Rate-Distortion-Optimal Subband Coding without Perfect-Reconstruction Constraints *†

M. Kıvanç Mıhçak[1], Pierre Moulin[1], Mihai Anitescu[2],
and Kannan Ramchandran[1]

[1] University of Illinois at Urbana-Champaign
Beckman Institute and ECE Department
405 N. Mathews Ave., Urbana, IL 61801

[2] University of Pittsburgh, Department of Mathematics
Thackeray 301, Pittsburgh, PA 15260

Contact Author: P. Moulin, tel: (217) 244-8366, fax: (217) 244-8371,
E-mail: *moulin@ifp.uiuc.edu*

May 31, 1999

EDICS Number: SP 2.4

Abstract

We investigate the design of subband coders without the traditional perfect-reconstruction constraint on the filters. The coder uses scalar quantizers, and its filters and bit allocation are designed so as to optimize a rate-distortion criterion. Convexity properties play a central role in the analysis. Our results hold for a broad class of rate-distortion criteria. First, we show that optimality can be achieved using filter banks that are the cascade of a (paraunitary) principal component filter bank for the input spectral process and a set of pre- and post-filters surrounding each quantizer. An algorithm for computing the globally optimal filters and bit allocation is given. We then develop closed-form solutions for the special case of two-channel coders under an exponential rate-distortion model. Finally, we investigate a constrained-length version of the filter design problem, which is applicable to practical coding scenarios. While the optimal filter banks are nearly perfect-reconstruction at high rates, we demonstrate an apparently surprising advantage of optimal FIR filter banks: they significantly outperform optimal perfect-reconstruction FIR filter banks at all bit rates.

*Work supported by NSF grant MIP-97-07633 (PM and KR) and by the Mathematical, Information and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, U.S. Department of Energy, under Contract W-31-109-Eng-38 (MA).

†Work presented in part at the 32nd CISS conference, Princeton, NJ, March 1998, and at the IEEE Symposium on Time–Frequency and Time–Scale Analysis, Pittsburgh, PA, Oct. 1998.

1 Introduction

Following the discovery of perfect-reconstruction (PR) filter banks in the mid-eighties [1, 2, 3, 4], subband coding research has focused primarily on the design and optimization of PR systems. The emergence of wavelet theory provided new insights into the design of PR filter banks and helped sustain this trend. Yet the use of subband coders without PR constraints is not without advantages. The flexibility provided by the use of such a broad class of filters means that subband coders without PR constraints generally outperform subband coders that use PR filter banks. This advantage has been well documented in research using statistical signal models. In a 1988 paper, Dembo and Malah [5] used modulated analysis filter banks and designed synthesis filters that minimize the mean-squared error of the reconstructed signal under a broad class of quantizer models. More recently, Kovačević [6] used (fixed-length) Lloyd-Max quantizers and designed synthesis filters in such a way that the reconstruction error is uncorrelated with the signal itself. Haddad and Park [7] also used Lloyd-Max quantizers, designed a set of scaling factors (which they called compensators) for synthesis filters, and minimized the mean-squared reconstruction error over these factors. Vaidyanathan and Chen [8] gave a Wiener filtering interpretation for such filter design problems. Gosse and Duhamel [9] considered a broader class of scalar quantizers, including uniform quantizers and entropy-constrained scalar quantizers, which have variable length and are optimal in a rate-distortion sense. They developed numerical methods for optimizing Finite Impulse Response (FIR) synthesis filter banks, given the analysis filter banks. Tuqan and Vaidyanathan [10] investigated the design of pre- and post-filters for *single-channel* quantization systems. As described in [10], under some assumptions the scalar quantizer could be replaced with an orthonormal subband coder. However, that special setup does not lend itself to joint optimization of the pre- and post-filters, orthonormal filter banks, and bit allocation.

The problem of interest in our paper is to design the analysis and synthesis filters and the bit allocation that jointly optimize the rate-distortion performance of subband coders that use scalar quantizers in each channel. The distortion measure is mean-squared reconstruction error. For a fixed bit budget, the resulting filter banks are termed minimum-mean-squared-error (MMSE) filter banks. Unlike previous work, our focus is on identifying basic statistical properties and deriving analytical expressions for the optimal filter banks and optimal bit allocation. These expressions are given in terms of the overall bit budget and second-order statistics of the input signal. Hence our objectives are similar to those in [11], where an analysis was developed for subband coders with PR constraints. The absence of PR constraints introduces interesting new challenges because the distortion-rate function includes both a signal and a quantization noise term, and several key tools from [11] are not applicable here.

This paper is organized as follows. Our working assumptions and the rate–distortion criterion for subband coder design are introduced in Sec. 2. Some fundamental statistical properties of the optimal filter banks are derived on Sec. 3; our first main result is Theorem 3.3, which shows that the optimal coder takes the form of the cascade of a principal component filter bank (PCFB) [12, 13] and a set of pre- and post-filters surrounding each quantizer. Expressions for these optimal filters are derived in Theorem 3.4, and an algorithm that computes the jointly optimal filters and bit allocation is given. The results are specialized to coders under an exponential rate-distortion model in Sec. 4. In Sec. 5, we present numerical results that illustrate our theoretical analysis and demonstrate some remarkable advantages of FIR filter banks over their PR counterparts. Conclusions are given in Sec. 6. Details of most proofs are given in the Appendix.

2 Statement of the Problem

2.1 Subband Coder Model

We consider M-channel subband coders with analysis filters $\{H_i(f), 0 \leq i < M\}$ and synthesis filters $\{\tilde{H}_i(f), 0 \leq i < M\}$. Figure 1 shows an equivalent representation of the codec in terms of the $M \times M$ analysis and synthesis polyphase matrices $\mathcal{H}(f)$ and $\tilde{\mathcal{H}}(f)$ [14]. Throughout, we use script fonts for polyphase matrices.

The input $x(n)$ to the subband coder is assumed to be real-valued, Gaussian, and wide-sense stationary, with zero mean and spectral density $S(f)$. Throughout, we assume that $S(f)$ is bounded away from zero. The total bit budget is R bits per sample, to be allocated to the quantizers in each channel. Quantizer Q_i in channel i operates on a signal $y_i(n)$ with variance σ_i^2, is scalar, and is allocated R_i bits, where $\frac{1}{M}\sum_{i=0}^{M-1} R_i = R$. We assume that the quantization noise is additive, white, and independent of the signal and that the quantization noise sources in different channels are mutually independent. This is a standard model that is valid for uniform quantizers at high bit rates, but not at low bit rates. (See [9] for more sophisticated quantization models at low bit rates.) Hence we assume that all R_i are large and positive. Since $x(n)$ is zero-mean Gaussian, all (unquantized) subband signals $y_i(n)$ are also zero-mean Gaussian. Each quantizer is assumed to have a distortion rate function $\sigma_i^2\overline{D}(R_i)$, where in this context, distortion is quantization noise variance, and the distortion–rate function $\overline{D}(.)$ for a unit-variance Gaussian random variable is *strictly positive, strictly monotonic decreasing*, and *strictly convex*. We also require $\ln \overline{D}(.)$ to be *concave*. The standard exponential model $\overline{D}(R_i) = \gamma\sigma_i^2 2^{-2R_i}$ for the rate-distortion function satisfies all of these assumptions and will be considered in some detail.

2.2 Design Criterion

Under the assumptions of Sec. 2.1, the reconstruction error $\hat{x}(n) - x(n)$ is a cyclostationary process with period M. For PR systems, the expected mean-squared error (MSE) takes the form [11]

$$\frac{1}{M}\sum_{n=0}^{M-1} E|\hat{x}(n) - x(n)|^2 = \frac{1}{M}\sum_{i=0}^{M-1} \overline{D}(R_i)\sigma_i^2||\tilde{h}_i||^2, \tag{2.1}$$

where $||\tilde{h}_i||^2$ represents the amplification factor for white noise passed through synthesis filter $\tilde{h}_i$.

In the MMSE filter bank approach, the PR constraints are relaxed so as to trade off systematic reconstruction errors (due to lack of PR) against quantization noise. The solution is *nearly* identical to the PR solution at high bit rates (low quantization noise), but notable improvements over PR designs have been demonstrated numerically at lower bit rates [9]. Here, we seek analytical expressions for the filters and bit allocation $\{R_i\}$ that jointly minimize the MSE. Let $\mathcal{S}(f)$ be the $M \times M$ spectral density matrix for the polyphase vector $\underline{x}(n)$, input to $\mathcal{H}(f)$ in Fig. 1. According to the model in Sec. 2.1, signal and quantization noise are independent. Hence, the MSE is the sum of the noise term (2.1) due to quantization and a signal term due to lack of PR:

$$\mathcal{E} = \frac{1}{M}\sum_{i=0}^{M-1} \overline{D}(R_i)\sigma_i^2||\tilde{h}_i||^2 + \frac{1}{M} Tr \int_{-0.5}^{0.5} (\tilde{\mathcal{H}}(f)\mathcal{H}(f) - I_M)\mathcal{S}(f)(\tilde{\mathcal{H}}(f)\mathcal{H}(f) - I_M)^\dagger \, df, \tag{2.2}$$

where

$$\begin{aligned} \sigma_i^2 &= \int_{-0.5}^{0.5} (\mathcal{H}(f)\mathcal{S}(f)\mathcal{H}(f)^\dagger)_{ii} \, df, \quad 0 \le i < M, \\ ||\tilde{h}_i||^2 &= \int_{-0.5}^{0.5} (\tilde{\mathcal{H}}^\dagger(f)\tilde{\mathcal{H}}(f))_{ii} \, df, \quad 0 \le i < M, \end{aligned} \tag{2.3}$$

and the superscript $\dagger$ stands for the Hermitian transpose operator. Hence the mathematical problem is to minimize the functional $\mathcal{E}$ over $\mathcal{H}$, $\tilde{\mathcal{H}}$, and $\{R_i\}$, subject to the constraint $R = \frac{1}{M}\sum_{i=0}^{M-1} R_i$. A solution is guaranteed to exist, because the admissible set of $\left(\mathcal{H}, \tilde{\mathcal{H}}, \{R_i\}\right)$ is nonempty and closed, and the cost function (2.2) is bounded from below. Also note that while $\mathcal{E}$ is strictly convex in $\{R_i\}$ (for fixed filters $\mathcal{H}$, $\tilde{\mathcal{H}}$) and quadratic in both $\mathcal{H}$ and $\tilde{\mathcal{H}}$ (for fixed $\{R_i\}$), $\mathcal{E}$ is not convex in $\left(\mathcal{H}, \tilde{\mathcal{H}}, \{R_i\}\right)$ jointly. This makes the joint optimization problem harder to solve.

2.3 Optimal Bit Allocation

In order to find the optimal bit allocation, the optimization problem (2.2) with bit rate constraints is transformed into the Lagrange optimization problem:

$$\text{Minimize } \mathcal{E} = \frac{1}{M}\sum_{i=0}^{M-1}\overline{D}(R_i)\sigma_i^2||\tilde{h}_i||^2 + \frac{1}{M} Tr\int_{-0.5}^{0.5}(\tilde{\mathcal{H}}(f)\mathcal{H}(f) - I_M)\mathcal{S}(f)(\tilde{\mathcal{H}}(f)\mathcal{H}(f) - I_M)^\dagger\, df - \mu\frac{1}{M}\sum_{i=0}^{M-1} R_i$$

over $\mathcal{H}, \tilde{\mathcal{H}}, \{R_i\}$, where $-\mu \geq 0$ is the Lagrange multiplier. For any choice of filters $\left(\mathcal{H}, \tilde{\mathcal{H}}\right)$, the optimal bit allocation $\{R_i\}$ satisfies the condition

$$\sigma_i^2||\tilde{h}_i||^2 \left.\frac{d\overline{D}(R)}{dR}\right|_{R_i} = \mu, \quad 0 \leq i < M. \tag{2.4}$$

Hence, the slope of the distortion-rate function $\frac{1}{M}\overline{D}(R_i)\sigma_i^2||\tilde{h}_i||^2$ at the encoder's operating point is the same for all channels. Here we have made the standard assumption that R is large enough so that the positivity constraints $R_i \geq 0$ are all inactive (i.e., $R_i > 0$ for $0 \leq i < M$). Because of the strict convexity of $\overline{D}(.)$, the optimal bit allocation condition (2.4) yields a unique solution, for fixed filters $\mathcal{H}$, $\tilde{\mathcal{H}}$.

3 Fundamental Properties of Optimal Subband Coder

3.1 Structure of Optimal Subband Coder

We have recently proven that filter banks that are optimal with respect to the PR criterion (2.1) enjoy two fundamental properties: total decorrelation of the subband channels, and spectral majorization [11]. These properties were previously known to apply only to paraunitary filter banks, in which case the solution is a PCFB [12, 13, 15]. Propositions 3.1 and 3.2 below show that these fundamental properties hold even when the PR constraint is relaxed and the appropriate cost function (2.2) is used. These results enable us to derive the structure of the optimal subband coder in Theorem 3.3.

Specifically, Proposition 3.1 implies that the $M \times M$ spectral density matrix $\mathcal{S}_y(f)$ for the vector $\underline{y}(n)$ of subband signals is diagonal; in other words, the optimal coder statistically decouples the subbands. Proposition 3.2 asserts that the spectral densities $\mathcal{S}_{y,ii}(f)$ for the individual subband signals $y_i(n)$ satisfy a certain ordering condition.

Proposition 3.1 (Total Decorrelation Is Necessary for Optimality.) *The system $\mathcal{H}, \tilde{\mathcal{H}}$, $\{R_i\}$ minimizes the MSE (2.2) only if the matrices $\mathcal{S}_y = \mathcal{H}\mathcal{S}\mathcal{H}^\dagger$ and $\tilde{\mathcal{H}}^\dagger\tilde{\mathcal{H}}$ are diagonal and $\{R_i\}$ satisfies (2.4).*

Proposition 3.2 (Spectral Majorization Is Necessary for Optimality.) *Let $\mathcal{H}, \tilde{\mathcal{H}}$, $\{R_i\}$ be a minimizer of the MSE (2.2), and $\mathcal{M} = \tilde{\mathcal{H}}^\dagger\tilde{\mathcal{H}}$. Without loss of generality, assume that $R_0 \geq R_1 \geq$*

$\cdots \geq R_{M-1}$. The normalized spectral densities $\frac{1}{\overline{D}(R_i)\sigma_i^2}\mathcal{S}_{y,ii}(f)$ for the subband signals satisfy the spectral majorization property

$$\frac{1}{\overline{D}(R_0)\sigma_0^2}\mathcal{S}_{y,00}(f) \geq \cdots \geq \frac{1}{\overline{D}(R_{M-1})\sigma_{M-1}^2}\mathcal{S}_{y,M-1,M-1}(f), \quad \forall f.$$

Likewise, the normalized quantities $\frac{1}{\overline{D}(R_i)||\tilde{h}_i||^2}\mathcal{M}_{ii}(f)$ satisfy the spectral majorization property

$$\frac{1}{\overline{D}(R_0)||\tilde{h}_0||^2}\mathcal{M}_{00}(f) \geq \cdots \frac{1}{\overline{D}(R_{M-1})||\tilde{h}_{M-1}||^2}\mathcal{M}_{M-1,M-1}(f), \quad \forall f.$$

The proof of Propositions 3.1 and 3.2 uses variational techniques and is briefly outlined below. The methodology parallels the proofs of Theorem 2.3 and Lemma 2.6 in [11]. The idea is to introduce *admissible variations* of the filters around the optimal solution, from which it can be inferred that *total decorrelation* and *spectral majorization* are necessary conditions for optimality. Each variation is a modification of the analysis and synthesis polyphase matrices, $\mathcal{H}$ and $\tilde{\mathcal{H}}$, by polyphase matrices $\mathcal{G}$ and $\mathcal{G}^{-1}$, respectively. The modified analysis and synthesis polyphase matrices under such variations are $\mathcal{H}_{new} = \mathcal{H}\mathcal{G}$ and $\tilde{\mathcal{H}}_{new} = \mathcal{G}^{-1}\tilde{\mathcal{H}}$, respectively. Note that under this variation, $\mathcal{H}_{new}\tilde{\mathcal{H}}_{new} = \mathcal{H}\tilde{\mathcal{H}}$, so the second term of the cost function (2.2) (signal term, due to lack of PR) remains unchanged.

Therefore, if $\left(\mathcal{H}, \tilde{\mathcal{H}}\right)$ is optimal, no variation $\left(\mathcal{H}\mathcal{G}, \mathcal{G}^{-1}\tilde{\mathcal{H}}\right)$ can possibly increase the first term of (2.2), $\frac{1}{M}\sum_{i=0}^{M-1}\overline{D}(R_i)\sigma_i^2||\tilde{h}_i||^2$. But this is precisely the cost function used in [11]. So the procedure used to prove Theorem 2.3 and Lemma 2.6 in [11] directly applies. We conclude that total decorrelation and spectral majorization are necessary conditions for optimality under the cost function (2.2). ■

We now show that the design in Fig. 2 is optimal. This result, which is stated more precisely in Theorem 3.3, is a fundamental property of optimal filter banks without PR constraints. All of our subsequent analysis is based on this theorem. The fact that $\mathcal{U}(f)$ is a PCFB means that the output $\underline{z}(n)$ of $\mathcal{U}(f)$ satisfies the spectral majorization property:

$$\mathcal{S}_{z,00}(f) \geq \cdots \geq \mathcal{S}_{z,M-1,M-1}(f).$$

Theorem 3.3 *Let $\mathcal{S} = \mathcal{V}\Lambda\mathcal{V}^\dagger$ be the eigenvector decomposition of $\mathcal{S}$, where the diagonal elements $\lambda_i(f)$ of Λ are spectrally majorized: $\lambda_0(f) \geq \lambda_1(f) \geq \ldots \geq \lambda_{M-1}(f)$ for all f. Without loss of generality, assume that $R_0 \geq R_1 \geq \cdots \geq R_{M-1}$. Let $\mathcal{C}$ be the class of subband coders in which the analysis and synthesis filter banks are of the form $\mathcal{H} = \mathcal{G}\mathcal{U}$ and $\tilde{\mathcal{H}} = \mathcal{U}^\dagger\tilde{\mathcal{G}}$, respectively, where*

$\mathcal{U} = \mathcal{V}^\dagger$ is a PCFB, and $\mathcal{G}$ and $\tilde{\mathcal{G}}$ are diagonal matrices with real, nonnegative functions on the diagonals; see Fig. 2. Then the optimal subband coder in class $\mathcal{C}$ achieves the global minimum of the MSE (2.2).

Proof: See Appendix A.

The particular subband coder in Fig. 2 was already shown to be optimal in the PR class [11]. In that case, we had the additional restriction $\tilde{G}_i(f) = 1/G_i(f)$, due to the PR constraint. It was also shown that under optimality conditions, $\mathcal{U}(f)$ must be a PCFB. As Theorem 3.3 shows, a closely related but weaker property holds for our problem. As discussed in Remark 5 in Sec. 3.2, it is generally not necessary for $\mathcal{U}(f)$ to be a PCFB for optimality. Hence, unlike in the PR case, the design in Fig. 2 may not be necessary for optimality.

In Sec. 3.2, we seek expressions for the best filters $G_i(f)$ and $\tilde{G}_i(f)$. The optimal solution is clearly guaranteed to be at least as good as the optimal PR solution, since the optimization is over a larger set.

3.2 Optimal Filters for a Given Bit Allocation

Our goal here is to find the pre- and post-filters $G_i(f)$ and $\tilde{G}_i(f)$ in Fig. 2 that minimize the MSE (2.2) for a given $\mathcal{U}(f)$ and bit allocation $\{R_i\}$. We shall *not* assume that either $\mathcal{U}(f)$ or the bit allocation $\{R_i\}$ is optimal. However, we shall assume that $\mathcal{U}(f)$ statistically decouples the subbands, as is the case for the PCFB in Fig. 2.

Let $S_i(f)$, $0 \le i < M$, be the spectral density of the signal $z_i(n)$ in channel i at the output of $\mathcal{U}(f)$. Since $\mathcal{H} = \mathcal{G}\mathcal{U}$ and $\tilde{\mathcal{H}} = \mathcal{U}^\dagger \tilde{\mathcal{G}}$, we have

$$\sigma_i^2 = \int_{-0.5}^{0.5} \left| G_i(f)\sqrt{S_i(f)} \right|^2 df \quad \text{and} \quad ||\tilde{h}_i||^2 = \int_{-0.5}^{0.5} \left| \tilde{G}_i(f) \right|^2 df, \quad 0 \le i < M. \tag{3.1}$$

Additionally, since $\mathcal{U}(f)$ statistically decouples the subbands,

$$\begin{aligned} & Tr\left[\left(\tilde{\mathcal{H}}(f)\mathcal{H}(f) - I_M \right) \mathcal{S}(f) \left(\tilde{\mathcal{H}}(f)\mathcal{H}(f) - I_M \right)^\dagger \right] \\ = \; & Tr\left[\mathcal{U}^\dagger(f) \left(\tilde{\mathcal{G}}(f)\mathcal{G}(f) - I_M \right) \mathcal{S}_z(f) \left(\tilde{\mathcal{G}}(f)\mathcal{G}(f) - I_M \right)^\dagger \mathcal{U}(f) \right] \\ = \; & Tr\left[\left(\tilde{\mathcal{G}}(f)\mathcal{G}(f) - I_M \right) \mathcal{S}_z(f) \left(\tilde{\mathcal{G}}(f)\mathcal{G}(f) - I_M \right)^\dagger \right] \\ = \; & \sum_{i=0}^{M-1} \left| \tilde{G}_i(f) G_i(f) - 1 \right|^2 S_i(f). \end{aligned} \tag{3.2}$$

Hence the MSE (2.2), viewed as a function of $\mathbf{G} = \{G_i(f)\}$ and $\tilde{\mathbf{G}} = \{\tilde{G}_i(f)\}$, takes the form

$$\mathcal{E}\left(\mathbf{G}, \tilde{\mathbf{G}}\right) = \frac{1}{M}\sum_{i=0}^{M-1} \overline{D}(R_i) \int_{-0.5}^{0.5} \left|G_i(f)\sqrt{S_i(f)}\right|^2 df \int_{-0.5}^{0.5} \left|\tilde{G}_i(f)\right|^2 df$$
$$+\frac{1}{M}\sum_{i=0}^{M-1} \int_{-0.5}^{0.5} \left|\tilde{G}_i(f)G_i(f) - 1\right|^2 S_i(f) df. \tag{3.3}$$

Applying the Cauchy-Schwartz inequality to the product of integrals in (3.3), we obtain

$$\mathcal{E}\left(\mathbf{G}, \tilde{\mathbf{G}}\right) \geq \overline{\mathcal{E}}\left(\mathbf{P}\right), \tag{3.4}$$

where we have introduced the *product filters*

$$P_i(f) = G_i(f)\tilde{G}_i(f),\ 0 \leq i < M; \quad \mathbf{P}(f) \triangleq \left[P_0(f), P_1(f), \ldots, P_{M-1}(f)\right]^T,$$

and the functional

$$\overline{\mathcal{E}}\left(\mathbf{P}\right) \triangleq \frac{1}{M}\sum_{i=0}^{M-1} \overline{D}(R_i) \left[\int_{-0.5}^{0.5} P_i(f)\sqrt{S_i(f)}\, df\right]^2 + \frac{1}{M}\sum_{i=0}^{M-1} \int_{-0.5}^{0.5} \left|P_i(f) - 1\right|^2 S_i(f)\, df. \tag{3.5}$$

By Theorem 3.3, $G_i(f)$ and $\tilde{G}_i(f)$ are real and nonnegative, so $P_i(f)$ is also nonnegative:

$$P_i(f) \geq 0, \quad 0 \leq i < M, \ \ f \in [-0.5, 0.5]. \tag{3.6}$$

The lower bound in (3.4) is attained if and only if

$$G_i(f)\sqrt{S_i(f)} = \alpha_i \tilde{G}_i(f), \quad 0 \leq i < M, \ \ f \in [-0.5, 0.5], \tag{3.7}$$

where α_i are arbitrary, nonnegative, real-valued constants. If (3.7) is satisfied, we have

$$\sigma_i^2 ||\tilde{h}_i||^2 = \left|\int_{-0.5}^{0.5} P_i(f)\sqrt{S_i(f)}\, df\right|^2, \quad 0 \leq i < M. \tag{3.8}$$

Our approach is to minimize the lower bound (3.5) over $\mathbf{P}$ subject to the nonnegativity constraints (3.6). Using (3.7), we then construct a set of filters $\mathbf{G}$, $\tilde{\mathbf{G}}$ that attains this minimum; hence this must be the solution to the minimization problem (3.3). Denote by

$$\mathbf{P_{opt}} = \mathrm{argmin}_{\{P_i(f) \geq 0\}} \overline{\mathcal{E}}\left(\mathbf{P}\right) \tag{3.9}$$

the solution to the constrained optimization problem (3.5) (3.6). The following theorem gives $\mathbf{P_{opt}}$ and constructs optimal filters $\mathbf{G_{opt}}$ and $\tilde{\mathbf{G}}_{\mathbf{opt}}$ in terms of the bit allocation $\{R_i\}$ and the spectral densities $\{S_i(f)\}$. If $\mathcal{U}(f)$ is a PCFB, and if the bit allocation is optimal, then $\mathbf{G_{opt}}$ and $\tilde{\mathbf{G}}_{\mathbf{opt}}$ give us a globally optimal solution to the original optimization problem (2.2).

Theorem 3.4 *For any given eigenvector matrix $\mathcal{U}(f)$ and $\{R_i\}$, the filters* $\mathbf{G}$ *and* $\tilde{\mathbf{G}}$ *in Fig. 2 that minimize the MSE (2.2) are given by*

$$G_{opt,i}(f) = c_i S_i^{-\frac{1}{4}}(f)\sqrt{P_{opt,i}(f)}, \quad 0 \le i < M, \;\; f \in [-0.5, 0.5] \tag{3.10}$$

and

$$\tilde{G}_{opt,i}(f) = c_i^{-1} S_i^{\frac{1}{4}}(f)\sqrt{P_{opt,i}(f)}, \quad 0 \le i < M, \;\; f \in [-0.5, 0.5], \tag{3.11}$$

where c_i are arbitrary positive scaling factors, and the product filters $P_{opt,i}(f) = G_{opt,i}(f)\tilde{G}_{opt,i}(f)$, are of the form

$$P_{opt,i}(f) = \max\left(0, 1 - \frac{\overline{D}(R_i)}{1 + |\mathcal{F}_i^+|\,\overline{D}(R_i)} \frac{\int_+ \sqrt{S_i(f)}\,df}{\sqrt{S_i(f)}}\right), \quad 0 \le i < M, \quad f \in [-0.5, 0.5]. \tag{3.12}$$

Here $\mathcal{F}_i^+ \subset [-0.5, 0.5]$ is the set of frequencies where $P_{opt,i}(f)$ is strictly positive, $\int_+$ denotes integration over $\mathcal{F}_i^+$, and $\int_-$ denotes integration over the complementary set. The normalized variances of the subband signals are

$$\frac{1}{c_i^2}\sigma^2_{opt,i} = \int_{-0.5}^{0.5} P_{opt,i}(f)\sqrt{S_i(f)}\,df = \frac{\int_+ \sqrt{S_i(f)}\,df}{1 + |\mathcal{F}_i^+|\,\overline{D}(R_i)}, \quad 0 \le i < M. \tag{3.13}$$

The MSE for the optimal filters $\mathbf{G}$, $\tilde{\mathbf{G}}$ *is given by*

$$\mathcal{E}_{min} = \frac{1}{M}\sum_{i=0}^{M-1} \frac{\overline{D}(R_i)}{1 + |\mathcal{F}_i^+|\,\overline{D}(R_i)} \left(\int_+ \sqrt{S_i(f)}\,df\right)^2 + \frac{1}{M}\sum_{i=0}^{M-1}\int_- S_i(f)\,df. \tag{3.14}$$

Proof: See Appendix B.

Since the support sets $\mathcal{F}_i^+$ depend on $P_{opt,i}(f)$, the expression (3.12) is apparently a complicated nonlinear system of equations. However, in Appendix C we develop a simple algorithm that solves this system. The key idea is that $P_{opt,i}(f)$ is zero for all frequencies f such that $S_i(f)$ is below a certain threshold. The weak signal components at frequencies $f \notin \mathcal{F}_i^+$ are of no significance in determining the optimal filters. Note that this property is unique to filter banks without PR constraints. The ability to entirely eliminate weak components of the signal bears a striking similarity with the well-known reverse water filling result in rate-distortion theory for stationary Gaussian processes [16, p. 641], according to which no bits should be allocated to frequency components that fall below a certain level that depends on the spectral density of the signal and on the bit budget available.

To illustrate Theorem 3.4 and the discussion above, we present a toy example. The signal $x(n)$ in Fig. 3 is applied to a two-channel filter bank. Its spectral density $S_x(f)$ is piecewise constant as given in Fig. 3(a), with $\delta \ll 1$. The frequency response of the lowpass and highpass filters of the PCFB $\mathcal{U}(f)$ is displayed in Fig. 3(b), and the resulting aliased spectra $S_0(f)$ and $S_1(f)$ are shown in Fig. 3(c). For any bit rate R_0, the solution to (3.12) is given by $\mathcal{F}_0^+ = [-0.5, 0.5]$, $P_{opt,0}(f) = 1 - \frac{\overline{D}(R_0)}{1+\overline{D}(R_0)}$. For any given R_1, if δ is small enough, half of the frequency components of $S_1(f)$ are weak and are filtered out. For $\delta \ll 1$, the solution to (3.12) is $\mathcal{F}_1^+ = [-0.5, -0.25] \cup [0.25, 0.5]$, and $P_{opt,1}(f) = 1 - \frac{\overline{D}(R_1)}{2+\overline{D}(R_1)}$ for $f \in \mathcal{F}_1^+$, and zero otherwise.

Remark 1 The *functional form* of the optimal filters (3.12) is independent of the particular rate-distortion criterion used.

Remark 2 Given a particular bit allocation $\{R_i\}$, the solution (3.12) is unique because it is the global minimizer of the strictly convex cost functional (3.5) over the convex set (3.6) (a simple extension of Proposition B.10, in [17, p. 571]). This implies uniqueness of the sets $\{\mathcal{F}_i^+\}$.

Remark 3 In virtually all examples involving AR processes we studied, a significant simplification arose as $P_{opt,i}(f)$ was strictly positive over the whole frequency interval:

$$\mathcal{F}_i^+ = [-0.5, 0.5], \quad 0 \le i < M. \tag{3.15}$$

From (3.12), we find that a necessary and sufficient condition for (3.15) to hold is

$$\min_{f\in[-0.5,0.5]} \sqrt{S_i(f)} > \frac{\overline{D}(R_i)}{1+\overline{D}(R_i)} \int_{-0.5}^{0.5} \sqrt{S_i(f)}\, df, \quad 0 \le i < M. \tag{3.16}$$

In other words, (3.15) holds if $\sqrt{S_i(f)}$ is not too small relative to its mean and if the bit rate R_i is large enough. This condition was not satisfied in the toy example of Fig. 3. If (3.15) holds, then (3.12) reduces to

$$P_{opt,i}(f) = 1 - \frac{\overline{D}(R_i)}{1+\overline{D}(R_i)} \frac{\int_{-0.5}^{0.5} \sqrt{S_i(f)}\, df}{\sqrt{S_i(f)}}, \quad 0 \le i < M, \quad f \in [-0.5, 0.5], \tag{3.17}$$

and the expression (3.14) for the minimal MSE simplifies to

$$\mathcal{E}_{min} = \frac{1}{M} \sum_{i=0}^{M-1} \frac{\overline{D}(R_i)}{1+\overline{D}(R_i)} \left[\int_{-0.5}^{0.5} \sqrt{S_i(f)}\, df\right]^2. \tag{3.18}$$

Remark 4 If $|\mathcal{F}_i^+| \neq 1$, it may not be necessary for $\mathcal{U}(f)$ to be a PCFB. To see this, consider the two-channel case, and assume there exists f such that the signal components at f are weak in both channels: $f \notin \mathcal{F}_0^+$ and $f \notin \mathcal{F}_1^+$. Then, switching $\mathcal{U}_{00}f)$ and $\mathcal{U}_{11}(f)$ has no effect on the cost functional.

Remark 5 Since the class $\mathcal{C}$ of filters considered includes biorthogonal filters as a special case, the optimal MSE (3.14) is *upper-bounded* by the minimal value

$$\mathcal{E} = \frac{1}{M} \sum_{i=0}^{M-1} \overline{D}(R_i) \left(\int_{-0.5}^{0.5} \sqrt{S_i(f)}\, df\right)^2$$

for the Infinite Impulse Response (IIR) biorthogonal case [11]. At high bit rates, $\mathcal{E}_{min}$ tends to the limit above. Observe that $G_{opt,i}(f)$ and $\tilde{G}_{opt,i}(f)$ can be approximated with arbitrary accuracy by FIR filters of sufficient length, so the performance of optimal FIR filter banks converges to (3.14) as the filter length tends to infinity. As discussed in [11], PR-FIR filter banks do not enjoy a similar property: they must satisfy the constraint $\det \mathcal{H}(f) \equiv 1$, and their performance does not converge to that of IIR biorthogonal filter banks (for which the constraint $\det \mathcal{H}(f) \equiv 1$ is not applicable). This has a remarkable implication, which is illustrated by experiments in Sec. 5: even at high bit rates, FIR filter banks of specified length can vastly outperform FIR biorthogonal filter banks of arbitrary length.

3.3 Computation of Jointly Optimal Filters and Bit Allocation

In Sec. 2.3, we derived the condition (2.4) for the optimal bit allocation given the filters. In Sec. 3.2, we obtained the optimal filters from the expression (3.12) for the optimal product filters $\mathbf{P}$, given the bit allocation $\{R_i\}$ and the eigenvector matrix $\mathcal{U}(f)$. This was done by minimizing the cost functional $\overline{\mathcal{E}}$ in (3.5) over $\mathbf{P}$. As mentioned in Remark 2, these optimal product filters are unique, given $\{R_i\}$ and $\mathcal{U}(f)$, because $\overline{\mathcal{E}}$ is strictly convex in $\mathbf{P}$. Moreover, $\overline{\mathcal{E}}$ is strictly convex in $\{R_i\}$ for every $\mathbf{P}$, which implies the uniqueness of the optimal $\{R_i\}$, given $\mathbf{P}$ and $\mathcal{U}(f)$.

We are now interested in computing jointly optimal filters and bit allocation. This can be done by using an alternating optimization approach in which bit allocation and filters are successively updated. Each iteration involves two steps: first, optimize the bit rates using standard convex programming (e.g., gradient-based) techniques, and second, optimize the filters using (3.10), (3.11), and (3.12). Such an algorithm would converge to the globally optimum solution under convexity conditions that are discussed below. The PCFB $\mathcal{U}(f)$ is fixed, so we view $\overline{\mathcal{E}}$ as a function of both $\{R_i\}$ and $\mathbf{P}$. The joint optimization problem is much harder than it appears at first sight. For instance, just because the optimal $\{R_i\}$ is unique given $\mathbf{P}$, and the optimal $\mathbf{P}$ is unique given $\{R_i\}$, we cannot assume that the jointly optimal solution is unique. [1] We now present a condition on the rate-distortion function $\overline{D}(R)$ that guarantees strict convexity of $\overline{\mathcal{E}}$ as a function of the pair $(\{R_i\}, \mathbf{P})$ and hence existence of a unique minimum. The convexity property implies that this minimum can be easily found by using numerical techniques such as the iterative algorithm above.

[1] A textbook example of this apparent paradox is the function of two variables $F(x_0, x_1) = (x_0 - 1)^2 x_1^2 + x_0^2 (x_1 - 1)^2$, which has a unique minimum for any given x_0 and a unique minimum for any given x_1, yet admits two joint minima $(0,0)$ and $(1,1)$. This function is strictly convex in x_0 as well as strictly convex in x_1, but is not convex in the pair (x_0, x_1).

Proposition 3.5 *The cost functional (3.5) is strictly convex in $\{R_i\}$ and* $\mathbf{P}$ *if*

$$2\left(\overline{D}'(R)\right)^2 < \overline{D}''(R)\left(\overline{D}(R)+1\right), \tag{3.19}$$

where $\overline{D}'(R)$ and $\overline{D}''(R)$ denote the first and second derivatives of $\overline{D}(R)$, respectively.

Proof: See Appendix D.

Note that our concavity requirement on $\ln \overline{D}(R)$ in Sec. 2.1 can be written as $\left[\overline{D}(R)\right]^2 \geq \overline{D}(R)\overline{D}''(R)$, so not many distortion rate-functions will satisfy both the log concavity assumption and the condition (3.19) for all R. Interestingly, under the classical exponential model $\overline{D}(R) = \gamma 2^{-2R}$ for the rate-distortion function, (3.19) reduces to the simple condition $\overline{D}(R) < 1$, namely,

$$R > \frac{\log_2 \gamma}{2}. \tag{3.20}$$

This condition is satisfied under our initial assumption of high bit rates. For entropy constrained scalar quantizers (ECSQ) with Gaussian inputs, $\frac{\log_2 \gamma}{2} = 0.71$ [16, p. 154].

4 Exponential Rate-Distortion Function

In this section, we derive analytical expressions for the jointly optimal filters and bit allocation, under the classical model $\overline{D}(R) = \gamma 2^{-2R}$ for the rate-distortion function. Recall from (3.20) that the high bit rate assumption guarantees uniqueness of the solution to this optimization problem. For this particular $\overline{D}(R)$, the optimal bit allocation condition (2.4) becomes

$$\sigma_i^2 ||\tilde{h}_i||^2 \left.\frac{d\overline{D}(R)}{dR}\right|_{R_i} = \gamma(-2\ln 2)\sigma_i^2||\tilde{h}_i||^2 2^{-2R_i} = \mu, \quad 0 \leq i < M. \tag{4.1}$$

The resulting distortions $\sigma_i^2||\tilde{h}_i||^2\overline{D}(R_i)$ are thus identical for all channels. Therefore,

$$2^{-2R_i}\sigma_i^2||\tilde{h}_i||^2 = 2^{-2R}\left[\prod_{j=0}^{M-1}\sigma_j^2||\tilde{h}_j||^2\right]^{1/M}, \quad 0 \leq i < M, \tag{4.2}$$

where $R = \frac{1}{M}\sum_{j=0}^{M-1} R_j$. This yields a closed–form expression for the optimal bit allocation in terms of the filters,

$$R_i = R + \log_2 \sigma_i||\tilde{h}_i|| - \frac{1}{M}\sum_{j=0}^{M-1}\log_2 \sigma_j||\tilde{h}_j||, \quad 0 \leq i < M. \tag{4.3}$$

Substituting (4.2) into (2.2), we obtain

$$\begin{aligned}\mathcal{E} \;=\; & \gamma 2^{-2R}\left[\prod_{i=0}^{M-1}\int_{-0.5}^{0.5}\left(\mathcal{H}(f)\mathcal{S}(f)\mathcal{H}^\dagger(f)\right)_{ii} df \int_{-0.5}^{0.5}\left(\tilde{\mathcal{H}}^\dagger(f)\tilde{\mathcal{H}}(f)\right)_{ii} df\right]^{1/M} \\ & + \frac{1}{M}Tr\int_{-0.5}^{0.5}\left(\tilde{\mathcal{H}}(f)\mathcal{H}(f) - I_M\right) S(f)\left(\tilde{\mathcal{H}}(f)\mathcal{H}(f) - I_M\right)^\dagger df,\end{aligned} \tag{4.4}$$

to be minimized over $(\mathcal{H}, \tilde{\mathcal{H}})$. (For PR systems, the optimization criterion is the classical coding gain criterion.) It remains to solve the resulting nonlinear system (3.12) (4.3). The solution of this system is given in Appendix E. Equation (E.1) gives the optimal filters under optimal bit allocation. This expression involves $M+1$ constants $\sigma^2_{opt,i}$ and $\mathcal{E}_q$, which are solutions to the nonlinear system (E.2) (E.3). These solutions are unique owing to (3.20). We have been able to solve the nonlinear system (E.2) (E.3) in closed form for the special case of two-channel filter banks at high bit rates. The result is stated below.

Proposition 4.1 *Assume without loss of generality that $c_i \equiv 1$.* [2] *For two-channel coders with exponential distortion–rate function, there exists $R^* > 0$ such that the following holds. For any $R > R^*$, the optimal pre- and post-filters are given by $G_{opt,i}(f) = S_i^{-\frac{1}{4}}(f)\sqrt{P_{opt,i}(f)}$ and $\tilde{G}_{opt,i}(f) = S_i^{\frac{1}{4}}(f)\sqrt{P_{opt,i}(f)}$, $i = 0, 1$, where*

$$P_{opt,0}(f) = 1 - \frac{\gamma 2^{-2R}}{\sqrt{S_0(f)}}\sigma^2_{opt,1}, \quad f \in [-0.5, 0.5],$$
$$P_{opt,1}(f) = 1 - \frac{\gamma 2^{-2R}}{\sqrt{S_1(f)}}\sigma^2_{opt,0}, \quad f \in [-0.5, 0.5], \tag{4.5}$$

and

$$\sigma^2_{opt,0} = \frac{\int_{-0.5}^{0.5}\sqrt{S_0(f)}\,df - \gamma 2^{-2R}\int_{-0.5}^{0.5}\sqrt{S_1(f)}}{1-(\gamma 2^{-2R})^2},$$
$$\sigma^2_{opt,1} = \frac{\int_{-0.5}^{0.5}\sqrt{S_1(f)}\,df - \gamma 2^{-2R}\int_{-0.5}^{0.5}\sqrt{S_0(f)}}{1-(\gamma 2^{-2R})^2}. \tag{4.6}$$

Proof: Here (E.3) takes the form

$$\mathcal{E}_q = \gamma 2^{-2R}\sigma^2_{opt,0}\sigma^2_{opt,1}. \tag{4.7}$$

Assume momentarily that condition (3.15) is satisfied. Substituting (4.7) into (E.9), we obtain

$$\sigma^4_{opt,i} - \sigma^2_{opt,i}\int_+ \sqrt{S_i(f)}\,df + \gamma 2^{-2R}\sigma^2_{opt,0}\sigma^2_{opt,1} = 0, \quad i = 0, 1.$$

Dividing both sides by $\sigma^2_{opt,0}$ when $i = 0$ and by $\sigma^2_{opt,1}$ when $i = 1$ gives the linear system

$$\sigma^2_{opt,0} + \gamma 2^{-2R}\sigma^2_{opt,1} = \int_{-0.5}^{0.5}\sqrt{S_0(f)}\,df,$$
$$\sigma^2_{opt,1} + \gamma 2^{-2R}\sigma^2_{opt,0} = \int_{-0.5}^{0.5}\sqrt{S_1(f)}\,df,$$

whose solution is (4.6). Substituting (4.7) into (E.1), we obtain (4.5). From (4.6), we have $\sigma^2_{opt,i} \to \int_{-0.5}^{0.5}\sqrt{S_i(f)}\,df$ as $R \to \infty$ $(i = 1, 2)$. Hence (4.3) implies that $R_i \to \infty$. Hence the right–hand side of (3.16) tends to zero, which justifies our initial assumption (3.15). ■

[2] Otherwise replace $\sigma^2_{opt,i}$ by $\sigma^2_{opt,i}/c_i^2$ in the expressions below.

5 Numerical Experiments

To illustrate our analyses, we applied an AR(1) input process with correlation coefficient $r = 0.8$ to a two-channel subband coder. The PCFB for this process is the traditional orthonormal filter bank with ideal low-pass and high-pass filters shown in Fig. 3b. We used the exponential rate-distortion model $\overline{D}(R) = \gamma 2^{-2R}$ for the quantizers, with $\gamma = 2.71$ (ECSQ). The optimization problem was solved for various rates R, and the results were compared with optimal IIR biorthogonal filter banks and optimal unconstrained-length FIR biorthogonal filter banks [11]. At all bit rates, we obtained $\mathcal{F}_i^+ = \left[-\frac{1}{2}, \frac{1}{2}\right]$. For $R = 1.76$, the bit rate in the high-pass channel becomes zero, in which case the design criterion (4.4) becomes invalid. As Fig. 4 shows, the performance of the optimal unconstrained-length MMSE filter banks is very close to optimal IIR biorthogonal filter banks at very high rates, but improvements become quite significant as R decreases. These results are consistent with numerous results in the literature using numerically optimized filter banks. Frequency responses are shown in Fig. 5 for an AR(1) process with correlation coefficient $r = 0.8$, and rate $R = 2.91$. The arbitrary scaling constants c_0 and c_1 for all three filter banks have been chosen so that the frequency responses are the same at $f = 0$ and at $f = 0.5$.

Remark 5 at the end of Sec. 3.2 motivated us to investigate the constrained-length version of this design and to quantify the improvements over FIR biorthogonal filter banks. A simple rectangular windowing technique was used to design constrained-length FIR filter banks from the optimum unconstrained-length solution. As shown in Fig. 6, the results are excellent at medium bit rates. At $R = 2.91$, the length-63 FIR filter bank outperforms optimal FIR biorthogonal filter banks of arbitrary length, and the length-103 FIR filter bank outperforms optimal IIR biorthogonal filter banks. Similar advantages hold at arbitrarily high bit rates, but longer FIR filters are needed to break the performance bounds for FIR and IIR biorthogonal filters. Refinements in the FIR filter design method are likely to yield further improvements.

6 Conclusion

We have studied the design of subband coders that are optimally adapted to second–order input signal statistics. Our main working assumptions are the use of scalar quantizers and a model for the rate–distortion characteristic of these quantizers. Absolutely no constraints are placed on the subband filters. The criterion for optimal design of the filters and bit allocation is overall rate-distortion performance of the subband coder. Previous results in the literature have addressed simplified versions of this problem, involving assumptions such as fixed analysis filters. However, joint optimization of analysis filters, synthesis filters, and bit allocation presents significant challenges. We have obtained answers to a number of basic theoretical questions.

First, we have shown that the output of the optimal coder satisfies fundamental properties of total decorrelation and spectral majorization, which were previously known to apply only to coders with PR constraints on the filters. Also, we have shown that the optimal subband coder in the special class of Fig. 2 achieves the global minimum of the distortion-rate function. This analysis stresses the important role played by PCFBs in such problems. We have derived analytical expressions for the optimal pre- and post-filters in Fig. 2. The functional form of these expressions is independent of the particular rate-distortion model used. Conditions for uniqueness of the jointly optimal pre- and post-filters and bit allocation have been identified in Proposition 3.5. A simple algorithm based on convexity properties of the cost function has been proposed to compute the globally optimal coder. Closed-form expressions have been obtained in the case of two-channel subband coders with exponential rate-distortion function.

The optimal filters do not admit a finite parameterization. In practice, FIR approximations may be sought. We have constructed simple FIR approximations and demonstrated both theoretically and numerically (see Fig. 6) the following remarkable property: these FIR approximations significantly outperform FIR PR filter bank of arbitrary length in the sense that even at high bit rates, the MSE performance gap does not tend to zero. Note that the ultimate performance in a constrained class of FIR filters would be obtained by using rate-distortion as the approximation criterion. However, determining optimal filters in this fashion is still an open problem. Another possible extension of our work would be to relax the high-rate assumption that led to the MSE model (2.2). However, as indicated in [9], low-bit-rate quantization noise models involve complex dependencies between noise samples. Optimization of the overall performance of the subband coder under such models would certainly present a formidable challenge.

APPENDIX

A Proof of Theorem 3.3

First we give two lemmas that are a direct consequence of Proposition 3.1.

Lemma A.1 *The optimal synthesis polyphase matrix is of the form*

$$\tilde{\mathcal{H}}(f) = \tilde{\mathcal{U}}(f)\tilde{\mathcal{G}}(f), \tag{A.1}$$

where $\tilde{\mathcal{U}}(f)$ is a paraunitary matrix and $\tilde{\mathcal{G}}(f)$ is a diagonal matrix with real, nonnegative entries.

The proof of Lemma A.1 is exactly the same as the proof of Corollary 2.4 in [11] and uses the fact that $\tilde{\mathcal{H}}^\dagger\tilde{\mathcal{H}}$ is diagonal, which is a consequence of Proposition 3.1. ■

Lemma A.2 *The optimal analysis polyphase matrix is of the form*

$$\mathcal{H}(f) = \mathcal{S}_y^{0.5}(f)\mathcal{W}(f)\Lambda^{-0.5}(f)\mathcal{V}^\dagger(f), \tag{A.2}$$

where $\mathcal{S} = \mathcal{V}\Lambda\mathcal{V}^\dagger$ is the eigenvector decomposition of $\mathcal{S}$, $\mathcal{W}$ is a paraunitary matrix, and $\mathcal{S}_y$ is the (diagonal) spectral density matrix for the vector $\underline{y}(n)$ of subband signals in Fig. 1.

Proof : Define the $M \times M$ matrix $\mathcal{W} = \mathcal{S}_y^{-0.5}\mathcal{H}\mathcal{V}\Lambda^{0.5}$. Hence $\mathcal{H}(f)$ takes the form (A.2). By Proposition 3.1, $\mathcal{S}_y = \mathcal{H}\mathcal{S}\mathcal{H}^\dagger$ is a diagonal matrix with nonnegative entries. This implies that $\mathcal{W}\mathcal{W}^\dagger = I$, and thus $\mathcal{W}$ is paraunitary. ■

To prove Theorem 3.3, we again use a variational technique. The outline of the proof is as follows. We fix the bit allocation $\{R_i\}$ and assume that $(\mathcal{H}, \tilde{\mathcal{H}})$ minimizes the MSE (2.2). Let $S_y = \mathcal{H}\mathcal{S}\mathcal{H}^\dagger$ be the spectral density matrix corresponding to this choice of $\mathcal{H}$. The variational technique is based on the parameterizations $\tilde{\mathcal{H}} = \tilde{\mathcal{U}}\tilde{\mathcal{G}}$ and $\mathcal{H} = \mathcal{S}_y^{0.5}\mathcal{W}\Lambda^{-0.5}\mathcal{V}^\dagger$ given by Lemmas A.1 and A.2. We construct a filter bank $(\mathcal{H}_{new}, \tilde{\mathcal{H}}_{new})$ that is in class $\mathcal{C}$ and satisfies $\mathcal{E}(\mathcal{H}_{new}, \tilde{\mathcal{H}}_{new}) = \mathcal{E}(\mathcal{H}, \tilde{\mathcal{H}})$. We then conclude that the optimal subband coder in class $\mathcal{C}$ achieves the global minimum of (2.2) over the unconstrained set of subband coders.

The proof is as follows. First consider the quantization noise term $\mathcal{E}_n(\mathcal{H}, \tilde{\mathcal{H}}) = \frac{1}{M}\sum_{i=0}^{M-1} \overline{D}(R_i)\sigma_i^2||\tilde{h}_i||^2$ in (2.2). From (2.3) and (A.1), we have

$$\sigma_i^2 = \int \mathcal{S}_{y,ii}(f)\, df \quad \text{and} \quad ||\tilde{h}_i||^2 = \int \tilde{\mathcal{G}}_{ii}^2(f)\, df, \quad i = 0, 1, \ldots, M-1. \tag{A.3}$$

Hence, for fixed $\{R_i\}$, $\mathcal{S}_y$ and $\tilde{\mathcal{G}}$, the noise term is unaffected by variations in $\tilde{\mathcal{U}}$ and $\mathcal{W}$.

The signal term (due to lack of PR) in (2.2) is given by

$$\begin{aligned}
\mathcal{E}_s(\mathcal{H},\tilde{\mathcal{H}}) &= \frac{1}{M} Tr \int \left(\tilde{\mathcal{H}}\mathcal{H} - I_M\right) \mathcal{S} \left(\tilde{\mathcal{H}}\mathcal{H} - I_M\right)^\dagger \\
&= \frac{1}{M} \int Tr \left(\tilde{\mathcal{H}}\mathcal{H} - I_M\right) \mathcal{S} \left(\tilde{\mathcal{H}}\mathcal{H} - I_M\right)^\dagger \\
&= \frac{1}{M} \int Tr \left(\tilde{\mathcal{U}}\tilde{\mathcal{G}}\mathcal{S}_y^{0.5}\mathcal{W}\Lambda^{-0.5}\mathcal{V}^\dagger - I_M\right) \mathcal{V}\Lambda^{0.5}\Lambda^{0.5}\mathcal{V}^\dagger \left(\tilde{\mathcal{U}}\tilde{\mathcal{G}}\mathcal{S}_y^{0.5}\mathcal{W}\Lambda^{-0.5}\mathcal{V}^\dagger - I_M\right)^\dagger \\
&= \frac{1}{M} \int Tr \left[\mathcal{V}\left(\mathcal{V}^\dagger\tilde{\mathcal{U}}\tilde{\mathcal{G}}\mathcal{S}_y^{0.5}\mathcal{W} - \Lambda^{0.5}\right)\left(\mathcal{V}^\dagger\tilde{\mathcal{U}}\tilde{\mathcal{G}}\mathcal{S}_y^{0.5}\mathcal{W} - \Lambda^{0.5}\right)^\dagger \mathcal{V}^\dagger\right] \\
&= \frac{1}{M} \int Tr \left[\left(\mathcal{V}^\dagger\tilde{\mathcal{U}}\tilde{\mathcal{G}}\mathcal{S}_y^{0.5}\mathcal{W} - \Lambda^{0.5}\right)\left(\mathcal{V}^\dagger\tilde{\mathcal{U}}\tilde{\mathcal{G}}\mathcal{S}_y^{0.5}\mathcal{W} - \Lambda^{0.5}\right)^\dagger\right], \qquad \text{(A.4)}
\end{aligned}$$

where the third equality uses the eigenvector decomposition for $\mathcal{S}$ and the parameterization of $\mathcal{H}$ and $\tilde{\mathcal{H}}$ given in Lemmas A.1 and A.2, and the last equality uses the trace invariance property, which holds for any unitary $\mathcal{V}$. Define the matrix

$$\mathcal{Q} = \tilde{\mathcal{G}}\mathcal{S}_y^{0.5}, \qquad \text{(A.5)}$$

which by our assumptions is diagonal, real, and nonnegative definite. Observe that its squared diagonal elements $Q_i^2(f)$ are the spectral density functions of the signal part of $\underline{v}(n)$ in Fig. 2. Both $\mathcal{Q}$ and $\Lambda^{0.5}$ are diagonal matrices with nonnegative real elements, and $\mathcal{V}^\dagger\tilde{\mathcal{U}}$ and $\mathcal{W}$ are complex paraunitary matrices. We now use a simple extension of the Hoffman-Wielandt theorem [18, Theorem 6.3.5] to the non-Hermitian case [18, Corollary 7.3.8] to bound the integrand of (A.4) from below:

$$Tr\left[\left(\mathcal{V}^\dagger\tilde{\mathcal{U}}\mathcal{Q}\mathcal{W} - \Lambda^{0.5}\right)\left(\mathcal{V}^\dagger\tilde{\mathcal{U}}\mathcal{Q}\mathcal{W} - \Lambda^{0.5}\right)^\dagger\right] \geq \sum_{i=0}^{M-1} \left(\overline{Q}_i - \lambda_i^{0.5}\right)^2, \qquad \text{(A.6)}$$

where $\{\overline{Q}_i\}$ is a frequency-dependent permutation of $\{Q_i\}$, so that $\overline{Q}_0(f) \geq \overline{Q}_1(f) \geq \cdots \geq \overline{Q}_{M-1}(f)$. Since $(\mathcal{H},\tilde{\mathcal{H}})$ is optimal, no variation in $\tilde{\mathcal{U}}$ and $\mathcal{W}$ can reduce $\mathcal{E}_s(\mathcal{H},\tilde{\mathcal{H}})$; hence $\mathcal{E}_s(\mathcal{H},\tilde{\mathcal{H}})$ must achieve the lower bound in the right-hand side of (A.6).

If $Q_i(f)$ are already ordered, then $\overline{Q}_i(f) \equiv Q_i(f)$, and equality is achieved in (A.6) for $\mathcal{V}^\dagger\tilde{\mathcal{U}} = \mathcal{W} = I_M$. We then choose $\mathcal{H}_{new} = \mathcal{S}_y^{0.5}\Lambda^{-0.5}\mathcal{V}^\dagger$ and $\tilde{\mathcal{H}}_{new} = \mathcal{V}\tilde{\mathcal{G}}$, which are clearly in class $\mathcal{C}$, and satisfy $\mathcal{E}_s(\mathcal{H}_{new},\tilde{\mathcal{H}}_{new}) = \mathcal{E}_s(\mathcal{H},\tilde{\mathcal{H}})$; hence $\mathcal{E}(\mathcal{H}_{new},\tilde{\mathcal{H}}_{new}) = \mathcal{E}(\mathcal{H},\tilde{\mathcal{H}})$.

If $Q_i(f)$ are not ordered, we proceed in two steps. The first step consists of normalizing the analysis and synthesis filters. We let $d_i = [\overline{D}(R_i)\sigma_i^2]^{1/2}$, $\mathcal{D} = \text{diag}(d_i)$, and normalize the filters as follows: $h_{temp,i}(n) = d_i^{-1}h_i(n)$, and $\tilde{h}_{temp,i}(n) = d_i\tilde{h}_i(n)$. In terms of polyphase matrices, we have $\mathcal{H}_{temp} = \mathcal{D}^{-1}\mathcal{H}$, and $\tilde{\mathcal{H}}_{temp} = \tilde{\mathcal{H}}\mathcal{D}$. We also let $S_{y,new} = \mathcal{H}_{temp}\mathcal{S}\mathcal{H}_{temp}^\dagger$ and $\tilde{\mathcal{G}}_{temp} = \tilde{\mathcal{G}}\mathcal{D}$, so $\tilde{\mathcal{H}}_{temp} = \tilde{\mathcal{U}}\tilde{\mathcal{G}}_{temp}$. This scaling changes neither the signal term $\mathcal{E}_s(\mathcal{H}_{temp},\tilde{\mathcal{H}}_{temp}) =$

$\mathcal{E}_s(\mathcal{H},\tilde{\mathcal{H}})$ nor the noise term, which can now be written as $\mathcal{E}_n(\mathcal{H}_{temp},\tilde{\mathcal{H}}_{temp}) = \mathcal{E}_n(\mathcal{H},\tilde{\mathcal{H}}) = \frac{1}{M}\sum_{i=0}^{M-1}\int \tilde{\mathcal{G}}^2_{temp,ii}(f)\,df$. Our second step consists of reordering the diagonal elements of $\tilde{\mathcal{G}}_{temp}(f)$. For each f, let $\tilde{G}_{new,i}(f)$ be a reordering of $\tilde{G}_{temp,i}(f)$, so that $\tilde{G}_{new,0}(f) \geq \tilde{G}_{new,1}(f) \geq \cdots \geq \tilde{G}_{new,M-1}(f)$. Note that $\overline{D}(R_i)\sigma^2_{new,i} \equiv 1$, so from Proposition 3.2, we also have $S_{y,new,0}(f) \geq S_{y,new,1}(f) \geq \cdots \geq S_{y,new,M-1}(f)$. Hence from (A.5), $Q_{new,i}(f) = S^{0.5}_{y,new,i}(f)\tilde{G}_{new,i}(f)$ are similarly ordered. We now let $\mathcal{H}_{new} = S^{0.5}_{y,new}\Lambda^{-0.5}\mathcal{V}^\dagger$ and $\tilde{\mathcal{H}}_{new} = \mathcal{V}\tilde{\mathcal{G}}_{new}$. The reordering of elements of $\tilde{\mathcal{G}}_{temp}$ does not affect the noise term:

$$\mathcal{E}_n(\mathcal{H}_{new},\tilde{\mathcal{H}}_{new}) = \frac{1}{M}\sum_{i=0}^{M-1}\int \tilde{\mathcal{G}}^2_{new,ii}(f)\,df = \frac{1}{M}\sum_{i=0}^{M-1}\int \tilde{\mathcal{G}}^2_{temp,ii}(f)\,df = \mathcal{E}_n(\mathcal{H},\tilde{\mathcal{H}}).$$

As well, (A.6) yields $\mathcal{E}_s(\mathcal{H}_{new},\tilde{\mathcal{H}}_{new}) = \mathcal{E}_s(\mathcal{H},\tilde{\mathcal{H}})$, where the equality follows from the ordering of $Q_{new,i}(f)$. So again, we obtain $(\mathcal{H}_{new},\tilde{\mathcal{H}}_{new})$ in class $\mathcal{C}$ that is as good as the original $(\mathcal{H},\tilde{\mathcal{H}})$. $\blacksquare$.

B Proof of Theorem 3.4

To find the solution to the constrained optimization problem (3.6) (3.9), we use the *Generalized Kuhn-Tucker Theorem* [19, p. 249]. We introduce the Lagrangian

$$\mathcal{L}(\mathbf{P},\boldsymbol{\lambda}) = \overline{\mathcal{E}}(\mathbf{P}) - \sum_{i=0}^{M-1}\int_{-0.5}^{0.5}\lambda_i(f)P_i(f)\,df,$$

where $\lambda_i(f) \geq 0, \quad 0 \leq i < M, \quad f \in [-0.5, 0.5]$ are Lagrange multipliers, and we define $\boldsymbol{\lambda}(f) \triangleq [\lambda_0(f), \lambda_1(f), \ldots, \lambda_{M-1}(f)]^T$. Let $\mathbf{P_{opt}}(f)$ be a solution to the constrained optimization problem (3.9). Then there exists $\boldsymbol{\lambda_{opt}}(f)$ such that the Lagrangian is stationary at $\mathbf{P_{opt}}(f)$. Setting the Fréchet derivative of the Lagrangian to zero, we obtain the necessary conditions (B.1) and (B.2) for optimality:

$$\left[\left(\nabla_{P_i}\overline{\mathcal{E}}\right)(f) - \lambda_{opt,i}(f)\right]\Big|_{P_i=P_{opt,i}} = 0, \quad 0 \leq i < M, \quad f \in [-0.5, 0.5], \tag{B.1}$$

where $\boldsymbol{\lambda_{opt}}(f)$ is such that the nonnegativity constraints (3.6) are satisfied. Additionally,

$$\lambda_{opt,i}(f)P_{opt,i}(f) = 0, \quad 0 \leq i < M, \; f \in [-0.5, 0.5]. \tag{B.2}$$

To find the solution to the optimization problem (3.6) (3.9), we examine whether the constraint $P_i(f) \geq 0$ is active or not, for each frequency $f \in [-0.5, 0.5]$. If the constraint is inactive for a particular frequency f, then $P_{opt,i}(f) > 0$, and successive application of (B.2) and (B.1) gives $\lambda_{opt,i}(f) = 0$ and $\nabla_{P_i}\overline{\mathcal{E}}\big|_{P_i=P_{opt,i}} = 0$. On the other hand, if the constraint is active at f, we have $P_{opt,i}(f)=0$ and $\nabla_{P_i}\overline{\mathcal{E}}\big|_{P_i=P_{opt,i}}(f)=\lambda_{opt,i}(f)$. This solution clearly satisfies the first-order Kuhn-Tucker conditions (B.1) and (B.2).

We can now compute $P_{opt,i}(f)$. When the constraint is inactive ($P_{opt,i}(f)>0$), we obtain from (3.5)

$$0=\left(\nabla_{P_i}\overline{\mathcal{E}}\right)\big|_{P_i=P_{opt,i}}(f)=\frac{2}{M}\overline{D}(R_i)\left[\int_{-0.5}^{0.5}P_{opt,i}(f)\sqrt{S_i(f)}\,df\right]\sqrt{S_i(f)}+\frac{2}{M}\left[P_{opt,i}(f)-1\right]S_i(f),$$
$$0\le i<M,\quad f\in\mathcal{F}_i^+,$$

whence

$$P_{opt,i}(f)=1-\frac{\overline{D}(R_i)}{\sqrt{S_i(f)}}\int_{-0.5}^{0.5}P_{opt,i}(f)\sqrt{S_i(f)}\,df,\quad 0\le i<M,\quad f\in\mathcal{F}_i^+. \tag{B.3}$$

When the constraint is active $\left(f\notin\mathcal{F}_i^+\right)$, we obtain

$$P_{opt,i}(f)=\max\left(0,1-\frac{\overline{D}(R_i)}{\sqrt{S_i(f)}}\int_{-0.5}^{0.5}P_{opt,i}(f)\sqrt{S_i(f)}\,df\right),\quad 0\le i<M,\quad f\in[-0.5,0.5]. \tag{B.4}$$

Multiplying both sides by $\sqrt{S_i(f)}$, integrating and simplifying, we obtain

$$\int_{-0.5}^{0.5}P_{opt,i}(f)\sqrt{S_i(f)}\,df=\frac{\int_+\sqrt{S_i(f)}\,df}{1+\left|\mathcal{F}_i^+\right|\overline{D}(R_i)},\quad 0\le i<M. \tag{B.5}$$

Substituting (B.5) into (B.4) yields the expression (3.12) for the optimal product filter.

We can now compute the optimal pre- and post-filters, $\{G_{opt,i}(f)\}$ and $\left\{\tilde{G}_{opt,i}(f)\right\}$. From (3.7), we have $G_{opt,i}(f)\sqrt{S_i(f)}=\alpha_i\tilde{G}_{opt,i}(f)=\alpha_i\frac{P_{opt,i}(f)}{G_{opt,i}(f)}$. Hence, (3.10) (3.11) follow with $c_i=\alpha_i$. We now substitute (3.10) and (3.11) into (3.1) and obtain

$$\frac{\sigma^2_{opt,i}}{c_i^2}=c_i^2||\tilde{h}_{opt,i}||^2=\int_{-0.5}^{0.5}P_{opt,i}(f)\sqrt{S_i(f)}\,df,\quad 0\le i<M. \tag{B.6}$$

Using (B.5) in (B.6) yields (3.13). Having derived the optimum filter responses (3.10)–(3.12), we compute the corresponding optimal value of the cost functional $\mathcal{E}_{min}$. Evaluating (3.5) at $\mathbf{P_{opt}}$, we obtain

$$\begin{aligned}\mathcal{E}_{min} &\triangleq \overline{\mathcal{E}}\left(\mathbf{P_{opt}}\right)\\ &= \frac{1}{M}\sum_{i=0}^{M-1}\overline{D}(R_i)\left[\int_{-0.5}^{0.5}P_{opt,i}(f)\sqrt{S_i(f)}\,df\right]^2\\ &\quad+\frac{1}{M}\sum_{i=0}^{M-1}\int_{-0.5}^{0.5}\left|P_{opt,i}(f)-1\right|^2S_i(f)\,df \qquad\qquad (\mathrm{B.7})\\ &= \frac{1}{M}\sum_{i=0}^{M-1}\overline{D}(R_i)\left[\frac{\int_+\sqrt{S_i(f)}\,df}{1+\left|\mathcal{F}_i^+\right|\overline{D}(R_i)}\right]^2\end{aligned}$$

$$+\frac{1}{M}\sum_{i=0}^{M-1}|\mathcal{F}_i^+|\left[\frac{\overline{D}(R_i)}{1+|\mathcal{F}_i^+|\,\overline{D}(R_i)}\int_+\sqrt{S_i(f)}\,df\right]^2+\frac{1}{M}\sum_{i=0}^{M-1}\int_- S_i(f)\,df$$

$$=\frac{1}{M}\sum_{i=0}^{M-1}\frac{\overline{D}(R_i)}{1+|\mathcal{F}_i^+|\,\overline{D}(R_i)}\left[\int_+\sqrt{S_i(f)}\,df\right]^2+\frac{1}{M}\sum_{i=0}^{M-1}\int_- S_i(f)\,df.$$

The expression above is (3.14) and is valid for any given bit allocation $\{R_i\}$. ■

C Computation of Support Sets $\mathcal{F}_i^+$

The sets $\mathcal{F}_i^+$ implicitly defined in (3.12) can be written as

$$\mathcal{F}_i^+ = \left\{f : -0.5 \le f \le 0.5 \text{ and } \sqrt{S_i(f)} > T_i^*\right\},$$

where $T_i^* \triangleq \frac{\overline{D}(R_i)}{1+|\mathcal{F}_i^+|\overline{D}(R_i)}\int_+\sqrt{S_i(f)}\,df$. Define the sets $F_i(T) = \left\{f : 0.5 \le f \le 0.5 \text{ and } \sqrt{S_i(f)} > T\right\}$ and the functions $Q_i(T) = \int_{F_i(T)}\sqrt{S_i(f)}\,df$. The threshold T_i^* is the solution to the nonlinear equation

$$T = f_i(T), \tag{C.1}$$

where

$$f_i(T) \triangleq \frac{\overline{D}(R_i)}{1+|F_i(T)|\,\overline{D}(R_i)}Q_i(T).$$

See Fig. 7 for an illustration. As noted in Remark 2 in Sec. 3.2, the sets $\mathcal{F}_i^+$ are unique, so the solution T_i^* to (C.1) is also unique.

$|F_i(T)|$ and $Q_i(T)$ are nonincreasing functions of T. If $|F_i(T)|$ and $Q_i(T)$ are differentiable with respect to T, then

$$\frac{d\,|F_i|}{dT} = -\sum_j \frac{1}{\left|d\sqrt{S_i(f)}/df\right|_{f=f_j}}, \qquad \frac{dQ_i}{dT} = T\frac{d\,|F_i|}{dT}, \quad 0 \le i < M,$$

where $\{f_j\}$ are the roots of $\sqrt{S_i(f)} = T$. After some algebraic manipulations, we find

$$\frac{df_i}{dT} = \frac{\overline{D}(R_i)}{1+|F_i(T)|\,\overline{D}(R_i)}\left[T - f_i(T)\right]\frac{d\,|F_i|}{dT}, \quad 0 \le i < M,$$

which implies that $\frac{df_i}{dT}$ has the same sign as $f_i(T) - T$. Since $f_i(0) = \frac{\overline{D}(R_i)}{1+\overline{D}(R_i)}\int_{-0.5}^{0.5}\sqrt{S_i(f)}\,df > 0$, we directly obtain the following:

Claim: For all $T < T_i^*$, we have $T < f_i(T) \le f_i(T_i^*) = T_i^*$. (C.2)

To establish this claim, we assumed differentiability of $|F_i(T)|$ and $Q_i(T)$. However, the existence of plateaus in $\sqrt{S_i(f)}$ (as in the example of Fig. 7a) would give rise to jumps in both $|F_i(T)|$ and

$Q_i(T)$. We now show that the claim holds even in this case. The functions $|F_i(T)|$ and $Q_i(T)$ are right-continuous and have the same set of discontinuities. We have

$$\lim_{\epsilon\to 0}|F_i(T-\epsilon)| = |F_i(T)| + |A_i(T)|, \quad \lim_{\epsilon\to 0} Q_i(T-\epsilon) = Q_i(T) + T\,|A_i(T)|, \quad 0 \le i < M,$$

where $A_i(T) \triangleq \left\{ f : -0.5 \le f \le 0.5 \text{ and } \sqrt{S_i(f)} = T \right\}$ are the level sets of $\sqrt{S_i(f)}$. It can easily be shown that jumps of $f_i(T)$ are given by

$$\Delta f_i(T) \triangleq f_i(T) - \lim_{\epsilon\to 0} f_i(T-\epsilon) = \frac{|A_i(T)|\,\overline{D}(R_i)}{1 + (|F_i(T)| + |A_i(T)|)\,\overline{D}(R_i)}\left[f_i(T) - T\right], 0 \le i < M. \quad \text{(C.3)}$$

Note that $\Delta f_i(T_i^*) = 0$ and that $\Delta f_i(T)$ has the same sign as $f_i(T) - T$. Hence the claim (C.2) still holds.

This suggests the use of the following iterative algorithm to find T_i^* in each channel (see Fig. 7b for an illustration):

1. Initialize $T_i^0 = 0$, and $k = 0$.
2. Use the update rule $T_i^{k+1} = f\left(T_i^k\right)$.

The claim (C.2) guarantees that the sequence T_i^k is nondecreasing and converges to T_i^*.

D Proof of Proposition 3.5

The cost functional (3.5) viewed as a function of both $\mathbf{P}$ and $\{R_i\}$ is the sum of M functionals

$$\overline{\mathcal{E}}\left(\{R_i\}, \mathbf{P}\right) = \sum_{i=0}^{M-1} F_i\left(R_i, P_i\right),$$

where

$$F_i\left(R, P\right) \triangleq \overline{D}(R)\left(\int_{-0.5}^{0.5} P(f)\sqrt{S_i(f)}\,df\right)^2 + \int_{-0.5}^{0.5}\left(P(f)-1\right)^2 S_i(f)\,df$$

depends on i via S_i. A sufficient condition for $\overline{\mathcal{E}}$ to be strictly convex in $(\{R_i\}, \mathbf{P})$ is that each F_i be strictly convex in (R, P). A sufficient condition for strict convexity of F_i is the positivity of its second order variation evaluated for all (R, P). (Analogously, positive definiteness of Hessians of functions defined on $\Re^n$ is a sufficient condition for the strict convexity of these functions). Let $(\delta R, \delta P)$ be any admissible variation of the variables (R, P). The claim will be proven by showing that the second derivative of $F\left(R + \alpha\delta R, P + \alpha\delta P\right)$ with respect to the real variable α, evaluated at $\alpha = 0$, is strictly positive for any admissible $(\delta R, \delta P)$. We have

$$\left.\frac{\partial^2 F_i}{\partial\alpha^2}\right|_{\alpha=0} = (\delta R)^2\,\overline{D}''(R)U_i^2 + 4\,(\delta R)\,\overline{D}'(R)U_i\,(\delta U_i) + 2\overline{D}(R)\,(\delta U_i)^2 + 2\int_{-0.5}^{0.5}\left(\delta P(f)\right)^2 S_i(f)\,df, \quad \text{(D.1)}$$

where

$$U_i \triangleq \int_{-0.5}^{0.5} P(f)\sqrt{S_i(f)}\,df, \quad \delta U_i \triangleq \int_{-0.5}^{0.5} \delta P(f)\sqrt{S_i(f)}\,df.$$

Using the Cauchy-Schwartz inequality on the rightmost term of (D.1), we have the following tight lower bound:

$$\left.\frac{\partial^2 F_i}{\partial \alpha^2}\right|_{\alpha=0} \geq (\delta R)^2\,\overline{D}''(R)U_i^2 + 4\,(\delta R)\,\overline{D}'(R)U_i\,(\delta U_i) + \left(2\overline{D}(R)+2\right)(\delta U_i)^2\,. \tag{D.2}$$

The right hand side of (D.2) is quadratic in δR and is strictly positive if the discriminant is negative:

$$8U_i^2\,(\delta U_i)^2\left[2\left(\overline{D}'(R)\right)^2 - \overline{D}''(R)\left(\overline{D}(R)+1\right)\right] < 0.$$

Hence (3.19) follows directly. ■

E Optimal Filters for Exponential Rate-Distortion Function

For notational simplicity, we use $\sigma^2_{opt,i}$ in place of $\sigma^2_{opt,i}/c_i^2$. (Equivalently, we assume that $c_i \equiv 1$.)

Claim : Under optimal bit allocation, the filters that minimize the MSE (4.4) are given by (3.10) and (3.11), with product filters

$$P_{opt,i}(f) = \max\left(0, 1 - \frac{\mathcal{E}_q}{\sigma^2_{opt,i}}\frac{1}{\sqrt{S_i(f)}}\right), \quad 0 \leq i < M, \quad f \in [-0.5, 0.5]. \tag{E.1}$$

The constants $\sigma^2_{opt,i}$ and $\mathcal{E}_q$ are solutions to the nonlinear system

$$\sigma^2_{opt,i} = \frac{1}{2}\int_+ \sqrt{S_i(f)}\,df \pm \sqrt{\left(\frac{1}{2}\int_+ \sqrt{S_i(f)}\,df\right)^2 - \left|\mathcal{F}_i^+\right|\mathcal{E}_q}, \quad 0 \leq i < M. \tag{E.2}$$

$$\mathcal{E}_q = \gamma 2^{-2R}\left[\prod_{i=0}^{M-1}\sigma^2_{opt,i}\right]^{2/M}. \tag{E.3}$$

$$\mathcal{F}_i^+ = \left\{f : -0.5 \leq f \leq 0.5 \text{ and } \sqrt{S_i(f)} > \frac{\mathcal{E}_q}{\sigma^2_{opt,i}}\right\}, \quad 0 \leq i < M. \tag{E.4}$$

The MSE for the optimal filters is given by

$$\mathcal{E}_{min} = \mathcal{E}_q + \frac{1}{M}\sum_{i=0}^{M-1}\left[\frac{\mathcal{E}_q}{\sigma^2_{opt,i}}\right]^2\left|\mathcal{F}_i^+\right| + \frac{1}{M}\sum_{i=0}^{M-1}\int_- S_i(f)\,df. \tag{E.5}$$

Proof: Substituting (B.6) in (B.4), we obtain

$$P_{opt,i}(f) = \max\left(0, 1 - \frac{\overline{D}(R_i)}{\sqrt{S_i(f)}}\,\sigma^2_{opt,i}\right), \quad 0 \leq i < M, \quad f \in [-0.5, 0.5]. \tag{E.6}$$

The first term in the right hand side of (4.4),

$$\gamma 2^{-2R}\left[\prod_{i=0}^{M-1}\sigma_i||\tilde{h}_i||\right]^{2/M}, \tag{E.7}$$

is inversely proportional to the classical coding gain for PR systems. For optimal filters, (B.6) is satisfied (with $c_i \equiv 1$), so (E.7) takes the form (E.3). Using (4.1) (equality of distortions in all channels), we obtain from (E.3)

$$\sigma^4_{opt,i}\overline{D}(R_i) = \mathcal{E}_q, \quad 0 \le i < M. \tag{E.8}$$

Substituting (E.8) into (E.6), we obtain (E.1). Multiplying both sides of (E.1) by $\sqrt{S_i(f)}$ and integrating, we obtain

$$\sigma^2_{opt,i} = \int_+ \sqrt{S_i(f)}\,df - \left|\mathcal{F}_i^+\right| \frac{\mathcal{E}_q}{\sigma^2_{opt,i}}, \quad 0 \le i < M.$$

Therefore, $\sigma^2_{opt,i}$ satisfies the quadratic equation

$$\sigma^4_{opt,i} - \sigma^2_{opt,i}\int_+ \sqrt{S_i(f)}\,df + \left|\mathcal{F}_i^+\right|\mathcal{E}_q = 0, \quad 0 \le i < M, \tag{E.9}$$

whose solution is given by (E.2). Equations (E.2), (E.3) constitute a nonlinear system of $M+1$ equations with $M+1$ unknowns $\{\sigma^2_{opt,i}\}, \mathcal{E}_q$. The nonlinearity of this system is compounded by the dependency of the domain of integration $\mathcal{F}_i^+$ on $\mathcal{E}_q$ and $\sigma^2_{opt,i}$. Given $\{\sigma^2_{opt,i}\}, \mathcal{E}_q$, one can solve for $\mathcal{F}_i^+$ using the algorithm in Appendix C. Given $\mathcal{F}_i^+$, the system (E.2) (E.3) can be solved numerically. The optimal pre- and post-filters are again given by (3.10) and (3.11), respectively.

To find the optimum value of the MSE, we substitute (B.6) and (E.1) into the first and second terms of (B.7), respectively. We obtain

$$\mathcal{E}_{min} = \frac{1}{M}\sum_{i=0}^{M-1}\overline{D}(R_i)\sigma^4_{opt,i} + \frac{1}{M}\sum_{i=0}^{M-1}\left[\frac{\mathcal{E}_q}{\sigma^2_{opt,i}}\right]^2\left|\mathcal{F}_i^+\right| + \frac{1}{M}\sum_{i=0}^{M-1}\int_- S_i(f)\,df. \tag{E.10}$$

Substituting (E.8) into the first term of (E.10), we obtain (E.5). ■

The discriminant in (E.2) is guaranteed to be nonnegative for all i because there always exists a solution to the optimization problem, as noted in the remark at the end of Sec. 2.2. This solution is unique because the condition (3.20) is assumed to be satisfied.

According to (E.2), there apparently exist two possible candidates for the solution $\sigma^2_{opt,i}$ in each channel, depending on the choice of the sign. Only one of them corresponds to the (unique) solution. At high bit rates, the solution tends to the IIR biorthogonal solution, and the optimal system tends to a PR system. In this case, $\sigma^2_{opt,i}$ tends to the PR solution $\int_{-0.5}^{0.5}\sqrt{S_i(f)}\,df$, meaning that all signs in (E.2) must positive.

Remark When the condition (3.15) is satisfied, the optimal product filters are given by

$$P_{opt,i}(f) = 1 - \frac{\mathcal{E}_q}{\sigma^2_{opt,i}}\frac{1}{\sqrt{S_i(f)}}, \quad 0 \le i < M, \;\; f \in [-0.5, 0.5],$$

in which case

$$\mathcal{E}_{min} = \mathcal{E}_q + \frac{1}{M}\sum_{i=0}^{M-1}\left(\frac{\mathcal{E}_q}{\sigma^2_{opt,i}}\right)^2, \tag{E.11}$$

where $\sigma^2_{opt,i}$ and $\mathcal{E}_q$ are the solutions to the nonlinear system (E.2),(E.3).

F Alternative Proof of Theorem 3.3

In this section, we use intensively the parameterization of $\mathcal{H}$ and $\mathcal{H}^\dagger$ as obtained from Corollary A.2:

$$\tilde{\mathcal{H}}(f) = \tilde{\mathcal{U}}(f)\tilde{\mathcal{G}}(f) \tag{F.1}$$

$$\mathcal{H}(f) = \mathcal{G}(f)\mathcal{W}(f)\Lambda^{-0.5}(f)\mathcal{V}^\dagger(f), \tag{F.2}$$

where $\mathcal{V}\Lambda\tilde{\mathcal{V}}$ is the eigenvalue decomposition of S, $\tilde{\mathcal{G}}$ and $\mathcal{G}$ are diagonal matrices with nonnegative entries and $\mathcal{U}$,$\tilde{\mathcal{U}}$, $\mathcal{W}$ are complex paraunitary matrices.

Lemma F.1

$$\tfrac{1}{M}\;\; Tr\int_{-0.5}^{0.5}(\tilde{\mathcal{H}}(f)\mathcal{H}(f) - I_M)\mathcal{S}(f)(\tilde{\mathcal{H}}(f)\mathcal{H}(f) - I_M)^\dagger\, df \ge \tag{F.3}$$

$$\tfrac{1}{M}\;\; \int_{-0.5}^{0.5}\sum \Lambda_{(i)}(f) + \mathcal{P}^2_{(i)}(f) - 2\Lambda^{0.5}_{(i)}(f)\mathcal{P}_{(i)}(f), \tag{F.4}$$

where $\mathcal{P} = \tilde{\mathcal{G}}\mathcal{G}$.

Proof For compactness of notation, we elude the dependence on f of the matrices involved. Using the parameterizations of $\mathcal{H}$ and $\tilde{\mathcal{H}}$, we obtain

$$Tr(\tilde{\mathcal{H}}\mathcal{H}\mathcal{S}\mathcal{H}^\dagger\mathcal{H}) = Tr(\tilde{\mathcal{U}}\tilde{\mathcal{G}}\mathcal{G}\mathcal{W}\Lambda^{-0.5}\mathcal{V}^\dagger\mathcal{V}\Lambda\mathcal{V}^\dagger\mathcal{V}\Lambda^{-0.5}\mathcal{W}^\dagger\mathcal{G}\tilde{\mathcal{G}}\tilde{\mathcal{U}}^\dagger) = \tag{F.5}$$

$$Tr(\tilde{\mathcal{U}}\tilde{\mathcal{G}}\mathcal{G}\mathcal{G}\tilde{\mathcal{G}}\tilde{\mathcal{U}}^\dagger) = Tr(\mathcal{P}^2), \tag{F.6}$$

the last equation arising from the invariance property $Tr(\mathcal{U}\mathcal{A}\mathcal{U}^\dagger) = Tr\mathcal{A}$, for $\mathcal{U}$ paraunitary. We also have that

$$Tr(\mathcal{P}^2) = \sum_i \mathcal{P}^2_i = \sum_i \mathcal{P}^2_{(i)} \tag{F.7}$$

since the sum is invariant to permutations of its own terms. At the same time,

$$Tr(\tilde{\mathcal{H}}\mathcal{H}\mathcal{S}) = Tr(\tilde{\mathcal{U}}\tilde{\mathcal{G}}\mathcal{G}\mathcal{W}\Lambda^{-0.5}\mathcal{V}^{\dagger}\mathcal{V}\Lambda\mathcal{V}^{\dagger}) = Tr(\mathcal{V}^{\dagger}\mathcal{U}\mathcal{P}\mathcal{W}\Lambda^{0.5}). \tag{F.8}$$

Thus,

$$Re(Tr(\tilde{\mathcal{H}}\mathcal{H}\mathcal{S})) = Tr(\tilde{\mathcal{U}}\tilde{\mathcal{G}}\mathcal{G}\mathcal{W}\Lambda^{-0.5}\mathcal{V}^{\dagger}\mathcal{V}\Lambda\mathcal{V}^{\dagger}) = Re(Tr(\mathcal{V}^{\dagger}\mathcal{U}\mathcal{P}\mathcal{W}\Lambda^{0.5})) \le \sum \mathcal{P}_{(i)}\lambda_{(i)}^{0.5} \tag{F.9}$$

where the last relation is a consequence of the Wielandt-Hoffman theorem for nonsymmetric matrices [18, Corollary 7.3.8]. We are now in position to write

$$\frac{1}{M} \quad Tr\int_{-0.5}^{0.5} (\tilde{\mathcal{H}}\mathcal{H} - I_M)\mathcal{S}(\tilde{\mathcal{H}}\mathcal{H} - I_M)^{\dagger}\, df = \tag{F.10}$$

$$\frac{1}{M} \quad Tr\int_{-0.5}^{0.5} Tr(\tilde{\mathcal{H}}\mathcal{H}\mathcal{S}\mathcal{H}^{\dagger}\mathcal{H}) + Tr(\mathcal{S}) - 2Re(Tr(\tilde{\mathcal{H}}\mathcal{H}\mathcal{S})) \ge \tag{F.11}$$

$$\frac{1}{M} \quad \int_{-0.5}^{0.5} \sum \Lambda_{(i)} + \mathcal{P}_{(i)}^2 - 2\Lambda_{(i)}^{0.5}\mathcal{P}_{(i)}, \tag{F.12}$$

by (F.6), (F.9). Since $Tr(\mathcal{S}) = \sum \Lambda_{(i)}$, the proof is complete. □

Lemma F.2 *Let $a \ge c$ and $b > 0$. Let $R(t)$ be the unique minimizer, for fixed t, of the expression*

$$E(R,t) = \overline{D}(R)(a+bt)^2 + \overline{D}(B-R)(c-bt)^2 \tag{F.13}$$

Then $\frac{d}{dt}E(R(t),t) < 0$, $\forall 0 < t < \frac{c}{b}$.

Proof For fixed t, the extremality conditions require $F(R,t) = \overline{D}'(R)(a+bt)^2 - \overline{D}'(B-R)(c-bt)^2 = 0$, which defines $R(t)$ implicitly. We have that $\frac{d}{dR}F(R,t) = \overline{D}''(R)(a+bt)^2 + \overline{D}''(B-R)(c-bt)^2 > 0$, $\forall t, R$, since, from our assumptions, $\overline{D}(R)$ is strictly convex. Thus, the solution to $F(R,t) = 0$ is unique, and $R(t)$ is differentiable by the implicit function theorem (the fact that one solution exists is the result of our high bit-rate assumption). Also, since $\overline{D}(R)$ is convex, $\overline{D}'(R)$ is increasing; thus $F(R,t) = 0$, together with our assumptions about a, b, c, imply that $R > B - R$.

Since $\frac{1}{\overline{D}(R)} = exp(-\ln(\overline{D}(R)))$ is a convex function, it follows that

$$-\frac{\overline{D}'(R)}{\overline{D}^2(R)} > -\frac{\overline{D}'(B-R)}{\overline{D}^2(B-R)} \tag{F.14}$$

and thus,

$$-\frac{\overline{D}'(R)(a+bt)^2}{\overline{D}^2(R)(a+bt)^2} > -\frac{\overline{D}'(B-R)(c-bt)^2}{\overline{D}^2(B-R)(c-bt)^2}. \tag{F.15}$$

In the previous expression, the denominators are positive and the numerators are positive and equal (by the extremal condition $F(R,t)=0$ and since $\overline{D}(R)$ is strictly decreasing). This results in

$$\overline{D}^2(R)(a+bt)^2 < \overline{D}^2(B-R)(c-bt)^2 \tag{F.16}$$

or

$$\overline{D}(R)(a+bt) < \overline{D}(B-R)(c-bt). \tag{F.17}$$

We now have

$$\begin{aligned}\frac{d}{dR}E(R(t),t) &= 2b(\overline{D}(R)(a+bt)-\overline{D}(B-R)(c-bt))+F(R,t)\frac{dR}{dt} &\text{(F.18)}\\ &= 2b(\overline{D}(R)(a+bt)-\overline{D}(B-R)(c-bt))<0 &\text{(F.19)}\end{aligned}$$

by (F.17) and $b>0$. The proof is complete. $\square$

Lemma F.3 *Assume that the bit rates are allocated such that* $R_0 \geq R_1 ... \geq R_{M-1}$. *Then there exists a choice of bit rates* $B_0 \geq B_1 ... \geq B_{M-1}$ *such that*

$$\frac{1}{M}\sum_{i=0}^{M-1}\overline{D}(R_i)\sigma_i^2||\tilde{h}_i||^2 \geq \frac{1}{M}\sum_{i=0}^{M-1}\overline{D}(B_i)(\int_{-0.5}^{0.5}\mathcal{P}_{(i)})^2. \tag{F.20}$$

Proof: We will first prove this result for two channels. In this case, we have that

$$\frac{1}{2}\sum_{i=0}^{1}\overline{D}(R_i)\sigma_i^2||\tilde{h}_i||^2 \geq = \frac{1}{2}\sum_{i=0}^{1}\overline{D}(R_i)\int_{-0.5}^{0.5}(\mathcal{G}_i)^2\int_{-0.5}^{0.5}(\tilde{\mathcal{G}}_i)^2 = \tag{F.21}$$

$$\frac{1}{2}\sum_{i=0}^{1}\overline{D}(R_i)\int_{-0.5}^{0.5}(\mathcal{G}_i)^2\int_{-0.5}^{0.5}(\tilde{\mathcal{G}}_i)^2 \geq = \frac{1}{2}\sum_{i=0}^{1}\overline{D}(R_i)(\int_{-0.5}^{0.5}(\mathcal{G}_i\tilde{\mathcal{G}}_i))^2 = \frac{1}{2}\sum_{i=0}^{1}\overline{D}(R_i)(\int_{-0.5}^{0.5}\mathcal{P}_i)^2 \tag{F.22}$$

as a result of the Cauchy inequality. Assume now that $a=\int_{-0.5}^{0.5}\mathcal{P}_0 \geq \int_{-0.5}^{0.5}\mathcal{P}_1 = c$ (otherwise we can simply switch the indices).

Let $\mathcal{F}=\{f\in[-0.5,0.5]|\mathcal{P}_0(f)<\mathcal{P}_1(f)\}$. If this set is of measure zero, then the lemma is proved. If $\mathcal{F}$ does not have measure zero, take $\mathcal{P}_{0t}=\mathcal{P}_0+t\chi(\mathcal{F})(\mathcal{P}_1-\mathcal{P}_0)$, and $\mathcal{P}_{1t}=\mathcal{P}_1-t\chi(\mathcal{F})(\mathcal{P}_1-\mathcal{P}_0)$, where $\chi(\mathcal{F})$ is the characteristic function of the set $\mathcal{F}$. Clearly $\mathcal{P}_{0t},\mathcal{P}_{1t}\geq 0$, $\forall t\in[0,1]$. Let $b=\int_{-0.5}^{0.5}\chi(\mathcal{F})(\mathcal{P}_1-\mathcal{P}_0)>0$. Clearly

$$b=\int_{-0.5}^{0.5}\chi(\mathcal{F})(\mathcal{P}_1-\mathcal{P}_0) \leq \int_{-0.5}^{0.5}\mathcal{P}_1 = c. \tag{F.23}$$

Let $R_0(t), R_1(t)$ be the results of adapting the bit rates to minimize

$$\begin{aligned}E(R_0,t) &= \frac{1}{2}(\overline{D}(R_0)(\int_{-0.5}^{0.5}\mathcal{P}_{0t})^2+\overline{D}(2R-R_0)(\int_{-0.5}^{0.5}\mathcal{P}_{1t})^2 &\text{(F.24)}\\ &= \frac{1}{2}(\overline{D}(R_0)(a+tb)^2+\overline{D}(2R-R_0)(c-tb)^2). &\text{(F.25)}\end{aligned}$$

From the results of our previous lemma, for $0 \le t \le 1 \le \frac{c}{b}$, we have that $\frac{d}{dt}E(R_0(t), t) < 0$, and thus $E(R_0(0), 0) > E(R_0(1), 1)$. But $\mathcal{P}_{01} = \max \mathcal{P}_0, \mathcal{P}_1 = \mathcal{P}_{(0)}$ and $\mathcal{P}_{11} = \min \mathcal{P}_0, \mathcal{P}_1 = \mathcal{P}_{(1)}$, which proves the claim for two channels.

For multiple channels, the two-channel approach is used to prove that we can replace $\mathcal{P}_0(f)$ and $\mathcal{P}_1(f)$ with $\max(\mathcal{P}_0(f), \mathcal{P}_1(f))$ and $\min(\mathcal{P}_0(f), \mathcal{P}_1(f))$, respectively, and then $\mathcal{P}_0(f)$ and $\mathcal{P}_2(f)$ with $\max(\mathcal{P}_0(f), \mathcal{P}_2(f))$ and $\min(\mathcal{P}_0(f), \mathcal{P}_2(f))$, and so on, until in channel 0 we have $\mathcal{P}_{(0)}$. The procedure is repeated on $\mathcal{P}_{(1)}$, $\mathcal{P}_{(2)}$ until all channels are exhausted. The proof is complete. □

The main result of this section is summarized in the following theorem

Theorem F.4 *Let $\mathcal{H}$ and $\tilde{\mathcal{H}}$ be two filters with the structure given by Corollary A.2. Let $\mathcal{V}\Lambda\mathcal{V}^\dagger = \mathcal{S}$ be the eigenvalue decomposition of $\mathcal{S}$ and $\mathcal{Q}(f)$ the permutation matrices such that $\mathcal{Q}(f)\Lambda(f)\mathcal{Q}(f)^\dagger = diag(\Lambda_{(i)}(f) = \Lambda^o(f)$ (the permutation-matrix-valued function that orders the eigenvalue of $\mathcal{S}$ in a decreasing order). Let $\mathcal{P} = \mathcal{G}\tilde{\mathcal{G}} > 0$ and $\mathcal{G}^o = diag(\sqrt{\mathcal{P}_{(i)}})$. Then, there exists a choice of bit rates B_i such that*

$$\mathcal{E}(\mathcal{H}, \tilde{\mathcal{H}}, R) \ge \frac{1}{M}\sum_{i=0}^{M-1}\overline{D}(B_i)(\int_{-0.5}^{0.5}\mathcal{P}_{(i)})^2 + \frac{1}{M}\int_{-0.5}^{0.5}(\sum \Lambda_{(i)}(f) + \mathcal{P}_{(i)}^2 - 2\Lambda_{(i)}^{0.5}\mathcal{P}_{(i)}) = \mathcal{E}(\mathcal{H}^o, \tilde{\mathcal{H}}^o, B), \tag{F.26}$$

where $\tilde{\mathcal{H}}^o = \mathcal{V}\mathcal{Q}^\dagger\mathcal{G}^o$ and $\mathcal{H}^o = \mathcal{G}^o\mathcal{Q}\Lambda^{-0.5}\mathcal{V}^\dagger$. For this configuration, the inequality is, therefore, attained.

Proof: The inequality follows instantly from Lemmas F.3 and F.1. To check that the inequality is attained, we need to verify only the part, following the details of the proof of Lemma F.1:

$$Re(Tr(\tilde{\mathcal{H}}^o\mathcal{H}^o\mathcal{S})) = Re(Tr(\mathcal{V}\mathcal{Q}^\dagger\mathcal{G}^o\mathcal{G}^o\mathcal{Q}\Lambda^{-0.5}\mathcal{V}^\dagger\mathcal{V}\Lambda\mathcal{V}^\dagger)) = \tag{F.27}$$

$$Re(Tr(\mathcal{V}\mathcal{Q}^\dagger\mathcal{G}^o\mathcal{G}^o\mathcal{Q}\Lambda^{0.5}\mathcal{V}^\dagger)) = Re(Tr(\mathcal{Q}^\dagger\mathcal{G}^o\mathcal{G}^o\mathcal{Q}\Lambda^{0.5})) = \tag{F.28}$$

$$Re(Tr(\mathcal{G}^o\mathcal{G}^o\mathcal{Q}\Lambda^{0.5}\mathcal{Q}^\dagger)) = Re(Tr(\mathcal{P}^o(\Lambda^o)^{0.5})) = \sum \Lambda_{(i)}^{0.5}\mathcal{P}_{(i)} \tag{F.29}$$

where we used $Tr(QAQ^\dagger) = Tr(A)$ for Q paraunitary. The proof is complete. □

F.1 Optimization Strategy

1. For $\mathcal{S}(f)$ given, compute the eigenvalue decomposition $\mathcal{V}(f)\Lambda(f)\mathcal{V}(f)^\dagger$.
2. Compute the permutation matrices $\mathcal{Q}(f)$ such that $\mathcal{Q}(f)\Lambda(f)\mathcal{Q}^\dagger(f) = diag(\Lambda_{(i)})$ (the permutations that order $\Lambda(f)$ decreasingly).

3. Minimize the functional (which is convex under the conditions specified)

$$\frac{1}{M}\sum_{i=0}^{M-1}\overline{D}(R_i)(\int_{-0.5}^{0.5}\mathcal{P}_i)^2+\frac{1}{M}\int_{-0.5}^{0.5}(\sum\Lambda_{(i)}(f)+\mathcal{P}_i^2-2\Lambda_{(i)}^{0.5}\mathcal{P}_i) \tag{F.30}$$

Note that the $\mathcal{P}_i$ need not be ordered, for the theorem guarantees that they will be ordered at the solution (or some degeneracy happens that will allow for an ordered solution anyway). Or alternatively, one may impose the ordering, resulting in a smaller search set but with more inequalities to enforce.

4. Choose as solutions $\tilde{\mathcal{H}}^o = \mathcal{V}\mathcal{Q}^\dagger\mathcal{G}^o$ and $\mathcal{H}^o = \mathcal{G}^o\mathcal{Q}\Lambda^{-0.5}\mathcal{V}^\dagger$, where $(\mathcal{G}^o)^2 = \mathcal{P}$.

References

[1] M. Smith and T. Barnwell III, "Exact reconstruction techniques for tree-structured subband coders," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, pp. 434–441, Mar. 1986.

[2] M. Vetterli, "Filter banks allowing perfect reconstruction," *Signal Proc.*, vol. 10, pp. 219–244, Mar. 1986.

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

[4] T. A. Ramstad, "Analysis/synthesis filter banks with critical sampling," in *Int. Conf. on Digital Sig. Proc.*, Florence, Italy, 1984.

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

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

[7] R. A. Haddad and K. Park, "Modeling, analysis, and optimum design of quantized M-band filter banks," *IEEE Trans. Signal Process.*, vol. 43, pp. 2540–2549, Nov. 1995.

[8] P. P. Vaidyanathan and T. S. Chen, "Statistically optimal synthesis filter banks for subband coders," in *Proc. ASILOMAR*, Asilomar, CA, 1994.

[9] K. Gosse and P. Duhamel, "Perfect–reconstruction versus MMSE filter banks in source coding," *IEEE Trans. Signal Process.*, vol. 45, pp. 2188–2202, Sept. 1997.

[10] J. Tuqan and P. P. Vaidyanathan, "Statistically optimum pre– and postfiltering in quantization," *IEEE Trans. Circ. and Sys.–II*, vol. 44, pp. 1015–1031, Dec. 1997.

[11] P. Moulin, M. Anitescu, and K. Ramchandran, "Theory of rate-distortion-optimal, constrained filter banks–application to IIR and FIR biorthogonal designs," to appear in *IEEE Transactions on Signal Processing*.
Available from `http://www.ifp.uiuc.edu/~moulin/Papers/adaptedFB98b.ps.Z`.

[12] M. Unser, "An extension of the Karhunen-Loève transform for wavelets and perfect–reconstruction filterbanks," in *Proc. SPIE Conf. on Vis. Commun. and Image Proc.*, vol. 2034, pp. 45–56, 1993.

[13] M. K. Tsatsanis and G. B. Giannakis, "Principal component filter banks for optimal multiresolution analysis," *IEEE Trans. Signal Process.*, vol. 43, pp. 1766–1777, Aug. 1995.

[14] P. P. Vaidyanathan, *Multirate Systems and Filter Banks.* Englewood Cliffs, NJ: Prentice-Hall, 1993.

[15] P. P. Vaidyanathan, "Theory of optimal orthonormal subband coders," *IEEE Trans. Signal Process.*, vol. 46, pp. 1528–1543, June 1998.

[16] N. S. Jayant and P. Noll, *Digital Coding of Waveforms.* Englewood Cliffs, NJ: Prentice-Hall, 1984.

[17] D. P. Bertsekas, *Nonlinear Programming.* Belmont, Mass., Athena Scientific, 1995.

[18] R. A. Horn and C. R. Johnson, *Matrix Analysis.* Cambridge University Press, Cambridge, 1996.

[19] D. G. Luenberger, *Optimization by Vector Space Methods.* John Wiley, New York, 1969.

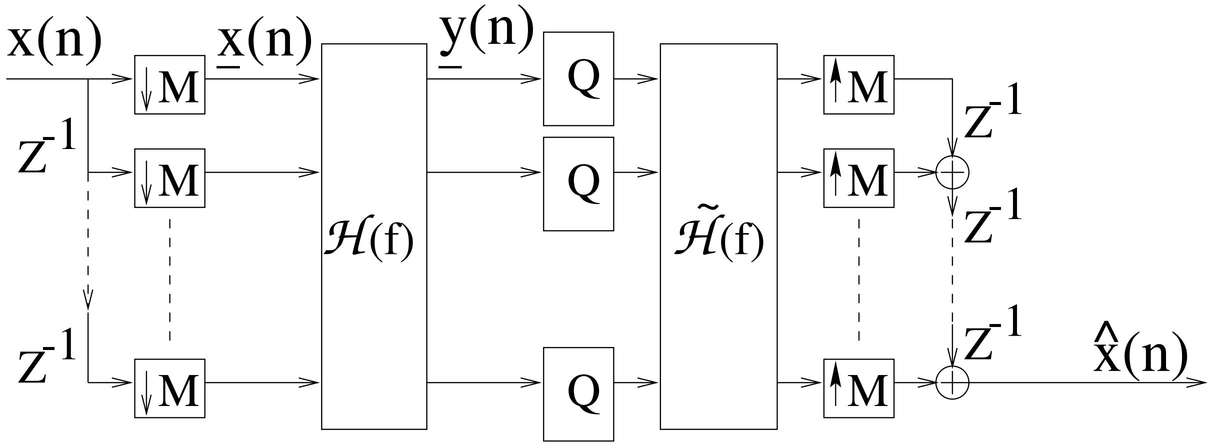

Figure 1: Polyphase representation of M–channel subband coder and decoder using analysis filters $H_i(f)$ and synthesis filters $\tilde{H}_i(f)$. The polyphase components of $H_i(f)$ (resp. $\tilde{H}_i(f)$) are contained in row (resp. column) i of the polyphase matrix $\mathcal{H}(f)$ (resp. $\tilde{\mathcal{H}}(f)$).

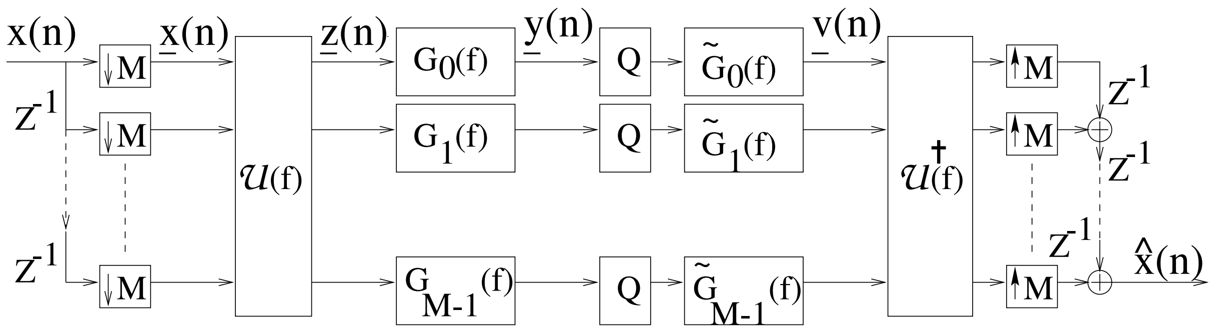

Figure 2: Cascade of a principal-component filter bank $\mathcal{U}(f)$ and a set of zero–phase prefilters $G_i(f)$ and postfilters $\tilde{G}_i(f)$ around each quantizer. The optimal subband coder in that class achieves the global minimum of the MSE (2.2).

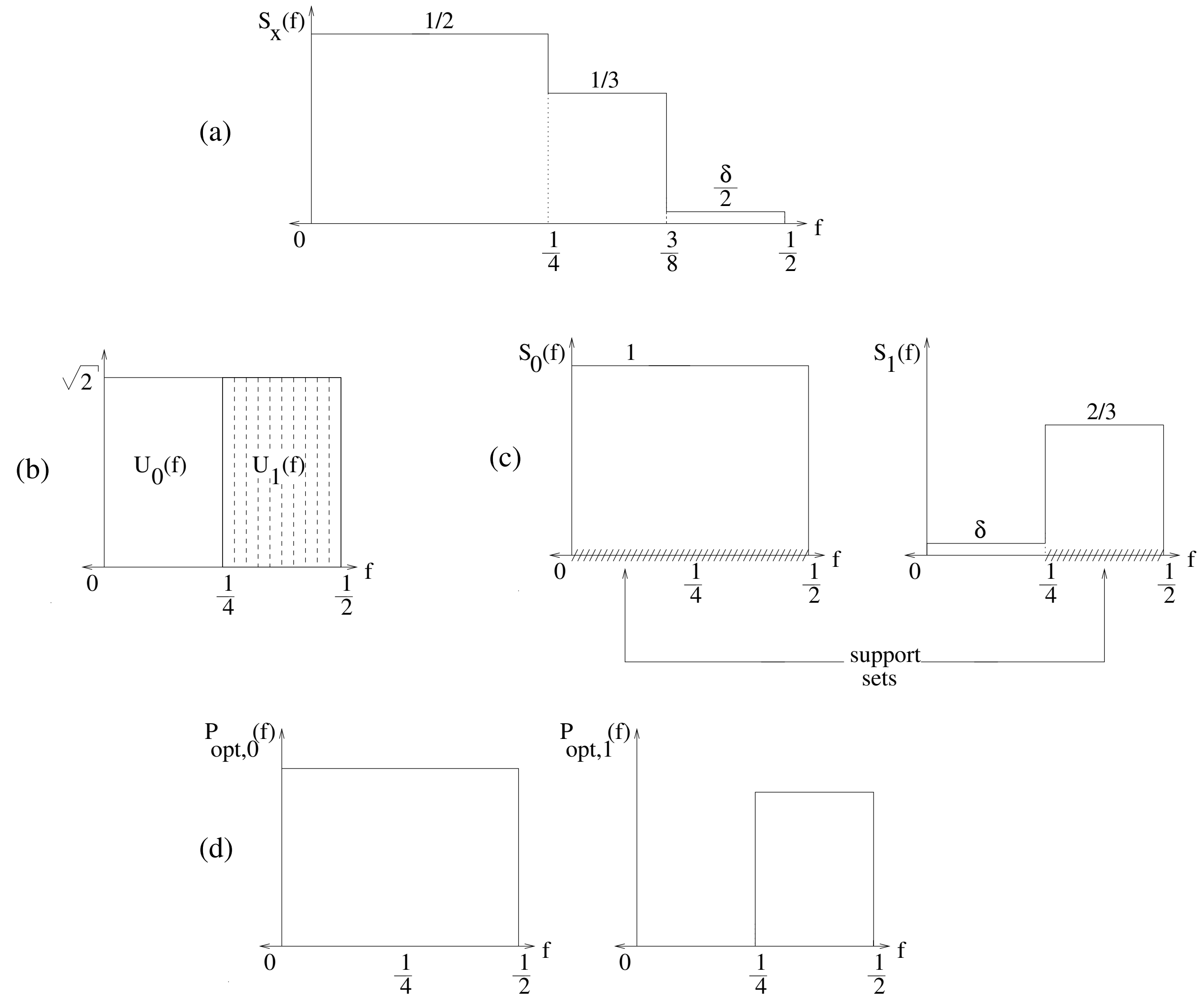

Figure 3: Toy example: (a) Spectral density of the input $x(n)$. (b) Frequency response of the two filters $U_0(f)$ and $U_1(f)$ in PCFB. These are the classical, "ideal" brickwall filters. (c) Corresponding spectral densities $S_0(f)$ and $S_1(f)$ together with the positive part of the support sets $\mathcal{F}_0^+$ and $\mathcal{F}_1^+$. (d) The optimal product filters $P_{opt,0}(f)$ and $P_{opt,1}(f)$. Note $P_{opt,1}(f)$ rejects the insignificant frequency components in the range $0 < |f| < 0.25$.

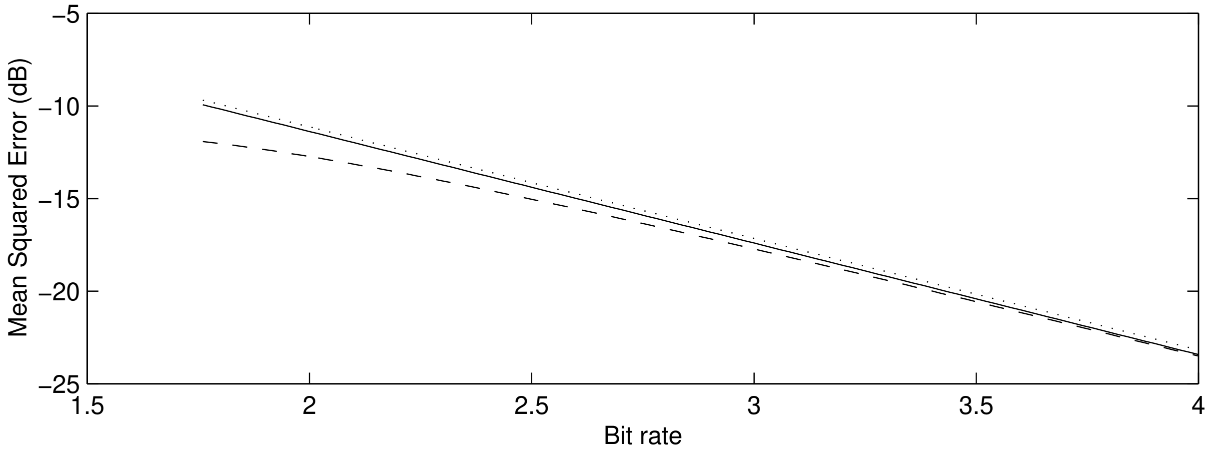

Figure 4: AR(1) process with correlation coefficient $r = 0.8$: optimum values of MSE as a function of overall bit rate R for different filter design methods, under exponential rate-distortion model. Solid line: IIR biorthogonal (half-whitening) filters [11]. Dotted line: FIR biorthogonal filters [11]. Dashed line: MMSE filters.

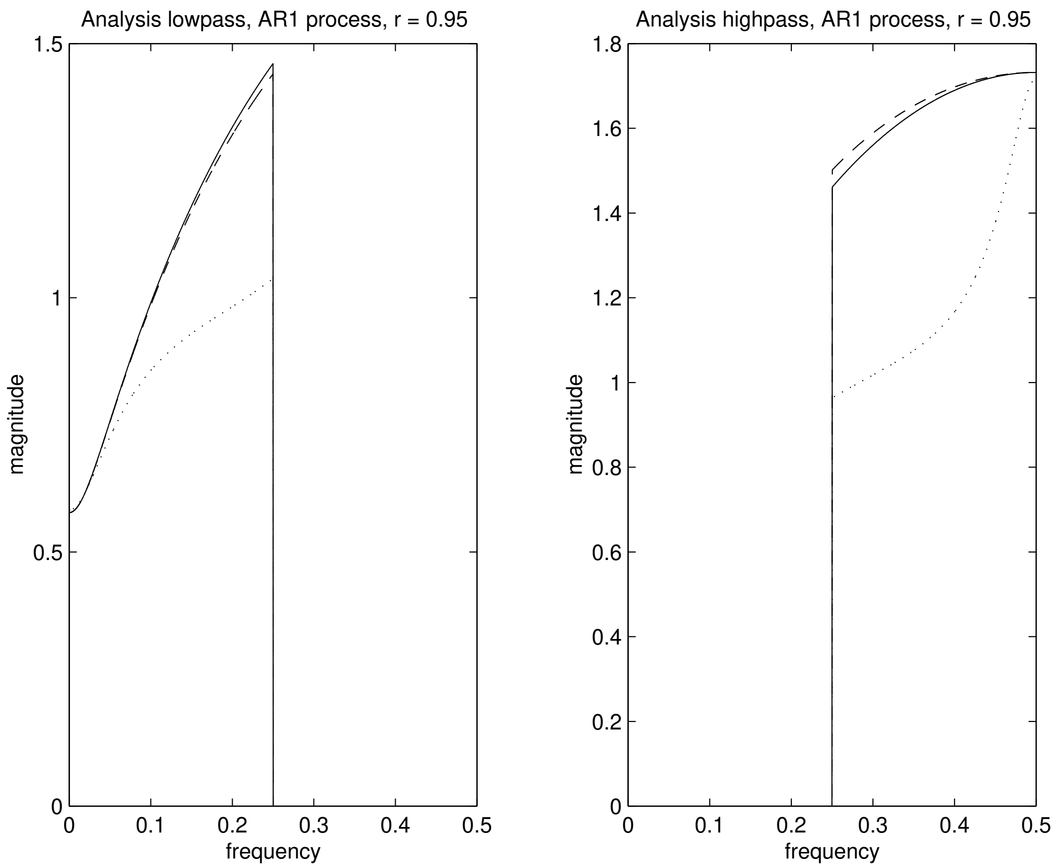

Figure 5: Frequency responses of optimal, unconstrained length analysis filters in two-channel coder, for AR(1) process at bit rate $R = 2.91$. Solid line: IIR biorthogonal (half-whitening) filters [11]. Dotted line: FIR biorthogonal filters [11]. Dashed line: MMSE filters.

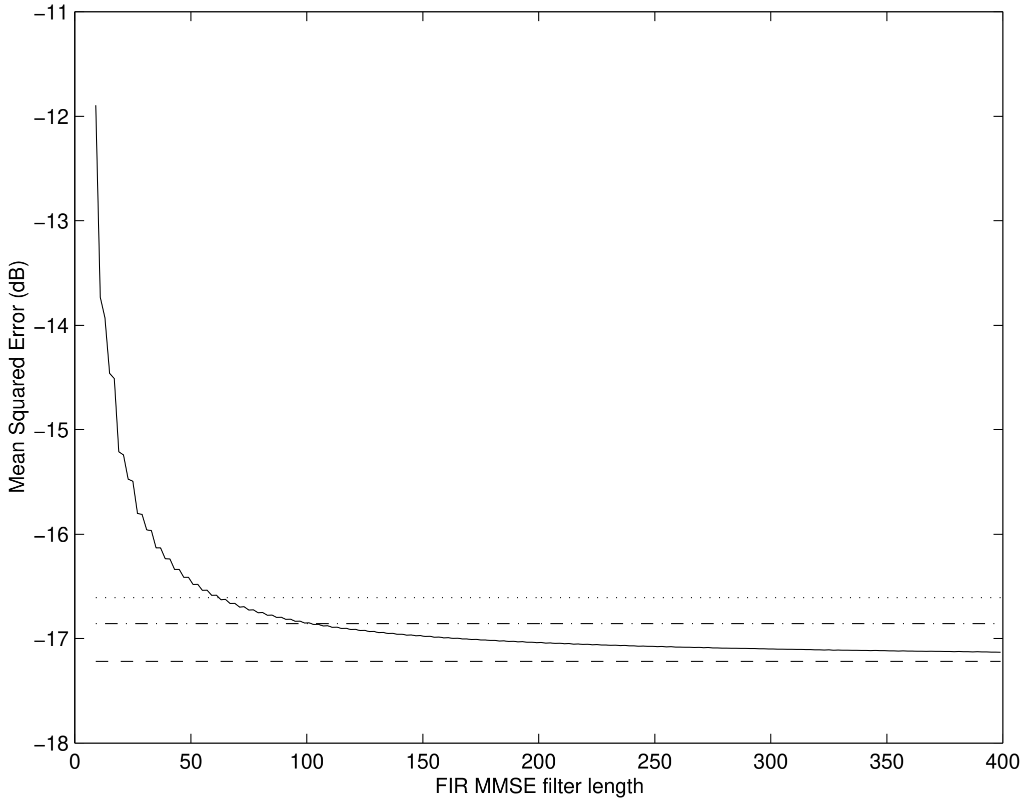

Figure 6: AR(1) process, $R=2.91$: convergence of MSE (solid curve) for FIR-MMSE filter banks to -17.22 dB limit (dashed line) for unconstrained-length MMSE filter banks. Compare with MSEs for IIR biorthogonal filter banks (dash-dotted line, -16.86 dB), and unconstrained-length FIR biorthogonal filter banks (dotted line, -16.61 dB).

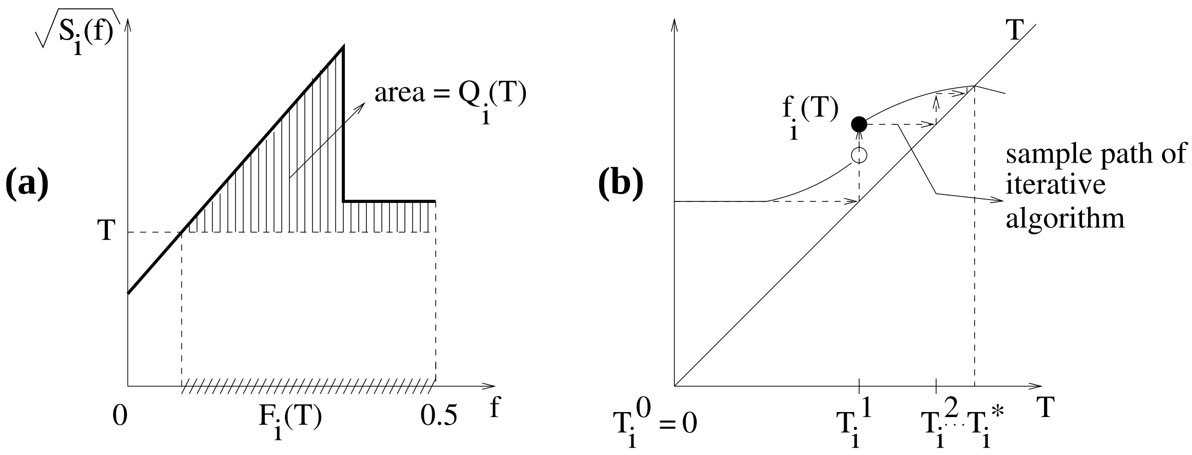

Figure 7: Computation of the threshold T_i^* that determines the set $\mathcal{F}_i^+$. (a) Square root of spectral density in channel i. The set $F_i(T)$ and the function $Q_i(T)$ are shown for a given value of the level T. (b) Function $f_i(T)$, and sample path of iterative algorithm converging to T_i^*, the solution to the equation $T = f_i(T)$.