main point we wanted to get across is that by using an unconventional system, one obtains an error at the output of the system not correlated with the input signal.

[4] Correlation coefficient is by definition $r = \frac{E\{(\mathbf{x}-\mu_x)(\mathbf{y}-\mu_y)\}}{\sigma_x\sigma_y}$, where μ_x, μ_y are the mean values of $\mathbf{x}$ and $\mathbf{y}$, respectively.

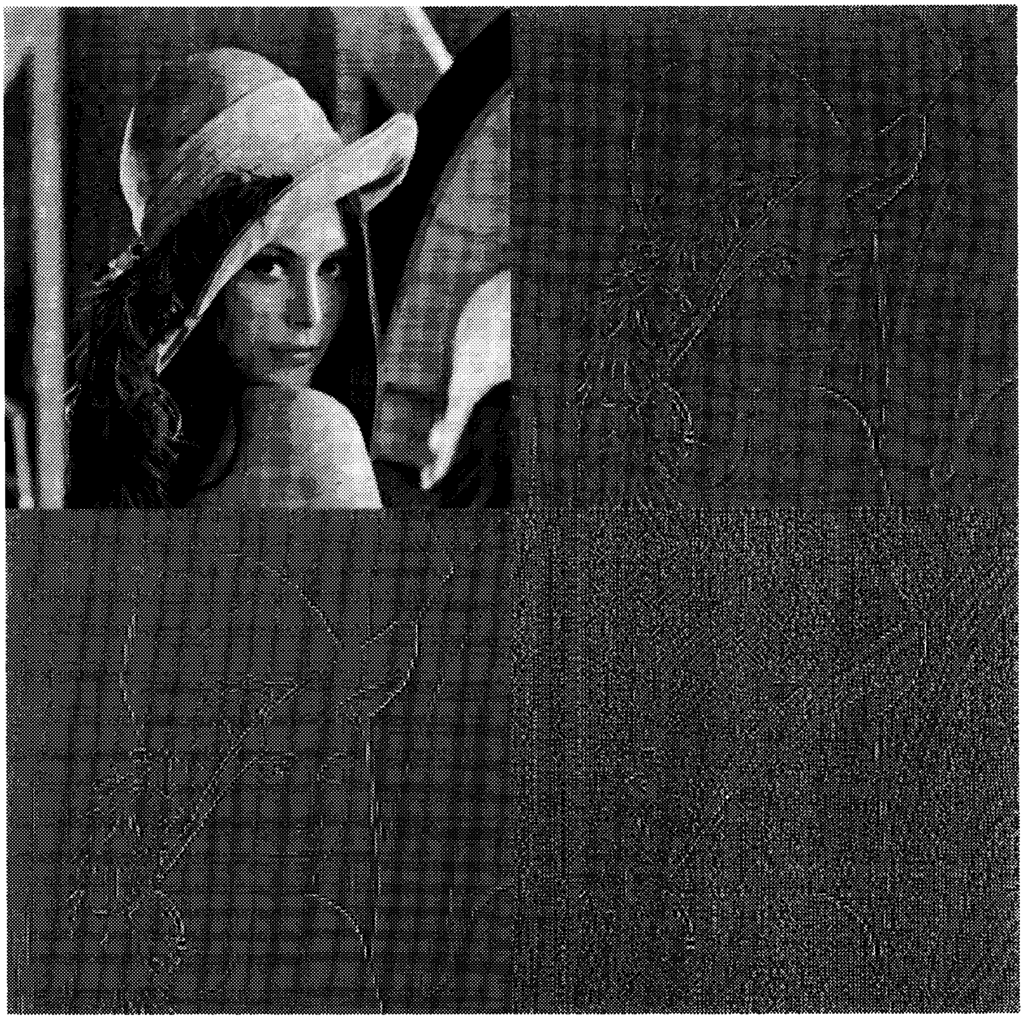


Fig. 6. Difference images for the test image "Lena" with unquantized subband 0 and 1-bit quantization of subbands 1, 2, and 3. Upper left-hand corner: Original image; lower left-hand corner: The difference between the input and the output of the conventional system; lower right-hand corner: the difference between the input and the output of the unconventional system; upper right-hand corner: The difference between the input and the output of the unconventional system followed by noise removal.

The advantage of such a system is that a random error is more easily dealt with than an error highly correlated to the input signal. As an example, a noise removal scheme will be employed to try to reduce the random error (see Fig. 3).

The technique used was proposed by Chan and Lim [19]. They use a cascade of 1-D filters in suitable directions (e.g., horizontal, vertical, and two main diagonals), to perform noise smoothing while preserving edges. The parameters one can vary in the technique are the window size and noise variance. Fig. 5(d) shows the output of the unconventional system followed by noise removal, while Fig. 6(d) shows the difference between that output and the original image. The overall impression when one compares the outputs of the CS versus US+NR is that the latter has the edges much better preserved (observe the rim of the hat and the edge of the cheek). This indicates that it is, indeed, possible to exploit the fact that the output error is random. Note that applying noise removal to CS does not result in any visible improvement.

As an objective measure of quality, one can compute MSE's. Table II gives MSE's between the original image and the output of the CS, US, and unconventional system followed by noise removal (US + NR), as given in Fig. 5. MSE's were computed using the following:

$$\text{MSE} = \sqrt{\frac{\sum_{i,j}(x_{i,j} - y_{i,j})^2}{l \cdot 255^2}} \tag{34}$$

where l is the size of the image (256^2 in this example), and $x_{i,j}$, $y_{i,j}$, are the values of the pixels of the input/output images, respectively.

TABLE II
MEAN SQUARE ERRORS BETWEEN THE ORIGINAL IMAGE AND THE OUTPUT OF THE CONVENTIONAL SYSTEM (CS), UNCONVENTIONAL SYSTEM (US), AND UNCONVENTIONAL SYSTEM FOLLOWED BY NOISE REMOVAL (US + NR)

System	MSE
CS	0.432812
US	0.435017
US + NR	0.425977

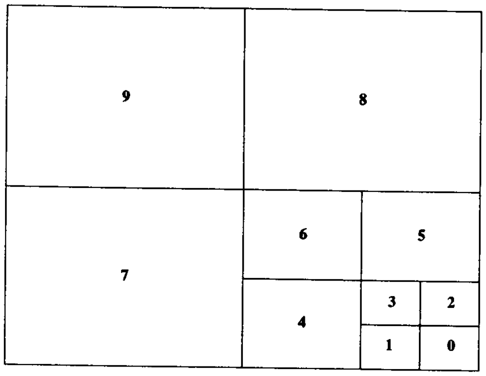


Fig. 7. Three-level logarithmic subband split.

B. Further Investigation of the Unconventional Versus the Conventional System

We want to examine now a more realistic coding scheme and compare the conventional versus the unconventional system. Note that here, we will not be dealing with the noise removal part anymore. That issue is left for future work. To that end, we will perform a three-level logarithmic splitting (this is actually a discrete wavelet transform) as given in Fig. 7. The filter used was the Daubechies' D_6 filter [20] (it has 12 taps) since it was shown in [21] that in wavelet-like schemes regularity of the filter is important. The justification for choosing a 12-tap filter is given later in this section. Note that this filter together with its highpass and synthesis filters is an orthonormal solution, which is beneficial for the bit allocation algorithm (see [21]). Note also that the toy example shown previously could be seen as a special case of this scheme where the lowest subband is not decomposed further.

TABLE III
QUANTIZER CHOICES FOR SUBBANDS 0–9 OBTAINED AS THE OUTPUT OF THE OPTIMAL BIT ALLOCATION ALGORITHM FROM [22]. THEY ARE GIVEN FOR TWO TARGET BIT RATES: 0.7 AND 1.5 bpp. THE NUMBERS IN THE TABLE REPRESENT THE NUMBER OF QUANTIZATION LEVELS OF THE LLOYD–MAX QUANTIZER FOR EACH SUBBAND

Subband	0.7bpp	1.5bpp
0	128	128
1	31	128
2	15	31
3	15	31
4	7	31
5	5	15
6	5	7
7	3	7
8	1	3
9	1	3

TABLE IV
GAIN FACTORS OF THE MODEL FOR THE LLOYD–MAX QUANTIZER FOR TWO TARGET BIT RATES: 0.7 AND 1.5 bpp

Subband	0.7bpp	1.5bpp
0	0.998198	0.998198
1	1.010118	1.000405
2	0.992122	1.022611
3	0.985918	1.022611
4	0.957444	0.996324
5	0.874389	0.991303
6	0.854066	0.994869
7	0.592204	0.934944
8	0.0	0.539889
9	0.0	0.542170

To efficiently allocate bits among subbands, we use the optimal bit allocation algorithm developed in [22]. The algorithm can be optimal for signal blocks that are dependent. It uses operational R-D curves for each subband (in our case) to yield an optimal choice of quantizers for a given budget (bit rate). The algorithm exploits the monotonicity property of the R-D curves. Due to this requirement we choose a set of possible quantizers for each subband such that each of the R-D curves is monotonic. The quantizers were designed in [18]. Therefore, we chose the following set of possible quantizers for the subbands 1–9: 3, 5, 7, 8, 15, 16, 31, 64, 63, 128 (where the numbers indicate the number of quantization levels of the Lloyd–Max quantizers). All of the quantizers are for the shape parameter $c = 0.5$ as explained in Section V-A. For the lowest band, however, we use $c = 0.75$ as proposed in [9] and the following set of quantizers: 31, 64, 63, 128. These sets will ensure that the individual R-D curves are all monotonic as required by the optimal bit allocation algorithm. The algorithm was run for a range of target bit rates from $0-2$ bpp. As an example, Table III shows the quantizers for target bit rates of 0.7 and 1.5 bpp (the actual bit rates are slightly different, namely, 0.676367 and 1.48054 bpp).

Once the subband signals are quantized, we can compute the gain factors according to (7). The scaling is performed on the signals directly, immediately after quantization, rather than incorporating them into the synthesis filters. Table IV shows the gain factors obtained for the target bit rates of 0.7 and 1.5 bpp. Note that for 1.5 bpp, almost all gain factors are around one, except for the last two (since the quantization is sufficiently fine). For 0.7 bpp, the last two values are zero, since these subbands are discarded (each one is quantized with one quantization level—its mean value). Note also that some gain factors exceed one. This can be attributed to numerical errors, as well to a quantizer mismatch. To check how good a quantizer fit we have, correlation coefficients were computed between the input subband signals and the random signals for those subbands that are encoded. The random signals are obtained as the difference between the actual quantized subband signal and the input subband signal multiplied by

TABLE V
CORRELATION COEFFIECIENTS BETWEEN THE INPUT SUBBAND SIGNALS AND THE RANDOM SIGNALS FOR THOSE SUBBANDS THAT ARE ENCODED. THE RANDOM SIGNALS ARE OBTAINED AS THE DIFFERENCE BETWEEN THE ACTUAL QUANTIZED SUBBAND SIGNAL AND THE INPUT SUBBAND SIGNAL MULTIPLIED BY ITS GAIN FACTOR. THEY ARE GIVEN FOR TWO TARGET BIT RATES: 0.7 AND 1.5 bpp. THE VALUES MISSING ARE FOR DISCARDED SUBBANDS

Subband	0.7bpp	1.5bpp
0	$1.107216e-02$	$1.107216e-02$
1	$-1.002162e-01$	$-2.569426e-02$
2	$-5.280836e-02$	$-1.652828e-01$
3	$-3.785217e-02$	$-1.164535e-01$
4	$-8.521539e-02$	$-2.454910e-02$
5	$-4.176539e-02$	$-5.352547e-02$
6	$1.551937e-02$	$-1.400486e-01$
7	$1.321345e-01$	$-3.997022e-02$
8		$1.421262e-01$
9		$1.369439e-01$

its gain factor. Table V lists correlation coefficients for each subband for the previously mentioned bit rates. As can be seen from the table, all correlation coefficients are close to zero, indicating that the random signals are indeed uncorrelated to their input signals.

Finally, to investigate the correlation between the input signal and the output error signal, we ran the coder for bit rates ranging from 0.1–2 bpp. Fig. 8 shows the correlation coefficient as a function of bit rate for the conventional system (dashed line) and the unconventional system (solid line). In the range 1.4 bpp and up, the unconventional system has a correlation coefficient that is at least one order of magnitude lower than the one for the conventional system. Under 1.4 bpp, the curve for the unconventional system starts approaching the conventional one. However, there is still a notable difference between them even at lower bit rates. The fact that the transition happens at 1.4 bpp is because this is the last bit rate at which all of the subbands are coded, meaning the last bit rate where the model is still valid. As soon as one of the subbands is discarded, which happens with subband 9 at target rate of 1.2 bpp (actual rate 1.1594 bpp), the model is no longer valid and we may expect correlated error at the output. This bad news is softened by the fact that the experimental results show that even when that happens, the output error for the unconventional system is still much less correlated than the output error for the conventional one (see Fig. 8).

We said at the beginning of this section that we would offer the justification for using the D_6 filter. We wanted to investigate the effect of different filters on the two systems. We have used the Daubechies family of regular filters of lengths 4–20, that is filters $D_2 - D_{10}$ [20]. The Haar filter was also included in the investigations since it could be seen as the filter D_1. The reason for using Daubechies filters rather then some other, more commonly used ones, is that the recent study by Rioul [21] showed that the regularity of the filters is indeed important in the discrete wavelet transform schemes (logarithmic decompositions). In some sense, this conclusion goes against the commonly used criterion for designing the subband filters, that is, sharp transition bands (frequency selectivity), since regular filters are very smooth and not very selective in frequency. Despite that, these regular filters were shown to perform better than the nonregular ones, both from the subjective and objective points of view [21]. We compared these filters at two bit rates, for 0.7 and 1.5 bpp. Table VI shows the variance of the output error signal as a function of the filter used, for the unconventional system at 0.7 and 1.5 bpp, and for the conventional one at 1.5 bpp. As can be seen from the table, the variance drops dramatically from the Haar, to D_2, to D_4 filters, and then levels off. The same behavior is observed for all three cases given in the table. Based on this, we decided to use the D_6 filter, since it is still considerably shorter than, for example, the D_{10} filter, but achieves almost the same performance (visually as well). Another thing to observe is that the variance of the conventional system for the same bit rate (1.5 bpp) is consistently smaller than the one of the unconventional one. However, visually, the images look almost the same (we tried this for other bit rates as well). Fig. 9(a) and (b) show these results graphically, where the variances are plotted as a function of the number of filter taps (2, 4, 8, 12, 16, 20).

Note that all the experiments in this section were performed on various images, we just chose "Lena" as an example. Also, visually, reconstructed images in both the conventional and the unconventional cases look almost the same. We did not include an example since when reproduced the images would look identical. The issue of whether it would be beneficial to employ some kind of noise removal after synthesis (since the unconventional system has a consistently less correlated error), is left for future work.

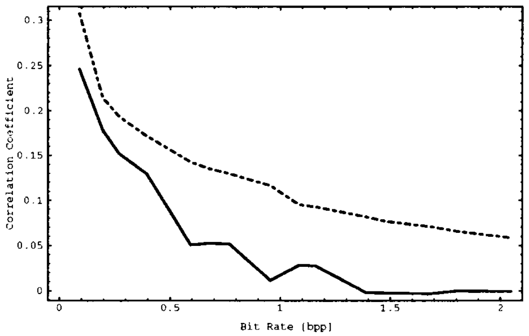


Fig. 8. Correlation coefficient as a function of bit rate for the conventional system (dashed line) and the unconventional system (solid line).

TABLE VI
VARIANCE OF THE OUTPUT ERROR SIGNAL AS A FUNCTION OF THE FILTER USED, FOR THE UNCONVENTIONAL (US) AND CONVENTIONAL SYSTEMS (CS)

Filter	US: 0.7bpp	US: 1.5bpp	CS:1.5bpp
$Haar$	76.73551	25.26492	20.69435
D_2	62.82273	23.39564	16.20846
D_4	49.52245	19.66615	13.91466
D_6	50.11194	19.06851	13.79680
D_8	50.24772	18.78234	13.78546
D_{10}	52.02554	17.69351	12.95259

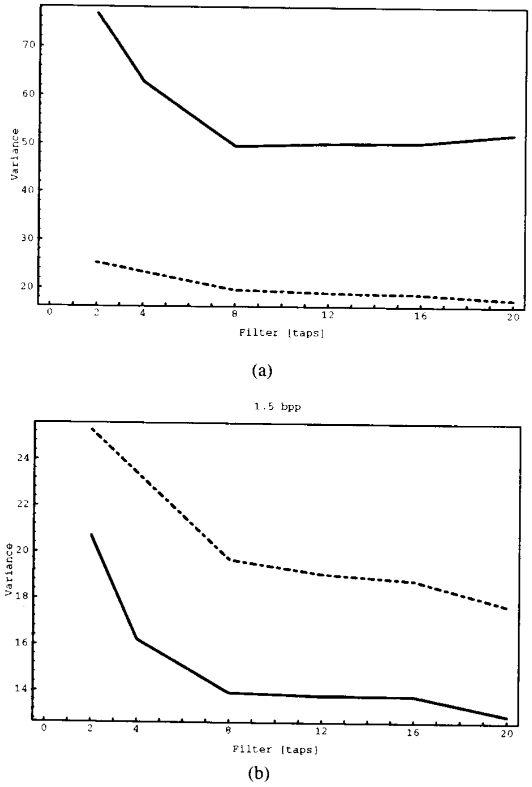


Fig. 9. Variance of the output error signal as a function of the filter used, for the: (a) unconventional system at 0.7 bpp (solid line) and 1.5 bpp (dashed line), and (b) unconventional system at 1.5 bpp (dashed line), and the conventional system at 1.5 bpp (solid line).

VI. Conclusion

In this paper, a new method for dealing with the effects of quantization in a subband system, was proposed. It uses the "gain plus additive noise" linear model for the Lloyd–Max quantizer. Based on this, it was demonstrated how, by appropriate scaling before synthesis, one can cancel all signal-dependent errors at the output of the system. The only remaining error is random in nature and not correlated with the input signal. As a result, it is possible to alleviate this error using an appropriate noise removal technique. The result was then extended to the case where the input is a multidimensional signal, and arbitrary sampling lattices are used, as well as to the QMF (alias cancellation) case. Experimental results show that the output error of the system is always less correlated with the input signal for the unconventional system, even when the model is no longer valid, that is, when some of the subbands are discarded. It is also shown that for a particular scheme employed (discrete wavelet transform with three octaves), a filter of 12 taps is recommended (similarly to [9]). In the same scheme, however, there was no considerable visual difference in the output images. In an exaggerated toy example, it was also shown that using noise removal after unconventional synthesis could be beneficial. This, however, is left for future work.

The results obtained in this paper are the first step in examining the joint effects of filtering and quantization in a subband system. Future investigations will concentrate on search for new quantizer models, as well as on more general unconventional synthesis filters (in the absence of such models). Two particular issues need to be addressed. First, the fact that visually, the results look quite different in the toy example versus the realistic coding scheme. At the same time, in the realistic coding scheme, statistical results indicate that it would be beneficial to use noise removal after the unconventional synthesis, while the reconstructed images look almost identical, contradicting this conclusion. Both of these aspects merit additional work in the future.

Acknowledgment

The author would like to thank N. Jayant for initiating the work on the effects of quantization in subband coding systems, and T. Pappas for pointing out the technique of Chan and Lim on noise removal [19], as well as for the software he provided. She would also like to thank M. Vetterli, K. Ramchandran, A. Ortega, and J. Smith for the software they provided for the optimal bit allocation.

Appendix
Derivation of the Gain Factor in (7)

The following derivation follows the one in [9]. It is included here for completeness. To find the expression for the gain factor when the input is not a zero-mean, unity variance random variable, first note that for the Lloyd–Max quantizer, $E(\mathbf{q}\mathbf{y}) = 0$ [16]. Then, write

$$\mathbf{y} = \mathbf{x} - \mathbf{q} = \alpha\mathbf{x} + \mathbf{r} \tag{35}$$

and compute

$$E(\mathbf{y}^2) = E(\mathbf{y}(\mathbf{x}-\mathbf{q})) = E(\mathbf{y}\mathbf{x}) - E(\mathbf{y}\mathbf{q}) = E(\mathbf{y}\mathbf{x}). \tag{36}$$

At the same time

$$\begin{aligned} E(\mathbf{y}\mathbf{x}) &= E(\mathbf{x}(\mathbf{x}-\mathbf{q})),\\ &= E(\mathbf{x}^2) - E(\mathbf{x}\mathbf{q}),\\ &= E(\mathbf{x}^2) - E(\mathbf{q}(\mathbf{y}+\mathbf{q})),\\ &= E(\mathbf{x}^2) - E(\mathbf{q}\mathbf{y}) - E(\mathbf{q}^2),\\ &= E(\mathbf{x}^2) - E(\mathbf{q}^2). \end{aligned} \tag{37}$$

Then, using (36) and (37)

$$E(\mathbf{y}^2) = E(\mathbf{x}^2) - E(\mathbf{q}^2) \tag{38}$$

$$E(\mathbf{q}\mathbf{x}) = E(\mathbf{q}(\mathbf{y}+\mathbf{q})) = E(\mathbf{q}\mathbf{y}) + E(\mathbf{q}^2) = E(\mathbf{q}^2) \tag{39}$$

$$\begin{aligned} E(\mathbf{q}\mathbf{x}) &= E((\mathbf{x}-\mathbf{y})\mathbf{x}) = E((\mathbf{x}-\alpha\mathbf{x}-\mathbf{r})\mathbf{x})\\ &= E(\mathbf{x}^2) - \alpha E(\mathbf{x}^2) - E(\mathbf{r}\mathbf{x}). \end{aligned} \tag{40}$$

Equating (39) and (40), one obtains

$$E(\mathbf{q}^2) = (1-\alpha)E(\mathbf{x}^2) - E(\mathbf{r}\mathbf{x}). \tag{41}$$

We can now write

$$E(\mathbf{rx}) - E(\mathbf{r})E(\mathbf{x}) = (1-\alpha)E(\mathbf{x}^2) - E(\mathbf{q}^2) - E(\mathbf{r})E(\mathbf{x}) \tag{42}$$

which, since $E(\mathbf{q}) = 0$, and $E(\mathbf{r}) = (1-\alpha)E(\mathbf{x})$, becomes

$$E(\mathbf{rx}) - E(\mathbf{r})E(\mathbf{x}) = (1-\alpha)\sigma_x^2 - \sigma_q^2. \tag{43}$$

It is obvious now that if we choose

$$\alpha = 1 - \frac{\sigma_q^2}{\sigma_x^2} \tag{44}$$

the input will not be correlated with the random noise part, that is

$$E(\mathbf{rx}) - E(\mathbf{r})E(\mathbf{x}) = 0. \tag{45}$$

References

[1] A. Croisier, D. Esteban, and C. Galand, "Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques," in *Proc. Int. Conf. Inform. Sci. Syst.*, Patras, Greece, Aug. 1976, pp. 443–446.

[2] F. Mintzer, "Filters for distortion-free two-band multirate filter banks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 626–630, June 1985.

[3] M. Smith and T. Barnwell III, "Exact reconstruction for tree-structured subband coders," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 431–441, June 1986.

[4] M. Vetterli, "A theory of multirate filter banks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 356–372, Mar. 1987.

[5] P. Vaidyanathan, "Quadrature mirror filter banks, M-band extensions and perfect reconstruction techniques," *IEEE Acoust., Speech, Signal Processing Mag.*, vol. 4, pp. 4–20, July 1987.

[6] T. Kronander, "New criteria for optimization of QMF banks to be used in an image coding system," in *Proc. IEEE Int. Symp. Circ. Syst.*, Portland, OR, 1989, pp. 1354–1357.

[7] T. Lookabaugh, M. Perkins, and C. Cadwell, "Analysis/synthesis systems in the presence of quantization," in *Proc. IEEE Int. Symp. Circ. Syst.*, Portland, OR, 1989, pp. 1341–1344.

[8] A. Dembo and D. Malah, "Statistical design of analysis/synthesis systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 328–341, Mar. 1988.

[9] P. Westerink, J. Biemond, and D. Boekee, "Scalar quantization error analysis for image subband coding using QMF's," *Signal Processing*, vol. 40, pp. 421–428, Feb. 1992.

[10] J. Kovačević, "Subband coding systems incorporating quantizer models," in *Proc. Data Compression Conf.*, Snowbird, UT, Mar. 1993. p. 486.

[11] N. Uzun and R. Haddad, "Modeling and analysis of quantization errors in two-channel subband filter structures," in *Proc. SPIE Conf. Visual Commun. Image Proc.*, Nov. 1992, pp. 1446–1457.

[12] R. Haddad and N. Uzun, "Modeling, analysis, and compenstion of quantization effects in M-band subband codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.*, Minneapolis, MN, Apr. 1993, pp. III: 173–176.

[13] H. Malvar, "Optimal pre- and post-filtering in noisy sampled data systems," Ph.D. thesis, Mass. Inst. of Technol., Aug. 1986.

[14] P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice Hall, 1992.

[15] J. Johnston, "A filter family designed for use in Quadrature Mirror Filter Banks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Denver, CO, 1980, pp. 291–294.

[16] N. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[17] J. Kovačević and M. Vetterli, "Non-separable multidimensional perfect reconstruction filter banks and wavelet bases for $\mathcal{R}^n$," *IEEE Trans. Inform. Theory, Special Issue Wavelet Transforms Multiresolution Signal Analysis*, vol. 38, pp. 533–555, Mar. 1992.

[18] P. Westerink, "Subband coding of images," Ph.D. thesis, Delft Univ. of Technol., Delft, The Netherlands, 1989.

[19] P. Chan and J. Lim, "One-dimensional processing for adaptive image restoration," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 117–125, Feb. 1985.

[20] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Commun. Pure Appl. Math.*, vol. 41, pp. 909–996, Nov. 1988.

[21] O. Rioul, "Regular wavelets: A discrete-time approach," *IEEE Trans. Signal Processing, Special Issue Wavelets Signal Processing*, vol. 41, no. 12, pp. 3572–3578, Dec. 1993.

[22] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," *IEEE Trans. Image Proc.*, vol. 3, no. 5, pp. 533–545, Sept. 1994.

Jelena Kovačević (S'88–M'91) was born in Yugoslavia in 1962. She received the Dipl. Electr. Eng. degree from the Electrical Engineering Department, University of Belgrade, Yugoslavia, in 1986, and the M.S. and Ph.D. degrees from Columbia University, New York, NY, in 1988 and 1991, respectively.

In November 1991, she joined AT&T Bell Laboratories, Murray Hill, NJ, as Member of Technical Staff. In the fall of 1986, she was a Teaching Assistant the University of Belgrade. From 1987 to 1991, she was a Graduate Research Assistant at Columbia University. In the summer of 1985, she worked for Gaz de France, Paris, France, during the summer of 1987 for INTELSAT, Washington, D.C., and in the summer of 1988 for Pacific Bell, San Ramon, CA. Her research interests include multirate signal processing, wavelets, and image and video coding.

Dr. Kovačević received the E.I. Jury Award at Columbia University for outstanding achievement as a graduate student in the areas of systems, communications, and signal processing. She is the coauthor (with M. Vetterli) of *Wavelets and Subband Coding* (Prentice Hall, 1994). She is an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and is guest co-editor (with I. Daubechies) of the Special Issue on Wavelets of the PROCEEDINGS OF THE IEEE. She is on the MDSP Technical Committee of the Signal Processing Society of the IEEE and is the General Co-Chair (with J. Allebach) of the Ninth Workshop on Image and Multidimensional Signal Processing.