JOINTLY OPTIMAL ANALYSIS AND SYNTHESIS FILTER BANKS FOR BIT CONSTRAINED SOURCE CODING

Are Hjørungnes and Tor A. Ramstad

Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106 USA
E-mail: {are, tor}@iplab.ece.ucsb.edu

ABSTRACT

A subband coder structure is fully optimized with respect to the minimum block mean squared error between the output and the input signals under a bit constraint. The analysis filter bank structure generates maximally decimated and equal bandwidth subbands. The subband quantizers are modeled as additive noise sources. To simplify the optimization an optimal multiple-input multiple-output system is first derived. Illustrative examples showing the system performance as well as filter transfer functions are given. The performance results are compared to the rate distortion curves.

1. INTRODUCTION

The optimization of low-rate subband coders suffers from two deficiencies. Firstly, the commonly used model for the quantizer is not strictly valid. Secondly, when the quantizer noise becomes sufficiently high, the perfect reconstruction property often assigned to the filter banks, is not longer optimal. Some efforts have been undertaken to overcome these obstacles.

A simple improvement of the quantizer model results when assuming that it does not only add noise to the signal, but also modifies the signal strength [1]. More accurate results can be obtained by using the true performance curve of the quantizer. Even that is not sufficient at low rates, because the assumption that the noise and signal are uncorrelated is far from exact. The traditional quantizer model will be used here even though this is not consistent with our low rate assumption.

What we address here is a joint optimization of the analysis and synthesis filter banks, for all rates, assuming the simple quantizer model is valid. A further refinement would result if more advanced quantizer models were applied. Some of the earlier works in this field include [2], where a unitary filter bank with perfect reconstruction is employed, and combined with pre- and post-filters, in [3] the analysis filters are kept constant while optimizing the bit allocation and the synthesis filters, and [4], where the total noise is minimized while maintaining perfect reconstruction through the use of inverse filters in the analysis and synthesis filter banks. In this paper we relax all the constraints from the cited articles, and minimize the overall output noise when the number of available bits is constrained.

The only model constraint adhered to in this paper is that all the subband channels occupy the same bandwidth and that the passband to baseband modulation in the analysis filter bank is obtained by integer factor decimation. The last requirement is enforced from an implementation point of view, although the results cited in this paper will be purely theoretical and give an upper bound for the performance of the system considered. The loss by using practical filters and a more realistic quantizer model are topics for an upcoming article.

2. PROBLEM FORMULATION

The system considered in this article is the subband coder model shown in Figure 1.

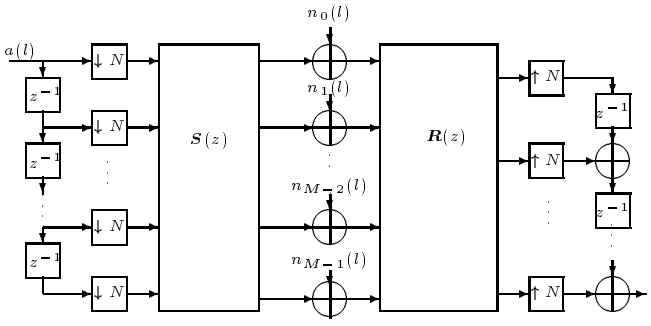

Figure 1: The subband coder considered.

The filter banks are implemented in polyphase form and the quantizers are replaced by additive noise sources.

Since the output of the subband coder system shown in Figure 1 is cyclostationary [5], it is difficult to analyze. By studying a corresponding vector system instead, the problem will become more manageable. The vector system is shown in Figure 2.

This work was supported by the Research Council of Norway.
The authors are on leave from Department of Telecommunications, Norwegian University of Science and Technology (NTNU), O.S. Bragstads plass 2B, N-7034 Trondheim, Norway.

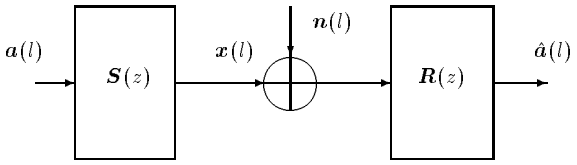

Figure 2: A block version of the subband coder.

The vector system's input signal $\boldsymbol{a}(l)$, is obtained by multiplexing the input time series $a(l)$ into

$$\boldsymbol{a}(l) = [a(lN), a(lN-1), \ldots, a(lN-(N-1))]^T. \quad (1)$$

The subband signals in Figure 2 are represented by the M-ary vector $\boldsymbol{x}(l) = [x_0(l), x_1(l), \ldots, x_{M-1}(l)]^T$ given by

$$\boldsymbol{x}(l) = \boldsymbol{s}(l) * \boldsymbol{a}(l), \quad (2)$$

where $*$ is the convolution operator and $\boldsymbol{s}(l)$ is the impulse response matrix of the analysis polyphase filter bank.

The quantizers are modeled as additive noise sources expressed in the following noise vector:

$$\boldsymbol{n}(l) = [n_0(l), n_1(l), \ldots, n_{M-1}(l)]^T. \quad (3)$$

The further assumptions made are the following: The vector time series $\boldsymbol{a}(l)$ and $\boldsymbol{n}(l)$ represent jointly wide sense stationary (WSS) vector series [5], which are mutually uncorrelated and have zero mean. The filters are linear, and are allowed to be non-causal with infinite impulse responses.

The partial statistical description of the signals $\boldsymbol{a}(l)$ and $\boldsymbol{n}(l)$ used here are given by the following power spectral density (PSD) matrices, respectively:

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{a}}(f) &= \sum_m E\left[\boldsymbol{a}(l+m)\boldsymbol{a}^H(l)\right] e^{-j2\pi fm},\\ \boldsymbol{\Sigma}_{\boldsymbol{n}}(f) &= \sum_m E\left[\boldsymbol{n}(l+m)\boldsymbol{n}^H(l)\right] e^{-j2\pi fm}.\end{aligned} \quad (4)$$

By using the inverse Fourier transform of the PSD matrices in Equation (4), the matrices $E\left[\boldsymbol{a}(l+m)\boldsymbol{a}^H(l)\right]$ and $E\left[\boldsymbol{n}(l+m)\boldsymbol{n}^H(l)\right]$ can be obtained. For high rates it is assumed that the quantizer noise can be modeled as white additive noise, and quantizer noise in one channel is uncorrelated with noise in all the other channels [6]. Therefore $\boldsymbol{\Sigma}_{\boldsymbol{n}}(f)$ is a diagonal matrix.

The Block mean squared error (MSE), used as the optimization criterion, is defined as

$$\epsilon_{N,M}(\gamma) = \mathrm{tr}\left(E\left[\boldsymbol{e}(l)\boldsymbol{e}^H(l)\right]\right), \quad (5)$$

where H, $\mathrm{tr}(\cdot)$, and $E[\cdot]$ are the Hermitian, trace, and expectation operators, respectively. Furthermore, $\boldsymbol{e}(l)$ is the error vector between the input and output vectors in Figure 2 given by

$$\begin{aligned}\boldsymbol{e}(l) =& \hat{\boldsymbol{a}}(l) - \boldsymbol{a}(l)\\ =& \sum_{m=-\infty}^{\infty} \boldsymbol{w}(l-m)\boldsymbol{a}(m) - \boldsymbol{a}(l)\\ &+ \sum_{p=-\infty}^{\infty} \boldsymbol{r}(l-p)\boldsymbol{n}(p),\end{aligned} \quad (6)$$

where $\boldsymbol{w}(l)$ is the impulse response matrix for the total system given by

$$\boldsymbol{w}(l) = \boldsymbol{r}(l) * \boldsymbol{s}(l). \quad (7)$$

Inserting $\boldsymbol{w}(l)$ and $\boldsymbol{e}(l)$ from Equations (7) and (6), respectively, into Equation (5), followed by manipulations, the following expression for the MSE described in the frequency domain is obtained:

$$\begin{aligned}\epsilon_{N,M}(\gamma) = \mathrm{tr}\Bigg(\int_{-\frac{1}{2}}^{\frac{1}{2}} \{&[\boldsymbol{I} - \boldsymbol{R}(f)\boldsymbol{S}(f)]\,\boldsymbol{\Sigma}_{\boldsymbol{a}}(f)\cdot\\ &[\boldsymbol{I} - \boldsymbol{R}(f)\boldsymbol{S}(f)]^H\\ &+\boldsymbol{R}(f)\boldsymbol{\Sigma}_{\boldsymbol{n}}(f)\boldsymbol{R}^H(f)\Big\}\, df\Bigg).\end{aligned} \quad (8)$$

$\boldsymbol{S}(f)$ and $\boldsymbol{R}(f)$ are the Fourier transform of the polyphase analysis and synthesis impulse response matrices, respectively.

The high rate assumptions allows the variance of the noise of quantizer number i, $\sigma^2_{n_i}$, to be modeled as

$$\sigma^2_{n_i} = \sigma^2_{x_i} h_i 2^{-2B_i}, \quad (9)$$

where $\sigma^2_{x_i}$ is the variance of subband signal number $i \in \{0, 1, \ldots, M-1\}$ (see Figure 2), B_i is the number of bits used in quantizer number i, and h_i is the quantizer performance factor [1]. Inversion of Equation (9), while asserting a non-negative bit count, gives

$$B_i = \max\left\{0, \frac{1}{2\ln 2}\ln\left(\frac{h_i\sigma^2_{x_i}}{\sigma^2_{n_i}}\right)\right\}. \quad (10)$$

From Equation (10) it is seen that the rate in quantizer number i is only dependent of the ratio between $\sigma^2_{x_i}$ and $\sigma^2_{n_i}$. Therefore the values of $\sigma^2_{n_i}$ can be chosen to an arbitrary value without loss of generality. By choosing the value of the quantizer noise variances equal to one, the filters will adjust to an appropriate value to give the minimum MSE overall performance. The optimal number of bits used in the quantizers will therefore be implicitly distributed through the system optimization.

It can be shown that the bit constraint can be expressed as

$$\Pr\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{S}(f)\boldsymbol{\Sigma}_{\boldsymbol{a}}(f)\boldsymbol{S}^H(f)\, df\right) \le B \cdot \prod_{i=0}^{M-1} \frac{\sigma^2_{n_i}}{h_i} = B', \quad (11)$$

where the operator Pr multiply the elements of the main diagonal of the matrix, and $B = \sum_{i=0}^{M-1} B_i$ is the total number of bits used on N source samples by the M quantizers. In Equation (11) the constant B' is also defined.

$\epsilon_{N,M}(\gamma)$ is a monotonic non-increasing function of the total number of bits used, B, which can be found through Equation (11). Therefore the constraint is assumed to be satisfied with equality without loss of generality. It is assumed that it is possible to use any positive real number of bits in the quantizers.

3. PROBLEM SOLUTION

The objective is to minimize the MSE given by Equation (8) subject to the bit constraint given by Equation (11) with respect to $\boldsymbol{S}(f)$ and $\boldsymbol{R}(f)$. It is not enough space to give the derivation of the optimal solution, so it will just be stated. A complete derivation will be presented in an future article.

The optimal transmitter and receiver matrices are given by

$$\begin{aligned} \boldsymbol{S}(f) &= \boldsymbol{V}(f)\boldsymbol{F}(f)\boldsymbol{U}^H(f), \\ \boldsymbol{R}(f) &= \boldsymbol{U}(f)\boldsymbol{G}(f)\boldsymbol{V}^H(f)\boldsymbol{\Lambda}_{\boldsymbol{n}}^{-1}(f). \end{aligned} \tag{12}$$

$\boldsymbol{U}(f)$ and $\boldsymbol{V}(f)$ are unitary matrices which diagonalize the input PSD matrix $\boldsymbol{\Sigma}_{\boldsymbol{a}}(f)$ and the Hermitian matrix $\boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1}(f)$, respectively, i.e.,

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{a}}(f)\boldsymbol{U}(f) &= \boldsymbol{U}(f)\boldsymbol{K}_{\boldsymbol{a}}(f), \\ \boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1}(f)\boldsymbol{V}(f) &= \boldsymbol{V}(f)\boldsymbol{\Lambda}_{\boldsymbol{n}}^{-1}(f). \end{aligned} \tag{13}$$

In Equation (13), $\boldsymbol{K}_{\boldsymbol{a}}(f)$ and $\boldsymbol{\Lambda}_{\boldsymbol{n}}^{-1}(f)$ are diagonal matrices that contain the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{a}}(f)$ and $\boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1}(f)$, respectively. In addition the elements of $\boldsymbol{K}_{\boldsymbol{a}}(f)$ and $\boldsymbol{\Lambda}_{\boldsymbol{n}}(f)$ are ordered as follows:

$$\begin{aligned} \kappa_0^{(N)}(f) &\geq \kappa_1^{(N)}(f) \geq \cdots \geq \kappa_{N-1}^{(N)}(f), \\ \lambda_0^{(M)}(f) &\leq \lambda_1^{(M)}(f) \leq \cdots \leq \lambda_{M-1}^{(M)}(f). \end{aligned} \tag{14}$$

The matrix $\boldsymbol{F}(f)$ is an $M \times N$ diagonal matrix where the magnitude of the diagonal elements are given by the square root of

$$\begin{aligned} |F_{i,i}(f)|^2 = \max\left(0, \alpha_i \sqrt{\frac{\lambda_i^{(M)}(f)}{\mu\kappa_i^{(N)}(f)}} - \frac{\lambda_i^{(M)}(f)}{\kappa_i^{(N)}(f)}\right), \\ i \in \{0, 1, \ldots, \min(M,N) - 1\}, \end{aligned} \tag{15}$$

where μ is a Lagrange multiplier for the constrained optimization problem, and α_i is a scaling factor which can be found by solving the following implicit equation:

$$\begin{aligned} \alpha_i = \sqrt{\int_{-\frac{1}{2}}^{\frac{1}{2}} \max\left[0, \alpha_i \sqrt{\frac{\lambda_i^{(M)}(f)\kappa_i^{(N)}(f)}{\mu}} - \lambda_i^{(M)}(f)\right]} \\ i \in \{0, 1, \ldots, \min(M,N) - 1]\}. \end{aligned} \tag{16}$$

The phase of the elements in $\boldsymbol{F}(f)$, see Equation (15), can be chosen arbitrarily.

The matrix $\boldsymbol{G}(f)$ in Equation (12) can be expressed by

$$\begin{aligned} \boldsymbol{G}(f) =& \boldsymbol{K}_{\boldsymbol{a}}(f)\boldsymbol{F}^H(f) \cdot \\ & \left[\boldsymbol{F}(f)\boldsymbol{K}_{\boldsymbol{a}}(f)\boldsymbol{F}^H(f) + \boldsymbol{\Lambda}_{\boldsymbol{n}}(f)\right]^{-1} \boldsymbol{\Lambda}_{\boldsymbol{n}}(f). \end{aligned} \tag{17}$$

The performance of the optimal system measured by block MSE, see Equation (5), is found to be

$$\epsilon_{N,M}(\gamma) = \begin{cases} \displaystyle\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=0}^{N-1} \frac{\kappa_i^{(N)}(f)\lambda_i^{(M)}(f)}{|F_{i,i}(f)|^2\kappa_i^{(N)}(f) + \lambda_i^{(M)}(f)}\, df, \\ \qquad\qquad \text{if } M \geq N, \\ \displaystyle\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=0}^{M-1} \frac{\kappa_i^{(N)}(f)\lambda_i^{(M)}(f)}{|F_{i,i}(f)|^2\kappa_i^{(N)}(f) + \lambda_i^{(M)}(f)}\, df \\ \qquad + \displaystyle\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=M}^{N-1} \kappa_i^{(N)}(f)\, df, \\ \qquad\qquad \text{if } M < N. \end{cases} \tag{18}$$

The constraint on bit per vector used can be expressed in the following way:

$$\sum_{i=0}^{\min(M,N)-1} \ln \int_{-\frac{1}{2}}^{\frac{1}{2}} |F_{i,i}(f)|^2 \kappa_i^{(N)}(f)\, df \leq B', \tag{19}$$

where B' is the constant defined in Equation (11).

For a given target bit-rate B in Equation (19), the objective is to find the Lagrangian multiplier μ, given implicitly through Equations (15) and (16), for equality in Equation (19). This value of μ is then inserted in Equation (18) to calculate the block MSE.

4. ILLUSTRATIVE EXAMPLES

As an example consider the case where the filter bank is maximally decimated, i.e., $N = 2$ / $M = 2$. The source which is coded is a Gaussian AR(3)-process for which the power spectral density is given in Figure 3 (a).

By using the noble identities the decimators in Figure 1 can be moved behind the polyphase matrix on the transmitter side. The frequency responses of the filters can then be found by taking the delay chain in front of the polyphase matrix into account. Notice from Figure 3 (b) that both filters extract two frequency bands. In parts (b) and (c) of the figures it is seen that only one of the filters is different from zero at any frequency. It can be shown that there will be no overlap in the frequency domain after the filters have been decimated. From Figure 3 (d) it is seen that the optimal filter banks do not have the perfect reconstruction property. The phase of the analysis filters is arbitrary and therefore can chosen to be linear. Then the synthesis filters will also have linear phase. It is also observed from the figure that in frequency intervals where the input signal has low energy no signal is sent through the system.

Figure 4 depicts the calculated block MSE (in dB) for the system with $N = 3$ as a function of rate when 3, 2 or 1 of the bands are quantized. From this figure it is seen that as the bit-rate is decreased the number of channels should be reduced. For example at 0.5 bit/sample only one filter should be used. In Figure 4 the performance of $N = 3$ / $M = 3$ will be the same as for the $N = 3$ / $M = 2$ case for

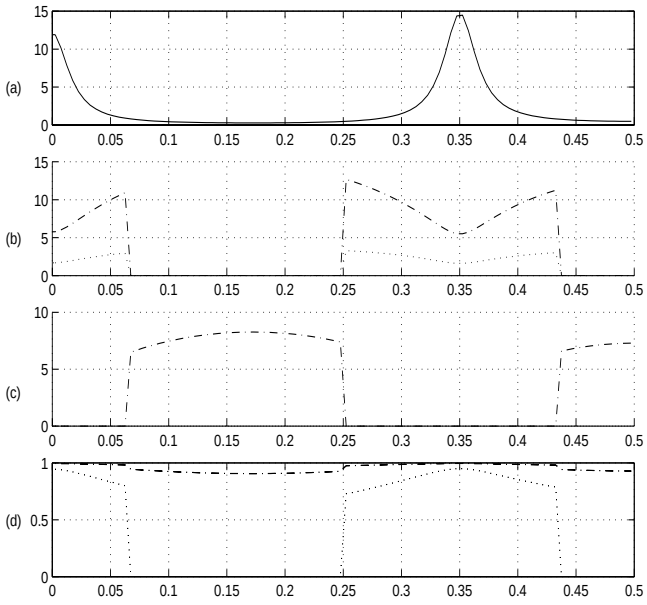

Figure 3: System example using $N = 2$ / $M = 2$. (a) PSD of the input signal. (b) Frequency response of the first channel on the transmitter side. (c) Frequency response of the second channel on the transmitter side. (d) Total frequency response through the system. In the last three plots ($-\cdot-\cdot-$) gives the result with $B/N = 3.28$ bits/sample, and SNR = 18.4 dB, while ($\cdots\cdots$) shows the result with $B/N = 1.12$ bits/sample, and SNR = 7.51 dB.

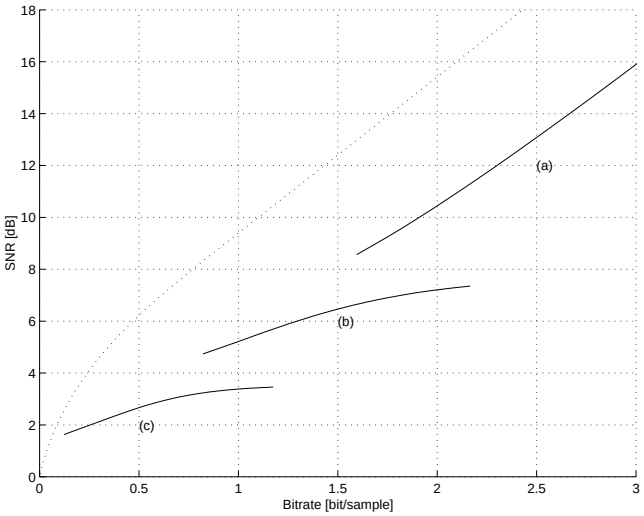

Figure 4: Performance of the system using $N = 3$ / $M = 3$ (a), $N = 3$ / $M = 2$ (b), and $N = 3$ / $M = 1$ (c). The PSD of the source to be coded is shown in Figure 3 (a). The upper curve is the rate distortion function.

rates below 1.3 bits/sample, and as $N = 3$ /$M = 1$ below 0.6 bits/sample. The reason for not showing all curves for all rates is that the quantizer model only is valid for high rates. The results which are obtained in the missing rate regions are very inaccurate because the assumptions made are not valid in these regions.

The high rate quantizer model underestimates the actual performance at high rates while at low rates the results give too good performance [1].

5. CONCLUSIONS

A jointly optimal analysis and synthesis filter bank and bit allocation is developed. The results show that the filter banks strongly depend on the power spectral density of the signal which is coded. The traditional way of using bandpass filters in each channel with only one contiguous passband in each subband is suboptimal for some PSDs. From the theory developed in this article it is seen that the frequency response of one filter of the minimum block MSE filter banks may have more than one passband. Another implication of the theory presented, is that when over-complete bases are considered with a linear system, i.e., $M > N$, $M - N$ of the filters will be set to zero. With the assumption made in this article there is nothing to gain by using over-complete bases. With bandwidth reduction, i.e. $M < N$, the performance of the linear system will be poor for high rates because perfect reconstruction is impossible when not full ranked matrices are used in linear systems. But by constraining the number of subbands to be less than the decimating factor the complexity of the system is reduced, and at low bit-rates the performance is the same as using $N = M$, because some of the quantizers are allocated zero bits anyway.

6. REFERENCES

[1] N. S. Jayant and P. Noll, Digital Coding of Waveforms, Principles and Applications to Speech and Video. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1984.

[2] P. Vaidyanathan and T. Chen, "Statistically optimal synthesis banks for subband coders," in Twenty-Eighth Asilomar Conference on Signals, Systems and Computers, 1994.

[3] K. Gosse, F. Moreau de Saint-Martin, and P. Duhamel, "Filter bank design for minimum distortion in presence of subband quantization," in Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP), pp. 1491–1494, 1996.

[4] K. C. Aas and C. T. Mullis, "Minimum mean-squared error transform coding and subband coding," IEEE Trans. Inform. Theory, vol. 42, pp. 1179–1192, July 1996.

[5] V. Sathe and P. P. Vaidyanathan, "Effects of multirate systems on the statistical properties of random signals," IEEE Trans. Signal Processing, vol. 41, pp. 131–146, Jan. 1991.

[6] A. Gersho and R. M. Gray, Vector Quantization and Signal Compression. Boston, MA, USA: Kluwer Academic Publishers, 1992.