Minimization of Quantization Noise Amplification in Biorthogonal Subband Coders

Anamitra Makur, *Senior Member, IEEE* and M. Arunkumar

***Abstract*—Quantization noise amplification (QNA) restricts the coding gain (CG) in biorthogonal subband coders. Here, a coloring filter is introduced to color the quantization noise. The optimal coloring filter eliminating/minimizing QNA, with or without order restriction, is found for a given finite-impulse response filter bank (FB). An efficient implementation of the coloring filter is proposed. With the coloring filter, the optimal biorthogonal ideal FB becomes the full whitening coder achieving maximum possible CG. Results verify the CG improvement due to coloring for existing FBs.**

***Index Terms*—Quantization noise, subband coding gain (CG).**

I. Quantization Noise in Subband Coding

A. Coding Gain

A SUBBAND CODER, consisting of a filter bank (FB) and quantization, offers coding gain (CG). The perfect reconstruction (PR) FB of a subband coder may be orthogonal or *biorthogonal* (or, general PR). In this section, the additive white noise model with uncorrelated noise (uniform quantizer at high rate) is used for the quantizers. CG of a biorthogonal subband coder with optimal bit allocation (OBA) is [1]

$$G = \sigma_x^2 \Bigg/ \left(\prod_{i=0}^{M-1} \sigma_{x_i}^2 \mathbf{f}_i^T \mathbf{f}_i \right)^{1/M} \tag{1}$$

where σ_x^2 is the source variance, M is the number of bands, $\sigma_{x_i}^2$'s are subband variances, and $\mathbf{f}_i$ is the impulse response column vector of the ith synthesis filter (vectors in this work are denoted by bold small letters). The additional terms $\mathbf{f}_i^T \mathbf{f}_i$, synthesis filter norms, appear because the white quantization noise passing through the synthesis FB becomes colored. Consequently, the overall noise variance is amplified, which is referred as *quantization noise amplification* (QNA).

The biorthogonal coder has an advantage and a disadvantage compared to the orthogonal coder. The advantage is that the biorthogonal FB passbands are not restricted to be flat, so passband shaping makes the subband spectrums partially flat before quantization, resulting in higher CG. The disadvantage is the QNA, which reduces the CG. In fact, CG itself $(G \geq 1)$ is no longer guaranteed [2], [3].

Manuscript received December 4, 2003; revised March 12, 2004. This paper was recommended by Associate Editor S.-M. Phoong.

A. Makur was with the Department of Electrical and Communication Engineering, Indian Institute of Science, Bangalore 560012, India. He is now with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: eamakur@ntu.edu.sg).

M. Arunkumar was with the Department of Electrical and Communication Engineering, Indian Institute of Science, Bangalore 560012, India. He is now with is presently with the Department of Electrical and Computer Engineering, Universtiy of Maryland, College Park, MD 20742 USA.

Digital Object Identifier 10.1109/TCSI.2004.835660

B. Proof of Amplification

While QNA ("noise gain" in [2], [3]) has been amply reported, to our knowledge it has not been explicitly shown that the synthesis norms indeed amplify the noise. We now show this. Let $\mathbf{E}(z)$ and $\mathbf{R}(z)$ be the analysis and synthesis polyphase matrices of some PR FB (matrices in this work are denoted by bold capital letters). Let $\mathbf{e}_i^T(z) = \mathbf{e}_i^{*H}(z)$ be the ith row of $\mathbf{E}(z)$, where T is transpose, $*$ is conjugation, and H is hermitian transpose. Let $\mathbf{r}_i(z)$ be the ith column of $\mathbf{R}(z)$. Then, from the Cauchy–Schwarz inequality

$$\left| \mathbf{e}_i^{*H}(z) \mathbf{r}_i(z) \right|^2 \leq \|\mathbf{e}_i^*(z)\|^2 \|\mathbf{r}_i(z)\|^2. \tag{2}$$

However, since PR implies $\mathbf{E}(z)\mathbf{R}(z) = \mathbf{I}$ (where $\mathbf{I}$ is identity matrix), $\mathbf{e}_i^T(z)\mathbf{r}_i(z) = 1$. (PR condition may involve a delay, which gives same result, and a scaling, which is assumed to be 1 here.) Using $\|\mathbf{e}_i^*(z)\| = \|\mathbf{e}_i(z)\|$, (2) becomes $1 \leq \|\mathbf{e}_i(z)\| \, \|\mathbf{r}_i(z)\|$. Integrating both sides on the unit circle and squaring, we obtain

$$1 \leq \left| \int_{-\pi}^{\pi} \|\mathbf{e}_i(e^{j\omega})\| \, \|\mathbf{r}_i(e^{j\omega})\| \frac{d\omega}{2\pi} \right|^2. \tag{3}$$

Applying the Cauchy–Schwarz inequality on the right-hand side

$$\left| \int_{-\pi}^{\pi} \|\mathbf{e}_i(e^{j\omega})\| \, \|\mathbf{r}_i(e^{j\omega})\| \frac{d\omega}{2\pi} \right|^2 \leq \int_{-\pi}^{\pi} \|\mathbf{e}_i(e^{j\omega})\|^2 \frac{d\omega}{2\pi} \int_{-\pi}^{\pi} \|\mathbf{r}_i(e^{j\omega})\|^2 \frac{d\omega}{2\pi}. \tag{4}$$

The integrals on the right-hand side are the analysis and synthesis filter norms $\mathbf{h}_i^T \mathbf{h}_i$ and $\mathbf{f}_i^T \mathbf{f}_i$, where $\mathbf{h}_i$ is the ith analysis filter impulse response. Therefore, combining (3) and (4), $1 \leq \mathbf{h}_i^T \mathbf{h}_i \mathbf{f}_i^T \mathbf{f}_i$. Scaling analysis/synthesis filters does not affect the CG, since the PR property requires an inverse scaling of synthesis/analysis filters, and the term $\mathbf{h}_i^T \mathbf{h}_i$ is implicitly present in $\sigma_{x_i}^2$ of (1). So, assume the norm of $\mathbf{h}_i$ is 1. Then, $1 \leq \mathbf{f}_i^T \mathbf{f}_i$, or the synthesis filter norms indeed amplify the noise. No QNA results for orthogonal FB, since its filters are of unit norm.

C. Related Work and Proposed Remedy

To tackle QNA for a given FB, the norm-weighted error energy [instantaneous version of (6)] is used in a rate-distortion based algorithm in [4]. The relaxation algorithm replaces the nearest-neighbor quantizer to minimize instead the reconstruction error in [5]. Similarly, the trellis-based and iterative vector

quantization in [6], [7] finds optimum sequence of codewords which, when passed through the synthesis FB, minimizes the reconstruction error. MINLAB or "minimum noise structures" (seen later as special cases of the proposed coloring filter) are used to eliminate QNA for the two-channel ladder-based FB in [2] and for a prediction-based lower-triangular transform in [3]. In Section II, we introduce a coloring filter to "color" the quantization noise to eliminate (for some FBs, minimize) QNA. We find the optimal coloring filter of both a given order and unrestricted order. Section III proposes two structures to implement the coloring filter and compares their implementation costs.

A different situation arises if the FB is designed afresh. Paper [4] proposes designing a nearly orthogonal FB to reduce QNA. The FBs of [2] and [3] have a built-in structure that, together with MINLAB, eliminates QNA. Such FBs offer a high CG; the 2-tap FB of [2] has higher G than the ideal orthogonal FB for any autoregressive (AR) order 1 or moving average order 1 source. While we are yet to find the finite-order optimal FB with coloring filter, we propose a sub-optimal design which performs very well.

The optimal biorthogonal FB in the ideal case [8] is known to be the ideal orthogonal FB followed by a *half-whitening* filter in each band. Its CG is superior to that of the ideal orthogonal coder. It is long known in signal compression that the maximum CG obtainable for a source with spectral flatness measure [9] γ_x^2 is ${}^{\infty}G = 1/\gamma_x^2$. This is equal to the CG achieved by various ideal coders, such as Coder A) ${}^{\infty}G_P$ for optimal predictive coder of infinite order, and Coder B) ${}^{\infty}G_{\mathrm{SBC}}$ for optimal subband coder with infinite number of bands [9]. Coder A essentially involves DPCM structure with quantization in the filtering loop, resulting in the quantization noise effectively passing through the inverse (synthesis) filter at the encoder. The optimal filter in this case is *full whitening (FW)* filter. It is also known that the predictive coder may be modified to obtain the *noise feedback (NF) coder* [9], where the quantization noise is arbitrarily shaped by passing through a coloring filter. The coder gain is never more than that of Coder A. In the special case when the quantization noise does not pass through any filter (D*PCM in [9]), CG involves decoder filter norm ("power transfer factor" in [9]), and the optimal filter is half-whitening filter.

It is, therefore, intuitive to use FW (similar to Coder A) in a biorthogonal subband coder, since whitening achieves maximum predictive CG. In the conventional case, the quantization noise passes through the synthesis filter. If FW is used at the analysis side, its inverse has to be used at the synthesis side. When white quantization noise would pass through this synthesis FB, it would become colored and would reduce the CG. Therefore, the filter norms force us to stop at half-whitening, in a fashion similar to D*PCM, and CG suffers. In our case, using the proposed coloring filter, the optimal biorthogonal coder will perform better than the conventional optimal coders. With a coloring filter, an obvious remedy is to color the quantization noise to the exact inverse of the synthesis filter. Then, the final noise spectrum becomes white. In Section IV we show that the optimal biorthogonal FB is indeed FW, and show that it achieves ${}^{\infty}G$.

Section V presents simulation results to illustrate the CG improvement using the coloring filter for both existing FBs and for approximate FW coder.

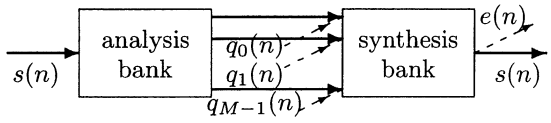


Fig. 1. Subband coder with signal and noise path.

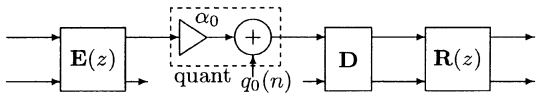


Fig. 2. Subband coder showing quantization model.

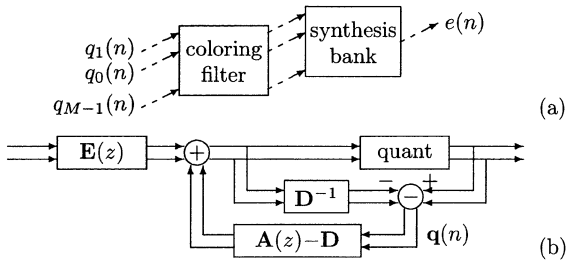


Fig. 3. Coloring filter. (a) Noise path. (b) Realization.

II. Minimizing Noise Amplification

A. Coloring Filter

Fig. 1 shows a conventional subband coder with both banks. $s(n)$ is the input while the output is $s(n) + e(n)$, where $e(n)$ is the reconstruction error. The uncorrelated quantization noise in each band is $q_i(n)$. Due to linearity of the FBs, the coder is shown as a superposition of a signal path (horizontal) producing perfect output, and a noise path (diagonal) through the synthesis FB producing $e(n)$.

In this section, more general *gain plus additive noise* model is used for the quantizers [9] (uniform or pdf-optimized quantizer at any rate). Fig. 2 shows the subband coder with this quantizer model. Each quantizer has a gain term $\alpha_i \leq 1$ and an uncorrelated noise source $q_i(n)$. As shown in [10], a synthesis bank $\mathbf{R}(z)\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with diagonal elements $(1/\alpha_0), (1/\alpha_1), \ldots, (1/\alpha_{M-1})$, is used to cancel the correlated noise such that $e(n)$ becomes uncorrelated to $s(n)$.[1] The noise path $\mathbf{q}(n) = [q_0(n)q_1(n)\ldots q_{M-1}(n)]^T$ passing through $\mathbf{R}(z)\mathbf{D}$ produces reconstruction error $e(n)$ uncorrelated to the input.

We introduce a coloring filter $\mathbf{A}(z)$ at the analysis side that shapes the uncorrelated quantization noise $\mathbf{q}(n)$, so that the noise path is now modified as shown in Fig. 3(a). This may be achieved by first extracting $\mathbf{q}(n)$ and then feeding it back through $\mathbf{A}(z) - \mathbf{D}$, as shown in Fig. 3(b). Analyzing Fig. 3(b), $\mathbf{q}(n)$ passes through $\mathbf{A}(z) - \mathbf{D}$ and then through the quantizer. But the quantizer is modeled as scaling by $\mathbf{D}^{-1}$ and adding $\mathbf{q}(n)$. Therefore, the noise path is modified such that $\mathbf{q}(n)$ now passes through $\mathbf{D}^{-1}(\mathbf{A}(z) - \mathbf{D}) + \mathbf{I} = \mathbf{D}^{-1}\mathbf{A}(z)$ before reaching $\mathbf{R}(z)\mathbf{D}$. The signal path remains unchanged. However, since $\mathbf{A}(z) - \mathbf{D}$ is at the feedback path, it should be implementable (no delay-free loop). If $\mathbf{A}(z) - \mathbf{D}$ is strictly causal,

[1] Since $(1/\alpha_i) \geq 1$, the synthesis norms of $\mathbf{R}(z)\mathbf{D}$ are not less than those of $\mathbf{R}(z)$. Hence, the QNA shown in Section I.B still persists (possibly more) in this structure.

then it is implementable since only past vectors $\mathbf{q}(n-i), i>0$, are used. This is the requirement if the quantizer is a vector quantizer. However, for scalar quantizers applied sequentially from bottom to top subbands, any subband may use components of $\mathbf{q}(n)$ of subbands *below* it. Consequently, $\mathbf{A}(z)-\mathbf{D}$ should be causal and its zeroth-order coefficient matrix should be strictly upper triangular (zero diagonal). Quantizing the bottom subband first and so forth is assumed since, as shown later, the sequence of performing quantization does not affect the performance.

In the presence of $\mathbf{A}(z)$, the reconstruction error $e(n)$ will be $\mathbf{q}(n)$ filtered through $\mathbf{R}(z)\mathbf{D}\mathbf{D}^{-1}\mathbf{A}(z)=\mathbf{R}(z)\mathbf{A}(z)$, which we term as the *effective noise filter*. Let $\mathbf{S_{qq}}(e^{j\omega})$ be the power-spectral density (psd) matrix of $\mathbf{q}(n)$, which becomes diagonal with elements $\sigma_{q_i}^2$ with white uncorrelated noise assumption. Then, the reconstruction error variance σ_e^2 is

$$\frac{1}{M}\mathrm{Tr}\left(\int_{-\pi}^{\pi}\mathbf{R}(e^{j\omega})\mathbf{A}(e^{j\omega})\mathbf{S_{qq}}(e^{j\omega})\mathbf{A}^H(e^{j\omega})\mathbf{R}^H(e^{j\omega})\frac{d\omega}{2\pi}\right)$$

where Tr denotes the trace. By interchanging the order of trace and integration, and using the fact that $\mathrm{Tr}(AB)=\mathrm{Tr}(BA)$, σ_e^2 becomes

$$\frac{1}{M}\int_{-\pi}^{\pi}\mathrm{Tr}(\mathbf{S_{qq}}(e^{j\omega})\mathbf{A}^H(e^{j\omega})\mathbf{R}^H(e^{j\omega})\mathbf{R}(e^{j\omega})\mathbf{A}(e^{j\omega}))\frac{d\omega}{2\pi}.$$

Let the effective noise filter be

$$\mathbf{R}(z)\mathbf{A}(z)=[\mathbf{r}_0'(z),\mathbf{r}_1'(z),\ldots,\mathbf{r}_{M-1}'(z)] \tag{5}$$

where $\mathbf{r}_i'(z)$ denotes its ith column. Then

$$\sigma_e^2=\frac{1}{M}\sum_{i=0}^{M-1}\sigma_{q_i}^2\int_{-\pi}^{\pi}\mathbf{r}_i'^H(e^{j\omega})\mathbf{r}_i'(e^{j\omega})\frac{d\omega}{2\pi} \tag{6}$$

where $\int_{-\pi}^{\pi}\mathbf{r}_i'^H(e^{j\omega})\mathbf{r}_i'(e^{j\omega})\,d\omega/2\pi$ is the effective noise filter norm. The above result in absence of $\mathbf{A}(z)$ is well known [3], [4]. Now, the aim is to minimize QNA, or σ_e^2, which may be achieved by minimizing each norm. By using OBA, the arithmetic mean in (6) is replaced by geometric mean. In this case, the minimization is achieved by minimizing the product of the norms. For some given FB, a coloring filter $\mathbf{A}(z)$ of certain class that achieves such minimization is termed as optimal. In the following sections, we address the design of optimal $\mathbf{A}(z)$ for different cases. First, the optimal $\mathbf{A}(z)$ of a given order is found. This result is derived for any bit allocation. Then, it is shown that the CG obtained using the optimal coloring filter of a given order is independent of the sequence of quantization. This result is valid only under OBA. Next, for a given FIR FB, the optimal coloring filter of unrestricted order is found. The order of this filter is at most equal to the analysis polyphase order. The above results are derived for the gain plus noise model for quantizers, and replacing $\alpha_i=1\ \forall i$ the corresponding results for the additive noise model may be obtained.

B. Optimal Coloring Filter of a Given Order

Here, we find what is the optimal coloring filter, restricted to order p finite-impulse response (FIR), for a given FIR PR FB.

Let $\mathbf{A}(z)$ be restricted to order p, and let a_{il}^j denote the jth-order coefficient of $a_{il}(z)$. Then, from (5)

$$\mathbf{r}_l'(z)=\sum_{i=0}^{M-1}\mathbf{r}_i(z)\sum_{j=0}^{p}z^{-j}a_{il}^j \tag{7}$$

for all l. The columns $\mathbf{r}_l(z)$ are the polyphase components of the FIR synthesis filters $\mathbf{f}_l$ (a column vector). Similarly, $\mathbf{r}_l'(z)$ represent the effective noise filters (also FIR) with impulse responses $\mathbf{f}_l'$. Further, $z^{-j}\mathbf{r}_l(z)$ represents the filter $\mathbf{f}_l$ delayed by jM samples. We use a short-hand notation $\mathbf{f}_l^{(jM)}$ to denote the column vector representing the impulse response of this filter. Let all these column vectors $\mathbf{f}_l^{(jM)}$ and $\mathbf{f}_l$ be zero padded at the end to make them all have the same length. Since $\{a_{il}^0\}$ is upper triangular, $a_{il}^0=0$ for $i>l$ and $a_{ll}^0=(1/\alpha_l)$. Then, from (7) it follows that

$$\mathbf{f}_l'=\frac{1}{\alpha_l}\mathbf{f}_l+\sum_{i=0}^{l-1}a_{il}^0\mathbf{f}_i+\sum_{j=1}^{p}\sum_{i=0}^{M-1}a_{il}^j\mathbf{f}_i^{(jM)} \tag{8}$$

or, $\mathbf{f}_l'=(1/\alpha_l)(\mathbf{f}_l-\hat{\mathbf{f}}_l)$ for $0\le l<M$. This may be interpreted as an equation giving the residual error $\alpha_l\mathbf{f}_l'$ in projecting the lth synthesis filter $\mathbf{f}_l$ onto the space spanned by the zeroth, first, $\ldots,(l-1)$th synthesis filters [second term of (8)] and all filters delayed by $M,2M,\ldots,pM$ samples [third term of (8)]. Define $\mathbf{X}=[\mathbf{f}_{M-1},\mathbf{f}_{M-2},\ldots,\mathbf{f}_0,\mathbf{f}_0^{(M)},\ldots,\mathbf{f}_{M-1}^{(M)},\ldots,\mathbf{f}_0^{(pM)},\ldots,\mathbf{f}_{M-1}^{(pM)}]$ such that

$$\hat{\mathbf{f}}_l=\mathbf{X}\begin{bmatrix}\mathbf{0}_{M-l}^T & \mathbf{b}_l^T\end{bmatrix}^T \tag{9}$$

or $\hat{\mathbf{f}}_l$ is a linear combination of some columns of $\mathbf{X}$, with $\mathbf{b}_l$ giving the weights of these columns, where $\mathbf{0}_i$ denotes zero column vector of length i. The polynomial entries of the lth column of $\mathbf{A}(z)$ have coefficients $-(1/\alpha_l)\mathbf{b}_l$, and a_{ll}^0 is $(1/\alpha_l)$. Let $\mathbf{X}_l$ denote the matrix $\mathbf{X}$ without the first $M-l$ columns. The following theorem provides the optimal solution.

Theorem 1: For a given FIR FB with $\mathbf{X}_l,\mathbf{f}_l$, and $\mathbf{b}_l$ defined as above, the optimal coloring filter restricted to order p minimizing the reconstruction error variance, for any bit allocation, is obtained from $\mathbf{b}_l=(\mathbf{X}_l^T\mathbf{X}_l)^{-1}\mathbf{X}_l^T\mathbf{f}_l$ for all l.

Proof: To minimize σ_e^2 for a given FB over a class of coloring filters, from (6), the effective noise filter norms $\mathbf{f}_l'^T\mathbf{f}_l'$ should be minimized. Since α_l is specified by the quantizer, this is same as minimizing $(\alpha_l\mathbf{f}_l')^T(\alpha_l\mathbf{f}_l')$. But $\alpha_l\mathbf{f}_l'$ is the error $\mathbf{f}_l-\hat{\mathbf{f}}_l$. This error norm $(\mathbf{f}_l-\hat{\mathbf{f}}_l)^H(\mathbf{f}_l-\hat{\mathbf{f}}_l)$ is minimized for each l by optimally projecting $\mathbf{f}_l$ onto the column space of $\mathbf{X}_l$, which leads to the optimal $\mathbf{b}_l$ as stated [11]. Since minimization of the lth norm depends only on the lth column of $\mathbf{A}(z)$, minimization can be done independently for each l. Since (6) is for any bit allocation, the above solution is optimal for any bit allocation. ■

The minimum norm resulting from the optimal solution is $(\mathbf{f}_l'^T\mathbf{f}_l')_{\min}=[\mathbf{f}_l^T\mathbf{f}_l-\mathbf{f}_l^T\mathbf{X}_l(\mathbf{X}_l^T\mathbf{X}_l)^{-1}\mathbf{X}_l^T\mathbf{f}_l]/\alpha_l^2$. Since any PR synthesis filter length is M or more, the number of rows of $\mathbf{X}_l$ is greater than its number of columns. Since the columns of $\mathbf{X}_l$ are synthesis filters and their jM shifted versions, from the PR

nature of the FB, $\mathbf{X}_l$ has full column rank. Therefore, $\mathbf{X}_l^T\mathbf{X}_l$ is invertible and hence positive definite, so that

$$\mathbf{f}_l^T\mathbf{X}_l\left(\mathbf{X}_l^T\mathbf{X}_l\right)^{-1}\mathbf{X}_l^T\mathbf{f}_l \geq 0. \tag{10}$$

Therefore, $(\mathbf{f}_l'^T\mathbf{f}_l')_{\min} \leq \mathbf{f}_l^T\mathbf{f}_l/\alpha_l^2$. Before using any coloring filter, the synthesis filter norms from (1) and footnote 1 are $\mathbf{f}_l^T\mathbf{f}_l/\alpha_l^2$. With the optimal coloring filter, they are $(\mathbf{f}_l'^T\mathbf{f}_l')_{\min}$. This ensures that σ_e^2 cannot increase with the optimal coloring filter. The CG with coloring filter, with OBA (using effective noise filter norms), is obtainable from (1) by replacing the synthesis norms by the effective noise filter norms. The CG improvement, relative to no coloring filter case, is

$$\frac{G_{\text{color}}}{G} = \left(\prod_{i=0}^{M-1}\alpha_l^{-2}\mathbf{f}_i^T\mathbf{f}_i \Big/ \prod_{i=0}^{M-1}\mathbf{f}_i'^T\mathbf{f}_i'\right)^{1/M} \geq 1. \tag{11}$$

The equality of (10) is achieved if $\mathbf{X}_l^T\mathbf{f}_l$ is zero, or all filters in $\mathbf{X}_l$ are orthogonal to $\mathbf{f}_l$. This is true for orthogonal FBs. Thus, no reduction in σ_e^2 is possible for orthogonal case, and the optimal solution of $\mathbf{b}_l$ is zero. However, $\mathbf{X}_l^T\mathbf{f}_l$ is normally not zero for biorthogonal FBs, and reduction of QNA is ensured.

C. Sequence of Performing Quantization

In Section II-B, it is arbitrarily assumed that quantization is done starting at the bottom subband, and performed in that sequence toward top. The following theorem shows that any other choice of sequence will have the same performance (though the optimal $\mathbf{A}(z)$ will be different). This result is similar to [3], which shows that its CG remains same for any sequence, but the implementation complexity does not.

Theorem 2: The CG improvement obtained with optimal coloring filter $\mathbf{A}(z)$, restricted to order p, for a given FIR FB, when the OBA is done is independent of the sequence in which the subbands are quantized.

Proof: Define $\mathbf{F}'$ as shown in (12) at the bottom of the page, where $\mathbf{f}_l'$ are the optimal solutions found from Theorem 1, while $\mathbf{f}_l'^{(iM)}$ are the minimum residual errors in projecting $\mathbf{f}_l^{(iM)}$ onto all columns to its right in $\mathbf{X}$. For simplicity, rename the ith columns of $\mathbf{F}'$ for $i \geq M$ as $\mathbf{f}_i'$. Extending (9), we may write

$$\mathbf{F}' = \mathbf{X}\mathbf{B}' \tag{13}$$

where the ith column of $\mathbf{B}'$ for $0 \leq i < M$ is obtained from (8) and (9) to be $[\mathbf{0}_i^T 1 - \mathbf{b}_{M-1-i}^T]^T$. The remaining columns for $i \geq M$ have a similar structure since they also are optimal projection coefficients. Therefore, $\mathbf{B}'$ is a square lower triangular matrix with all diagonal elements equal to 1.

From (11), with OBA the CG depends on the product of the norms of $\mathbf{f}_i'$. $\mathbf{f}_i'$, ith column of $\mathbf{F}'$, is the residual after optimal projection onto the space spanned by all columns $>i$ of $\mathbf{X}$. From the orthogonality principle, this residual is orthogonal to the space. But any $\mathbf{f}_j'$ for $j > i$ is, from (13), a linear combination of columns $\geq j$ of $\mathbf{X}$. Hence, $\mathbf{f}_j'$ is from this space, and is orthogonal to $\mathbf{f}_i'$. Therefore, all the columns of $\mathbf{F}'$ are mutually orthogonal, or $\det(\mathbf{F}'^T\mathbf{F}') = \prod_{i<M}\alpha_i^2\prod_i \mathbf{f}_i'^T\mathbf{f}_i'$. Consequently, the product of the effective noise filter norms $\prod_{i=0}^{M-1}\mathbf{f}_i'^T\mathbf{f}_i'$ is

$$\frac{\det(\mathbf{F}'^T\mathbf{F}')}{\prod_{i<M}\alpha_i^2\prod_{i\geq M}\mathbf{f}_i'^T\mathbf{f}_i'} = \frac{\det(\mathbf{X}^T\mathbf{X})}{\prod_{i<M}\alpha_i^2\prod_{i\geq M}\mathbf{f}_i'^T\mathbf{f}_i'} \tag{14}$$

since $\det(\mathbf{F}'^T\mathbf{F}') = \det(\mathbf{B}'^T\mathbf{X}^T\mathbf{X}\mathbf{B}') = \det(\mathbf{X}^T\mathbf{X}), \mathbf{B}'$ is lower triangular with unity diagonal, or $\det(\mathbf{B}') = 1$.

Now, consider any other sequence of performing quantization. This is equivalent to doing a column permutation of the first M columns of $\mathbf{X}$ or taking $\mathbf{X}\mathbf{P}$ (where $\mathbf{P}$ is a permutation matrix), and then restricting the coloring filter to have the same upper triangular zeroth-order matrix. Let $\mathbf{f}_l''$ be the optimal solution from Theorem 1 for this case. Construct $\mathbf{F}''$ in the same way as (12) except replacing $\mathbf{f}_l'$ by $\mathbf{f}_l''$. Let $\mathbf{F}'' = \mathbf{X}\mathbf{P}\mathbf{B}''$. Note that $\mathbf{f}_i'' = \mathbf{f}_i'$ for $i \geq M$ since $\mathbf{P}$ permutes only the first M columns. By the same reasoning as above, the product of the new norms $\prod_{i=0}^{M-1}\mathbf{f}_i''^T\mathbf{f}_i''$ is $\det(\mathbf{F}''^T\mathbf{F}'')/(\prod_{i<M}\alpha_i^2\prod_{i\geq M}\mathbf{f}_i'^T\mathbf{f}_i')$. The denominator is identical, and the numerator equals $\det(\mathbf{B}''^T\mathbf{P}^T\mathbf{X}^T\mathbf{X}\mathbf{P}\mathbf{B}'') = \det(\mathbf{X}^T\mathbf{X})$ since $\det(\mathbf{P})$ is ± 1. Since this is the same as (14), the proof is complete. ■

D. Optimal Coloring Filter Without Order Restriction

Theorem 1 assumes $\mathbf{A}(z)$ to be FIR of order p. Below we find the optimal coloring filter without order restriction for a given FIR FB, which we show to be FIR of order no more than the analysis polyphase order. The sequence of performing quantization is taken to be from top to bottom (since performance is sequence independent with OBA) such that the zeroth-order matrix of $\mathbf{A}(z)$ is lower triangular.

Consider, an FIR synthesis polyphase matrix $\mathbf{R}(z)$ with monomial determinant. It is known that any such $\mathbf{R}(z)$ can be factorized into a paraunitary matrix $\mathbf{G}(z)$ and a unimodular matrix $\mathbf{U}(z)$ [12]

$$\mathbf{R}(z) = \mathbf{G}(z)\mathbf{U}(z). \tag{15}$$

Since the determinant of a unimodular matrix is unity, it always has causal inverse and hence the zeroth-order coefficient matrix of a unimodular matrix is always invertible. Let $\mathbf{U}^0$ be the zeroth-order coefficient of $\mathbf{U}(z)$. Let $\mathbf{U}^0 = \mathbf{Q}\mathbf{D}\mathbf{L}$ be the QR factorization of $\mathbf{U}^0$, where $\mathbf{Q}$ is unitary, $\mathbf{D}$ is diagonal, and $\mathbf{L}$ is lower triangular with ones along the diagonal. Define

$$\mathbf{S} = \mathbf{Q}\mathbf{D}. \tag{16}$$

The following theorem states the solution to the optimal coloring filter in this case. In the proof, we first show that this solution minimizes the *sum* of the effective noise filter norms. Then from the uniqueness of the solution of Theorem 1 we argue that both solutions are same. Therefore, this solution must be optimal.

$$\left[\alpha_{M-1}\mathbf{f}_{M-1}', \ldots, \alpha_0\mathbf{f}_0', \mathbf{f}_0'^{(M)}, \ldots, \mathbf{f}_{M-1}'^{(M)}, \ldots, \mathbf{f}_0'^{(pM)}, \ldots, \mathbf{f}_{M-1}'^{(pM)}\right] \tag{12}$$

Theorem 3: For some given FIR FB, the optimal coloring filter of unrestricted order is $\mathbf{U}^{-1}(z)\mathbf{S}$, where $\mathbf{U}(z)$ is the unimodular part of the synthesis polyphase matrix, and $\mathbf{S}$ is defined in (16).

Proof: Let $\mathbf{A}_{\rm opt}(z)$ be the claimed optimal filter $\mathbf{U}^{-1}(z)\mathbf{S}$ and let $\mathbf{A}(z)$ be any coloring filter. Since the effective noise filter is $\mathbf{G}(z)\mathbf{U}(z)\mathbf{A}(z)$, the sum of the effective noise filter norms is the trace of the zeroth-order coefficient matrix of $\tilde{\mathbf{A}}(z)\tilde{\mathbf{U}}(z)\tilde{\mathbf{G}}(z)\mathbf{G}(z)\mathbf{U}(z)\mathbf{A}(z)$ where $\tilde{}$ denotes para-conjugation. This equals $[\tilde{\mathbf{A}}(z)-\tilde{\mathbf{A}}_{\rm opt}(z)+\tilde{\mathbf{A}}_{\rm opt}(z)]\tilde{\mathbf{U}}(z)\mathbf{U}(z)[\mathbf{A}(z)-\mathbf{A}_{\rm opt}(z)+\mathbf{A}_{\rm opt}(z)]$ since $\mathbf{G}(z)$ is paraunitary. Further, it is equal to

$$[\tilde{\mathbf{A}}(z)-\tilde{\mathbf{A}}_{\rm opt}(z)]\tilde{\mathbf{U}}(z)\mathbf{U}(z)[\mathbf{A}(z)-\mathbf{A}_{\rm opt}(z)]+\tilde{\mathbf{A}}_{\rm opt}(z)\times\tilde{\mathbf{U}}(z)\mathbf{U}(z)[\mathbf{A}(z)-\mathbf{A}_{\rm opt}(z)]+[\tilde{\mathbf{A}}(z)-\tilde{\mathbf{A}}_{\rm opt}(z)]\cdot\tilde{\mathbf{U}}(z)\mathbf{U}(z)\mathbf{A}_{\rm opt}(z)+\tilde{\mathbf{A}}_{\rm opt}(z)\tilde{\mathbf{U}}(z)\mathbf{U}(z)\mathbf{A}_{\rm opt}(z). \quad (17)$$

Consider $\mathbf{A}(z)$ that minimizes each term of (17) separately.

1) The first term is positive semidefinite for all values of z. So, its zeroth-order coefficient matrix is also positive semidefinite. So its trace is nonnegative, and is minimum at $\mathbf{A}_{\rm opt}(z)$.
2) The fourth term is independent of $\mathbf{A}(z)$.
3) The second term is the para-conjugate of the third term. So the zeroth-order coefficient matrix of one is the hermitian transpose of the other. Since the filter coefficients are real valued, their trace is the same. The second term equals $\mathbf{S}^H[\mathbf{U}(z)\mathbf{A}(z)-\mathbf{U}(z)\mathbf{A}_{\rm opt}(z)]=\mathbf{S}^H\mathbf{S}[\mathbf{S}^{-1}\mathbf{U}(z)\mathbf{A}(z)-\mathbf{I}]$ where the simplifications are obtained by using $\mathbf{A}_{\rm opt}(z)$ from Theorem 3. Zeroth-order matrix of $\mathbf{S}^{-1}\mathbf{U}(z)$ is $\mathbf{S}^{-1}\mathbf{U}^0=\mathbf{L}$, and zeroth-order matrix of $\mathbf{A}(z)$ is also lower triangular with 1's along the diagonal. Further, from (16), $\mathbf{S}^H\mathbf{S}$ is diagonal. Therefore the zeroth-order matrix of the second term has zero diagonal, or its trace is zero.

This proves that $\mathbf{A}_{\rm opt}(z)$ minimizes the sum of the effective noise filter norms.

Since the given FB is FIR, $\mathbf{U}^{-1}(z)$ (and hence $\mathbf{A}_{\rm opt}(z)$) is FIR of order, say, p. p is at most equal to the order of $\mathbf{E}(z)$. From Theorem 1, the optimal coloring filter of order p for this quantization sequence minimizes each norm, hence minimizes the sum of the norms. From the derivation using optimal projections, the optimal coloring filter of Theorem 1 is unique for a particular sequence of quantization. Therefore, the solution obtained in Theorem 1 for order p must be $\mathbf{A}_{\rm opt}(z)$. This completes the proof. ■

Thus, the order p coloring filter of Theorem 1 becomes the overall optimal coloring filter of Theorem 3 for a sufficiently large p. The effective noise filter amplifying the quantization noise is $\mathbf{R}(z)\mathbf{A}(z)=\mathbf{G}(z)\mathbf{S}$ for the optimal coloring filter $\mathbf{A}_{\rm opt}(z)$, and while $\mathbf{G}(z)$ is orthogonal, $\mathbf{S}$ may result in some residual QNA depending on the given FB. Obviously, the complete elimination of QNA for some given FB is achieved if and only if $\mathbf{U}^0$, the zeroth-order coefficient of the unimodular part of the synthesis polyphase matrix (or $\mathbf{U}^{0-1}$, the zeroth-order coefficient of the unimodular part of the analysis polyphase), is lower triangular with 1's along the diagonal (or its permuted version). In this case, $\mathbf{S}=\mathbf{I}$ and $\mathbf{A}_{\rm opt}(z)$ is the unimodular part of $\mathbf{E}(z)$.

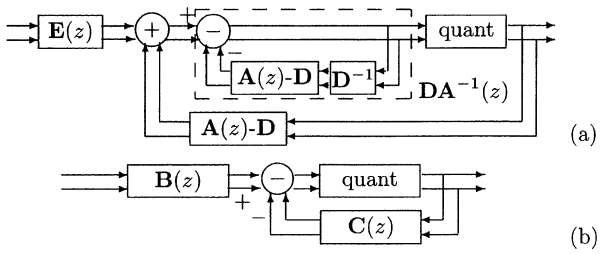


Fig. 4. Realizing the coloring filter. (a) Intermediate step. (b) Final step.

The ladder-based FB of [2] when the ladder elements are causal polynomials, and the transform of [3], are special cases of the proposed coloring filter. The analysis polyphase matrix is unimodular with zeroth-order coefficient matrix being lower triangular with 1's along the diagonal in [2]. The analysis transform is lower triangular with 1's along the diagonal in [3]. Therefore, from Theorem 3, the optimal coloring filter $\mathbf{A}_{\rm opt}(z)$ is the analysis polyphase matrix itself in both cases, completely eliminating QNA.

III. Low-Complexity Implementation of Optimal Coloring Filter

Realization of Fig. 3(b), henceforth referred to as the NF structure, requires additionally implementing the filter $\mathbf{A}_{\rm opt}(z)$ and hence causes an overhead at the analysis side. However, since $\mathbf{E}(z)$ is inverse of $\mathbf{R}(z)$, and $\mathbf{A}_{\rm opt}(z)$ is inverse of the unimodular part of $\mathbf{R}(z)$ except for a constant scaling matrix (from Theorem 3), there is some commonality between these two filters. Exploiting this, we propose another structure where a part of filtering of quantization noise is done along with analysis filtering itself.

Fig. 3(b) may be redrawn as shown in Fig. 4(a), where the coloring filter is realized separately for the unquantized and the quantized signals. The filter $1/[\mathbf{I}+(\mathbf{A}(z)-\mathbf{D})\mathbf{D}^{-1}]=\mathbf{D}\mathbf{A}^{-1}(z)$ for the unquantized signal may be moved before the adder and merged with both $\mathbf{E}(z)$ and $\mathbf{A}(z)-\mathbf{D}$ to obtain the structure of Fig. 4(b), referred as the quantized sample feedback (QSF) structure, where $\mathbf{B}(z)=\mathbf{D}\mathbf{A}^{-1}(z)\mathbf{E}(z)$ and $\mathbf{C}(z)=\mathbf{D}\mathbf{A}^{-1}(z)\mathbf{D}-\mathbf{D}$. As earlier, the zeroth-order coefficient matrix of $\mathbf{C}(z)$ is required to be strictly lower triangular (or a permutation of it for other sequencing) for scalar quantizers, or zero for vector quantizer. Note that the quantized samples are also being used in the analysis filtering. This is what causes the quantization noise to be colored.

The cost of implementation of NF and QSF structures for the optimal coloring filter is discussed now. We estimate the cost of implementing an M-input M-output transfer function of order N as realizing $N+1$ matrices of size $M\times M$, since it represents M filters of length $M(N+1)$ each (neglecting any zero coefficient, such as for the triangular matrix restriction).

Let $\mathbf{R}(z)$ be causal of order N_R, $\mathbf{R}(z)=\mathbf{G}(z)\mathbf{U}(z)$ from (15), and $|\mathbf{R}(z)|=z^{-N}$. Then, the orthogonal $\mathbf{G}(z)$ is of order N since $|\mathbf{G}(z)|=z^{-N}$. The maximum possible degree of $\mathbf{R}(z)$ is MN_R, so $N\le MN_R$. Let the unimodular $\mathbf{U}(z)$ be of order N_U. Since $\mathbf{R}(z)$ is the product of $\mathbf{G}(z)$ and $\mathbf{U}(z)$, $N_R\le N+N_U$. On the other hand, N_U is upper bounded by N_R [12]. Therefore, $\max\{0,N_R-N\}\le N_U\le N_R$.

TABLE I
IMPLEMENTATION COSTS OF VARIOUS STRUCTURES

structure	to realize	no. of matrix	use if
BC (fig.2)	$\mathbf{E}(z)$	N_E+1	
NF (fig.3b)	$\mathbf{E}(z), \mathbf{A}(z)$	$N_E+N_{U^{-1}}+2$	$N > N_R$
QSF (fig.4b)	$\mathbf{B}(z), \mathbf{C}(z)$	$N+N_U+2$	$N < N_R$

Choose $\mathbf{E}(z) = \mathbf{R}^{-1}(z) = \mathbf{U}^{-1}(z)\tilde{\mathbf{G}}(z)$ with Laurent polynomial order N_E. $\mathbf{E}(z)$ may be noncausal but it does not affect the cost. Orders of $\mathbf{U}^{-1}(z)$ and $\tilde{\mathbf{G}}(z)$ are $N_{U^{-1}}$ and N, and $N \le MN_E, \max\{0, N_E - N\} \le N_{U^{-1}} \le N_E$.

The cost of implementing the analysis side of the baseline coder (BC) with no coloring filter is realizing $\mathbf{E}(z)$ or realizing N_E+1 matrices. Cost of NF is realizing $\mathbf{E}(z)$ and $\mathbf{A}_{\text{opt}}(z)$. Since $\mathbf{A}_{\text{opt}}(z) = \mathbf{U}^{-1}(z)\mathbf{S}$ from Theorem 3, order of $\mathbf{A}_{\text{opt}}(z)$ is $N_{U^{-1}}$. Cost of QSF is realizing $\mathbf{B}(z) = \mathbf{D}\mathbf{S}^{-1}\tilde{\mathbf{G}}(z)$ and $\mathbf{C}(z) = \mathbf{D}\mathbf{S}^{-1}\mathbf{U}(z)\mathbf{D} - \mathbf{D}$, with orders N and N_U. The synthesis side is identical for all cases. These costs are compared in Table I.

While it is possible for N to exceed N_R, such a FB is rare. In such a case, the NF structure costs less than the QSF. For example, assume for some FB $N_R = N_E = N_U = N_{U^{-1}} = n, N = n+1$. Then, the NF costs $2n+2$ matrices while the QSF costs $2n+3$. For FBs where its delay N equals its order N_R, NF and QSF are typically equally efficient. For example, the Gandhi 8-8b FB [6] has $N_R = N_E = N = N_U = N_{U^{-1}} = 3$. Therefore, both NF and QSF costs eight matrices. Here, the overhead is substantial since the BC costs only four matrices.

For FBs having $N < N_R$, QSF typically is better than NF. The last column of Table I crudely recommends how to choose between NF and QSF. The lower is the delay N of a FB, the lesser QSF costs than NF. Further, the overhead for QSF compared to BC also becomes smaller. Take the lowest delay bank, or an unimodular FB, as an example. Gandhi 12-12a [6] has $N_R = N_E = N_U = N_{U^{-1}} = 5, N = 0$, therefore QSF requires only seven matrices whereas NF requires 12 matrices. There is still some overhead since BC needs six matrices.

Interestingly, the overhead may become zero or negative if the synthesis order is less than the analysis order. Take another unimodular example $N = 0$ with synthesis order $N_R = N_U = 2$ whose inverse is fourth order, $N_E = N_{U^{-1}} = 4$ (such as the example of (12)–(13) of [12] with z replaced by z^2). While BC requires five matrices, QSF needs four matrices in spite of the coloring filter. Thus it is possible that realizing a coloring filter may show an actual saving. Note that in all cases, NF has a higher cost than BC.

To summarize, the QSF structure is typically better than the NF structure unless the FB delay is more than its order. The coloring filter overhead is larger for larger delay banks, and is equal to the analysis cost at the worst (with NF structure). On the other extreme, it may actually save the analysis cost (with QSF structure).

IV. OPTIMAL BIORTHOGONAL CODER WITH COLORING FILTER

A. Optimal FB in the Ideal Case

In [1], it is reasoned that $\mathbf{E}(z)$ of any biorthogonal FB can be expressed as $\mathbf{\Lambda}(z)\mathbf{E}'(z)$ where $\mathbf{E}'(z)$ is paraunitary. Let $\mathbf{E}'(z)$ be any orthogonal solution that completely decorrelates the subbands. $\mathbf{\Lambda}(z)$ is chosen to be diagonal (in a fashion similar to the conventional biorthogonal coder of [8]). It is shown later that the diagonal restriction does not compromise on performance since this coder, indeed, achieves the maximum possible CG, ${}^{\infty}G$. The PR synthesis polyphase for this coder is $\tilde{\mathbf{E}}'(z)\mathbf{\Lambda}^{-1}(z)$. The coder structure is same as in Fig. 3(b). Here, the additive white noise model with uncorrelated noise (uniform quantizer at high rate) is used for the quantizers ($\mathbf{D} = \mathbf{I}$).

The zeroth-order coefficient matrix of $\mathbf{A}(z)$ is restricted to be upper triangular with 1's along the diagonal. Let $\lambda_i(z)$ be the ith diagonal element of $\mathbf{\Lambda}(z)$. The effective noise filter is $\tilde{\mathbf{E}}'(z)\mathbf{\Lambda}^{-1}(z)\mathbf{A}(z)$. Since $\tilde{\mathbf{E}}'(z)$ does not introduce any QNA, the QNA due to $\mathbf{\Lambda}^{-1}(z)$ is compensated by choosing $\mathbf{A}(z)$ equal to $\mathbf{\Lambda}(z)$. Since diagonal elements of $\mathbf{A}(z)$ are restricted to be monic causal polynomials (1 as the constant coefficient), all $\lambda_i(z)$ should be monic causal. The following theorem gives the optimal solution.

Theorem 4: The optimal solution to the ideal subband coder with analysis polyphase $\mathbf{\Lambda}(z)\mathbf{E}'(z)$ where $\mathbf{\Lambda}(z)$ is diagonal with monic causal entries $\lambda_i(z)$, and with optimal coloring filter $\mathbf{A}(z) = \mathbf{\Lambda}(z)$, is when $\lambda_i(z)$ is FW filter for the ith subband signal after $\mathbf{E}'(z)$.

Proof: Let $S_{y_iy_i}(e^{j\omega})$ be the psd of the ith subband $y_i(n)$ after $\mathbf{E}'(z)$ but before $\lambda_i(z)$. Then, the reconstruction error variance introduced at the output of the ith subband or the ith term on the right-hand side of (6) [note that (6) is same for any $\mathbf{D}$, even $\mathbf{D} = \mathbf{I}$] is

$$\sigma_{e_i}^2 = c2^{-2b_i} \int_{-\pi}^{\pi} S_{y_iy_i}(e^{j\omega}) |\lambda_i(e^{j\omega})|^2 \frac{d\omega}{2\pi} \qquad (18)$$

where $c2^{-2b_i}$ comes from the b_i-bit quantizer [9], the integral is the subband variance $\sigma_{x_i}^2$, and the effective noise filter norm is 1 due to complete elimination of QNA, $\mathbf{A}(z) = \mathbf{\Lambda}(z)$. $\lambda_i(z)$, a monic causal transfer function that minimizes the integral of (18) (hence $\sigma_{e_i}^2$) is known from the theory of linear prediction [9]. $\lambda_i(z)$ is the optimal linear predictor of $y_i(n)$, and in the ideal case $|\lambda_i(e^{j\omega})| = k_i S_{y_iy_i}^{-1/2}(e^{j\omega})$ where k_i^2 is the variance of the output of the ideal predictor when the input is $y_i(n)$. This indeed is the FW filter of $y_i(n)$. ■

Note that the CG depends only on the magnitude of $\lambda_i(z)$'s, their phase being irrelevant. So, by choosing $\lambda_i(z)$ as minimum phase (linear predictor), it can be ensured that stable inverse filter exists. Note that for this coder $\mathbf{A}(z) = \mathbf{\Lambda}(z)$, so $\mathbf{B}(z) = \mathbf{E}'(z)$ and $\mathbf{C}(z) = \mathbf{\Lambda}^{-1}(z) - \mathbf{I}$, and it may be realized efficiently using the QSF structure (which becomes DPCM on each subband due to diagonal $\mathbf{A}(z)$). Since $\mathbf{A}(z)$ is diagonal, the sequence of performing quantization does not affect the optimal solution.

While in the presence of QNA, the ideal biorthogonal coder only achieves half-whitening, the proposed ideal biorthogonal coder with coloring filter achieves FW. Therefore, we name it the *FW* coder [13]. Since FW of any subband amounts to its optimal linear prediction, the FW coder resembles the coder using linear prediction of subband signals. Such coder uses a subband decomposition (equivalent to $\mathbf{E}'(z)$ in the FW coder) followed by a linear prediction in each subband (equivalent to a diagonal $\mathbf{\Lambda}(z)$ in the FW coder). It is worth mentioning that such a coder

for ideal subband decomposition but finite-order prediction is dealt in [14], and for finite-order subband decomposition but ideal prediction is found in [15]. We now find the CG of the FW coder.

B. CG of the FW Coder

Recalling that $y_i(n)$ is the input to the shaping filter $\lambda_i(z)$ and $x_i(n)$ is the output, and that there is no QNA, the reconstruction error variance from (6) is $\sigma_e^2 = \sum_{i=0}^{M-1} c2^{-2b_i}\sigma_{x_i}^2/M$. Using OBA, it becomes equal to $c2^{-2\overline{b}}(\prod_{i=0}^{M-1}\sigma_{x_i}^2)^{1/M}$ where $\overline{b}$ is the average bit-rate. Since $\lambda_i(z)$ is the ideal linear predictor, $\sigma_{x_i}^2 = \gamma_{y_i}^2\sigma_{y_i}^2$ where $\sigma_{y_i}^2$ and $\gamma_{y_i}^2$ are the variance and spectral flatness measure of $y_i(n)$ having psd $S_{y_iy_i}(e^{j\omega})$. Therefore, the CG of the FW coder relative to a PCM coder having the same rate is $G_{\mathrm{FW}} = \sigma_x^2/(\prod_{i=0}^{M-1}\sigma_{y_i}^2\gamma_{y_i}^2)^{1/M}$. In the following theorem we show that G_{FW} is indeed the maximum possible CG.

Theorem 5: The FW coder attains the gain ${}^{\infty}G$.

Proof: Spectral flatness measure is defined as $\gamma_{y_i}^2 = \exp[\int_{-\pi}^{\pi}\ln S_{y_iy_i}(e^{j\omega})d\omega/2\pi]/\sigma_{y_i}^2$. Therefore G_{FW} may be written as $\sigma_x^2/\exp[(1/M)\sum_{i=0}^{M-1}\int_{-\pi}^{\pi}\ln S_{y_iy_i}(e^{j\omega})d\omega/2\pi]$. Let us consider the denominator. Let $S_{xx}(e^{j\omega})$ be the input psd defined on $\omega \in S = (-\pi,\pi]$. Since $\mathbf{E}'(z)$ completely decorrelates the subbands, the analysis filters are flat-top selecting nonoverlapping partitions S_i of S such that $S = \cup_i S_i$. From the Nyquist-M property of S_i it follows that $\int_S S_{y_iy_i}(e^{j\omega})d\omega/2\pi = \int_{S_i} MS_{xx}(e^{j\omega})d\omega/2\pi$ since $S_{y_iy_i}(e^{j\omega})$, the subband psd, is stretched and re-ordered $S_{xx}(e^{j\omega})$ in S_i. Therefore, $\int_S \ln S_{y_iy_i}(e^{j\omega})d\omega/2\pi = \int_{S_i} M\ln S_{xx}(e^{j\omega})d\omega/2\pi$. Since $S = \cup_i S_i$, we also have $\sum_{i=0}^{M-1}\int_{S_i}\ln S_{xx}(e^{j\omega})d\omega/2\pi = \int_S \ln S_{xx}(e^{j\omega})d\omega/2\pi$. From above two results, $(1/M)\sum_{i=0}^{M-1}\int_{-\pi}^{\pi}\ln S_{y_iy_i}(e^{j\omega})d\omega/2\pi = \int_{-\pi}^{\pi}\ln S_{xx}(e^{j\omega})d\omega/2\pi$. Thus $G_{\mathrm{FW}} = \sigma_x^2/\exp[\int_{-\pi}^{\pi}\ln S_{xx}(e^{j\omega})d\omega/2\pi] = {}^{\infty}G$. ■

Note that ${}^{\infty}G$ is achieved by the FW coder (with ideal filters) for any value of M, the number of subbands, unlike ${}^{\infty}G_{\mathrm{SBC}}$ of conventional subband coder B which achieves ${}^{\infty}G$ when M approaches infinity. Also, since the orthogonal part $\mathbf{E}'(z)$ of the FW coder does not need majorization unlike the optimal orthogonal FB, filters can be contiguous which are simpler. Obviously, G_{FW} is superior to the ideal biorthogonal coder [8]. It may be shown that CGs of ideal orthogonal coder and ideal biorthogonal coder are both equal to G_{FW} only if all $\gamma_{y_i}^2$ for the optimal orthogonal analysis FB are 1 (all subband psds are flat).

V. Simulation Results

A. Performance for Existing FBs

Using Theorem 3, the optimal coloring filter and the CG improvement G_{color}/G for several existing 2-channel biorthogonal FBs has been found. For a 2-channel FB, we use the Lightstone measure of nonorthogonality [4], $\int_0^{\pi}(|H_0(e^{j\omega})|^2 + |H_1(e^{j\omega})|^2 - 2)^2 d\omega$, where $H_i(e^{j\omega})$ are the analysis filters scaled to have unit energy. For orthogonal FBs this measure is 0. Since an orthogonal FB has no CG improvement, one may be curious to know how the CG improvement behaves with increasing nonorthogonality measure of an FB. Table II shows the CG improvement of (11) using the coloring filter relative to no coloring filter case, and also the measure of nonorthogonality, for some existing FBs (Moulin 3-9 [5] is same as Egger 3-9).

TABLE II
Theoretical Coding Gain Improvement Over Existing FBs

filter bank	CG improvement	non-orthogonality
Egger 3-9 [16]	0.1288 dB	0.3913
Egger 4-12 [16]	0.5195 dB	1.2874
LeGall 3-5 [17]	0.1633 dB	0.3887
Moulin 1-3 [5]	0.8805 dB	2.9671
Moulin 5-11 [5]	1.2689 dB	2.5587
Gandhi 12-12a [6]	1.0849 dB	1.4535
Gandhi 12-12b [6]	0.0835 dB	0.0082
Gandhi 8-8a [6]	0.7017 dB	0.4755
Gandhi 8-8b [6]	8.02×10^{-8} dB	1.21×10^{-9}

Simulation of these FBs, without and with $\mathbf{A}_{\mathrm{opt}}(z)$, is performed on actual samples of an AR(1) source with $\rho = 0.95$ for various rates. In general, the obtained CG improvement values approach the theoretical values of Table II. For example, Fig. 5 illustrates the variation of SNR with rate for the FBs Egger 4-12, Moulin 1-3, Moulin 5-11 and Gandhi 12-12a. The CG improvements obtained in this work are superior to that of the iterative scheme as well as the closed-loop scheme, and are comparable to that of the trellis-based scheme, of [6], [7]. However, the proposed coder's complexity is less than the iterative scheme and much less compared to the trellis-based scheme.

B. Performance of Coloring Filter With Order Restriction

Using Theorem 1, the FIR coloring filter for different orders for the existing FBs is found. Fig. 6 shows the variation of the theoretical CG improvement with the order of the coloring filter for the FBs Moulin 5-11 and Gandhi 12-12a. It is worth noting that excellent improvement is obtained with first-order $\mathbf{A}(z)$.

C. FW Coder Performance

In the second column of Table III, theoretical CGs of various ideal coders are presented for $M = 2$ channels for the earlier AR(1) source. For this source, ${}^{\infty}G = 1/(1-\rho^2) = 10.11$ dB. The proposed FW coder gives more than 4 dB improvement over the orthogonal coder, and about 2 dB improvement over the half-whitening coder.

While the FW coder requires ideal filters, it is of interest to find the achievable performance for its finite-order approximation. As the third column of Table III illustrates, the CG improvements obtained for ideal case are true even for the FIR FBs. Note also that the performance gap between second and third columns is only 0.3 dB for FW case with moderate-order filters. The orthogonal FB is of order 11 and approximates the uniform contiguous brick-wall filter shape, since for this source the ideal orthogonal FB is the uniform contiguous FB. The half-whitening FB uses the same FIR orthogonal filters cascaded with diagonal half-whitening filters of order 3 (optimal linear predictors for the square-root of the subband *psd*s). The FW FB uses the same FIR orthogonal filters cascaded with diagonal FW filters of order 3 (optimal linear predictors for the subband *psd*s), and $\mathbf{A}(z) = \mathbf{\Lambda}(z)$. Fig. 7 shows the analysis filter responses of the above three coders for finite-order case. The dip of the low-pass analysis filter near zero frequency clearly shows the

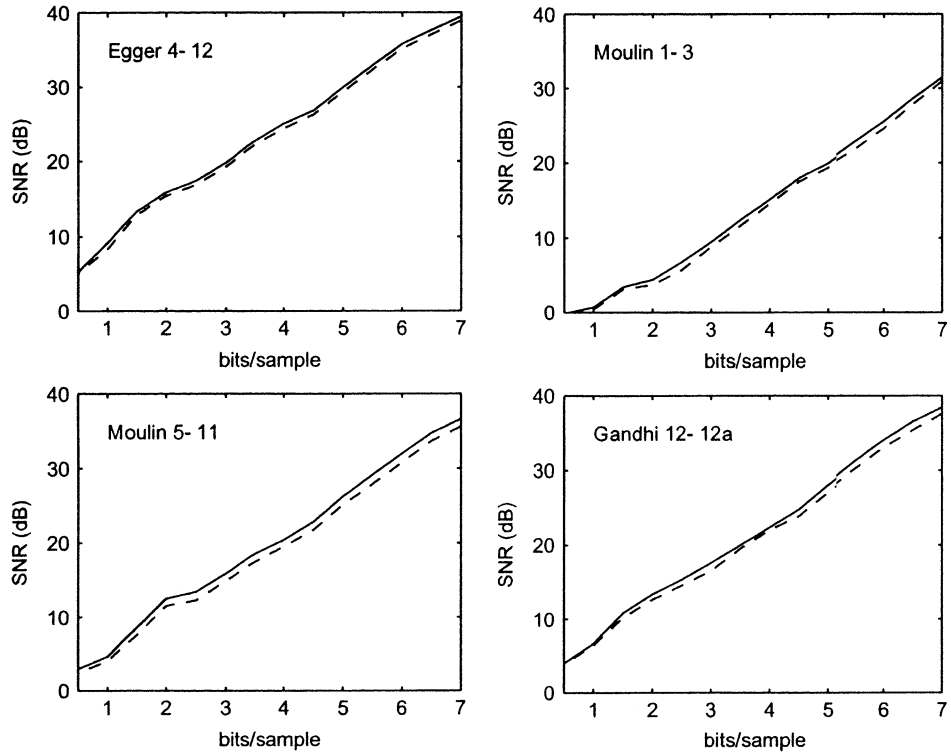


Fig. 5. SNR-rate plot without (dash) and with (solid) coloring filter.

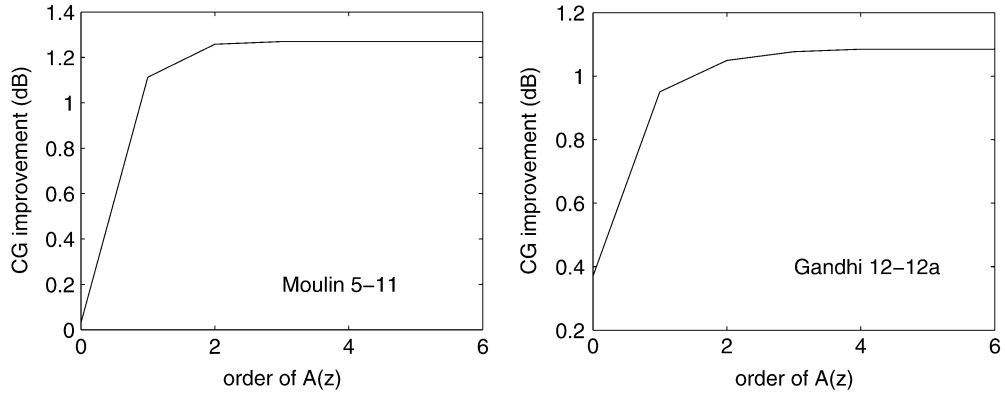


Fig. 6. CG improvement versus order of coloring filter.

TABLE III
THEORETICAL CODING GAIN COMPARISON

coder	ideal CG	FIR CG
orthogonal	$G_O = 5.96$ dB	5.56 dB
half-whitening [8]	$G_B = 8.16$ dB	7.75 dB
full whitening	$G_{FW} = 10.11$ dB	9.78 dB

whitening action of the latter two coders. The FW filter has almost twice as much dip in dB as the half-whitening filter.

The above results show that the FW idea may be used to increase the CG of a given orthogonal FB. First, the subband psds after the orthogonal FB $S_{y_i y_i}(z)$ may be found, and a diagonal $\mathbf{\Lambda}(z)$ with FIR $\lambda_i(z)$ may be designed from the optimal linear predictor obtained by solving the normal equations or the Levinson Durbin recursion [9]. For unknown/nonstationary input, adaptive linear prediction may be used along with overhead bits to transmit the predictor coefficients to the synthesis side. Second, the optimal $\mathbf{A}(z)$ for this FB is equal to $\mathbf{\Lambda}(z)$. Such a coder will give more CG, since $\lambda_i(z)$ achieves additional CG due to shaping, and there is no QNA. Further, the computational overhead is only in implementing $\lambda_i(z)$ as seen from Section III. The third column of Table III illustrates that the CG of an orthogonal FB (of order 11) is increased by more than 4 dB by adding $\lambda_i(z)$ and $\mathbf{A}(z)$ (of order 3). Since such CG is not possible even in infinite-order conventional subband coders, we believe this strategy helds great potential in signal compression.

Fig. 8 shows the SNR versus rate plots for the above three coders (finite-order case) as well as the PCM coder for the earlier source. When applied on actual signal samples, the obtained CG improvements are verified in Fig. 8 to be the same as before. Simulations in this section use quantizers with uniform decision intervals and centroid reconstruction levels, and positive integer approximation to OBA obtained using additive noise model of quantizer and effective noise filter norms. The obtained CG may

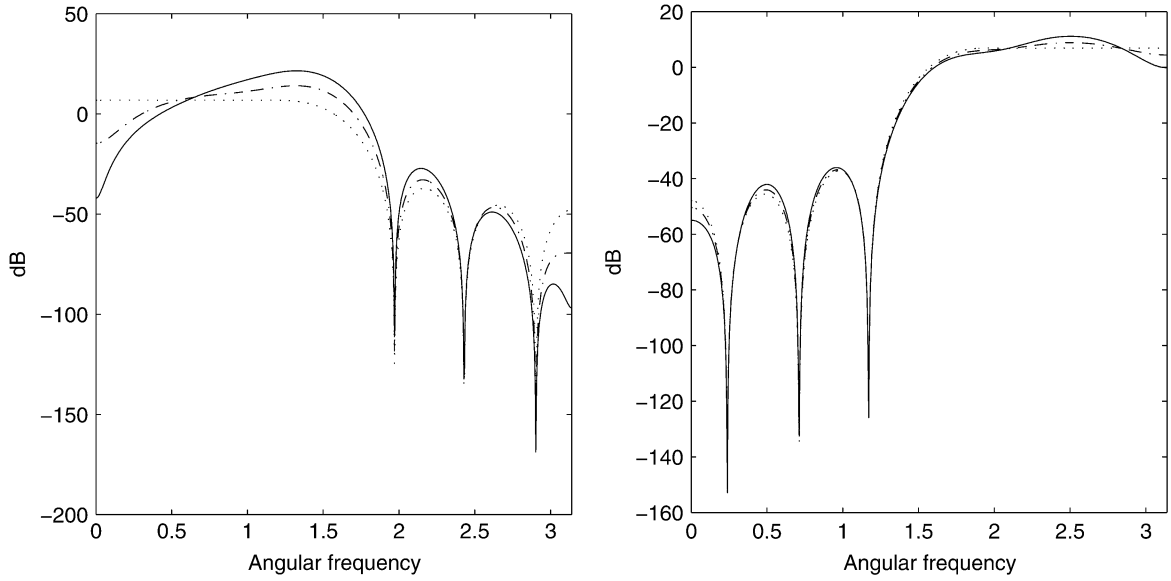


Fig. 7. Analysis filter response ($\mathrm{left} = \mathrm{lpf}, \mathrm{right} = \mathrm{hpf}$) of orthogonal (dot), half-whitening (dash-dot), and FW (solid) coders.

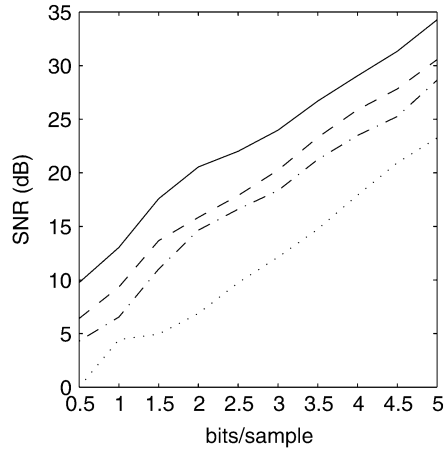


Fig. 8. SNR versus rate for the PCM (dot), orthogonal (dash-dot), half-whitening (dash), and FW (solid) coders.

differ from the theoretical one, especially at low rates. This is because the quantizer model, the high rate quantizer assumption for signal statistics (the variance of a subband plus colored quantization noise is same as the variance of the subband alone, similar to [2]), and the positive integer approximation of bit allocation, becomes less accurate. Note that the CG advantage is obtained even at low rates.

VI. Conclusion

A coloring filter $\mathbf{A}(z)$ is introduced to minimize the QNA in biorthogonal subband coders. For a given FIR FB, the optimal coloring filter of a given order and of unrestricted order is found. With OBA, the CG improvement is independent of the sequence in which the subbands are quantized. Feeding back the quantized samples in the analysis FB gives a low complexity structure for this coloring filter. The ideal biorthogonal FB with such coloring filter is found to perform FW and to attain the maximum possible CG ${}^{\infty}G$. Simulation on AR source shows that appreciable CG improvement is obtained using coloring filter with several existing FIR biorthogonal FBs even at low rates. The CG improvement of the FW coder is shown to be significant and is maintained even for finite-order FBs operating at low rates.

References

[1] P. P. Vaidyanathan and A. Kirac, "Results on optimal biorthogonal filter banks," *IEEE Trans. Circuits Syst. II*, vol. 45, pp. 932–947, Aug. 1998.

[2] S. Phoong and Y. Lin, "MINLAB: Minimum noise structure for ladder-based biorthogonal filter banks," *IEEE Trans. Signal Processing*, vol. 48, pp. 465–476, Feb. 2000.

[3] ——, "Prediction-based lower triangular transform," *IEEE Trans. Signal Processing*, vol. 48, pp. 1947–1955, July 2000.

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

[5] P. Moulin, "A multiscale relaxation algorithm for SNR maximization in nonorthogonal subband coding," *IEEE Trans. Image Processing*, vol. 4, pp. 1269–1281, Sept. 1995.

[6] R. Gandhi, "Filter Bank Design and quantization techniques for subband coding," Ph.D. thesis, Univ. California, Santa Barbara, CA, Mar. 2000.

[7] R. Gandhi and S. K. Mitra, "Quantization to maximize SNR in nonorthogonal subband coders," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, May 2001, pp. 3689–3692.

[8] P. Moulin, M. Anitescu, and K. Ramchandran, "Theory of rate-distortion-optimal, constrained filterbanks-application to IIR and FIR biorthogonal designs," *IEEE Trans. Signal Processing*, vol. 48, pp. 1120–1131, Apr. 2000.

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

[10] J. Kovacevic, "Subband coding systems incorporating quantizer models," *IEEE Trans. Image Processing*, vol. 4, pp. 543–553, May 1995.

[11] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[12] P. P. Vaidyanathan, "How to capture all FIR perfect reconstruction QMF banks with unimodular matrices," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 1990, pp. 2030–2033.

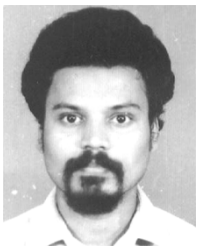
[13] M. Arunkumar and A. Makur, "Optimal biorthogonal filter banks with minimization of quantization noise amplification," in *Proc. 11th IEEE SSP Workshop*, Aug. 2001, pp. 603–606.

[14] S. Rao and W. A. Pearlman, "Analysis of linear prediction, coding, and spectral estimation from subbands," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1160–1178, July 1996.

[15] S.-L. Tan and T. R. Fischer, "Linear prediction of subband signals," *IEEE J. Select Areas Comm.*, vol. 12, pp. 1576–1583, Dec. 1994.

[16] O. Egger and W. Li, "Subband coding of images using asymmetrical filter banks," *IEEE Trans. Image Processing*, vol. 4, no. 4, pp. 478–485, Apr. 1995.

[17] D. LeGall and A. Tabatabai, "Subband coding of images using symmetric short kernel filters and arithmetic coding techniques," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1988, pp. 761–764.



Anamitra Makur received the B.Tech. degree from the Indian Institute of Technology, Kharagpur, India, the M.S. and Ph.D. degrees (in 1990) from the California Institute of Technology, Pasadena.

He worked in the Departement of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India, until 2002 in various capacities, the most recent being a Professor. He also held visiting positions at the University of California, Santa Barbara, and the University of Kaiserslautern, Germany. Since 2002, he is an Associate Professor at the Nanyang Technological University, Singapore. His current research interests include image/video/signal compression, subband coding, filterbank design, watermarking, and image/video processing.

Dr. Makur was awarded the 1998 Young Engineer Award from the Indian National Academy of Engineering, and is currently an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING.

M. Arunkumar received the B.Tech. degree in electronics and communication from the College of Engineering, Trivandrum, India, in 1998, and the M.E. degree in signal processing from the Indian Institute of Science, Bangalore, India, in 2001. He is currently working toward the Ph.D. degree at University of Maryland, College Park.

His research interests include source coding, signal compression, image processing, and recovering three-dimensional structure from two-dimensional images.