Design of FIR Digital Filters with Discrete Coefficients via Convex Relaxation

Wu-Sheng Lu
Dept. of Electrical & Computer Eng.
University of Victoria
Victoria, BC, Canada V8W 3P6
Email: wslu@ece.uvic.ca

***Abstract*— Digital filters with discrete coefficients that can be expressed as sums of power of two (SP2) are of practical use because they admit fast implementations that do not require multiplications. In this paper, a new method for the design of finite-impulse-response (FIR) digital filters with SP2 coefficients by convex relaxation is proposed. The major difference of the proposed method from the semidefinite programming relaxation (SDPR) method proposed in the literature is that a sequential convex quadratic programming relaxation (QPR) in conjunction with a low-bit descent search technique replaces SDPR, yielding much reduced algorithmic complexity. Design examples are presented to illustrate the proposed algorithm and to demonstrate its near optimal performance against a weight least squares error measure.**

I. Introduction

Digital filters with discrete coefficients are of practical use because they admit fast implementations that do not require multiplications but superposition of shifted versions of the input. Throughout the term "discrete coefficient" is referred to a real value that can be expressed as sum of finite power-of-two terms. For brevity we call it a SP2 (sum of power of two) coefficient. There has been considerable research interest in the design of digital filters with SP2 coefficients in the past [1]–[7]. A technical difficulty encountered in designing an optimum filter with SP2 coefficients is its exponential complexity. This is because the design problem is essentially an integer programming (IP) problem which is known to be NP-hard [8]. In [7], a semidefinite programming (SDP) relaxation approach is proposed, in which the IP problem involved is "relaxed" to an SDP problem that is a convex optimization problem and is solvable with polynomial complexity. Design practice has indicated that the suboptimal solutions obtained using SDP relaxation (SDPR) for a variety of filter types and filter lengths are of excellent quality [7]. A problem with the SDPR-based method is its complexity. Although in theory the solution method is of polynomial complexity and is indeed considerably more efficient than IP-based design algorithms, the amount of computation remains to be more than affordable for high-order filters.

In this paper, the design problem is addressed using a new relaxation method that differs from the SDPR method in a major design step where a $\{-1,\ 1\}$-optimization problem is relaxed to a sequence of quadratic programming (QP) problems which can be solved substantially more efficiently than their SDP counterparts, especially when the order of the filter is high. Design examples are presented to illustrate and evaluate the proposed algorithm as compared with several existing solutions including the one obtained using the SDP relaxation.

II. A New Design Method Based on Convex Relaxation

Four steps are involved in the proposed design method: (i) design of an FIR filter with continuous coefficients that approximates a desired frequency response in a cetrain optimal sense; (ii) a subsequent formulation of the design of an FIR filter with SP2 coefficients as a $\{-1,\ 1\}$-optimization problem; (iii) a sequential convex QP relaxation of the $\{-1,\ 1\}$-optimization problem, and (iv) solution enhancement by low-bit descent search. In what follows we describe these steps in order.

A. Design of an FIR Filter with Continuous Coefficients

There are many methods for the design of this type of filters [9]. For illustration purposes, suppose we apply one of the available methods to design a linear-phase FIR filter of odd length that approximates a desired frequency response $H_d(\omega)$ such that the weighted least square (WLS) error

$$e = \int_0^\pi W(\omega)|H(e^{j\omega}) - H_d(\omega)|^2\, d\omega \tag{1}$$

is minimized, and the FIR filter obtained is represented by

$$H_c(z) = \sum_{k=0}^{N-1} h_k z^{-k} \tag{2}$$

B. A Weighted Least-Square $\{-1,\ 1\}$-Optimization Problem

Instead of $H_c(z)$ in (2), we are interested in an FIR transfer function

$$H(z) = \sum_{k=0}^{N-1} d_k z^{-k} \tag{3}$$

where each coefficient d_k is a length-L binary number in two's complement representation [9], i.e.,

$$d_k = -\beta_{k0} + \sum_{i=1}^{L} \beta_{ki} 2^{-i} \tag{4}$$

with $\beta_{ki} \in \{0,\ 1\}$ for $i = 0,\ 1,\ \ldots,\ L$. For a given filter length N and length budget L, we seek to determine SP2 coefficients $\{d_k,\ k = 0,\ \ldots,\ N-1\}$ such that the WLS error in (1) is minimized. This is a *discrete* QP problem with a total of $N(L+1)$ binary variables. Even for a moderate filter length N in the range [41, 81] and length budget L in the range [8, 16], the number of binary variables can easily exceed 500, and the computational complexity of solving a *discrete* QP problem of such a size is very high.

Our approach to the problem is to place a *SP2 layer* surrounding the optimal continuous coefficients (i.e., h_k's in (2)) and formulate a reduced-size $\{-1,\ 1\}$-optimization problem whose solution is obtained by optimally choosing the SP2 coefficients *within* the layer. Given a real-valued coefficient h_k and length budget L, we find the largest SP2 lower bound of h_k, denoted by $\underline{d}_k$, and the smallest SP2 upper bound of h_k, denoted by $\bar{d}_k$. The relation of coefficient h_k to these bounds is illustrated in Fig. 1,

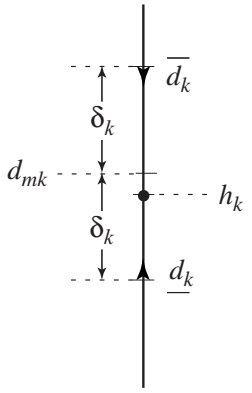

Fig. 1 Relation of h_k to the bounds.

where

$$d_{mk} = (\bar{d}_k + \underline{d}_k)/2,\ \ \delta_k = (\bar{d}_k - \underline{d}_k)/2 \tag{5}$$

It can be readily verified that

$$d_k = d_{mk} + \delta_k b_k \tag{6}$$

assumes the values of $\underline{d}_k$ or $\bar{d}_k$ when b_k assumes the values of -1 or 1, respectively. It is generally expected that the impulse response of the optimal FIR filter with SP2 coefficients should lie in a small vicinity of the impulse response of the optimal FIR filter with continuous coefficients. From Fig. 1 and (6), we see that a SP2 layer has been placed on each side of $\{h_k\}$ and, for each k, any of the two SP2 bounds can be selected by specifying the binary variable b_k. In this way, the problem at hand becomes an $\{-1,\ 1\}$-optimization problem with N binary variables. More layers can be placed if necessary, at the cost of dealing with mN binary variables for the m-layer case.

For the design of linear-phase FIR filters of odd length N, $\{h_k\}$, $\{d_{mk}\}$ and $\{b_k\}$ are all symmetrical with respect to the midpoint, hence (3) and (6) imply

$$H(e^{j\omega}) = e^{-jN_1\omega}[A_m(\omega) + \boldsymbol{b}^T\boldsymbol{c}(\omega)] \tag{7}$$

where $N_1 = (N-1)/2$, and $A_m(\omega)$ is the real-valued trigonometric polynomial obtained from

$$\sum_{k=0}^{N-1} d_{mk} e^{-jk\omega} = e^{-jN_1\omega} A_m(\omega)$$

In (7), vectors $\boldsymbol{b}$ and $\boldsymbol{c}(\omega)$ are defined by

$$\begin{aligned} \boldsymbol{b} &= [b_0 \ \ b_1 \ \ \cdots \ \ b_{N1}]^T \\ \boldsymbol{c}(\omega) &= [2\delta_0 \cos N_1\omega \ \ 2\delta_1 \cos(N_1-1)\omega \ \ \cdots \ \ \delta_{N_1}]^T \end{aligned}$$

Under these circumstances, the WLS error in (1) can be evaluated as

$$e = \boldsymbol{b}^T\boldsymbol{Q}\boldsymbol{b} + 2\boldsymbol{b}^T\boldsymbol{q} + \text{const}$$

where

$$\boldsymbol{Q} = \int_0^\pi W(\omega)\boldsymbol{c}(\omega)\boldsymbol{c}^T(\omega)\, d\omega \tag{8a}$$

$$\boldsymbol{q} = \int_0^\pi W(\omega)[A_m(\omega) - A_d(\omega)]\boldsymbol{c}(\omega)\, d\omega \tag{8b}$$

In (8b), $A_d(\omega)$ is obtained from the desired frequency response $H_d(\omega) = e^{-jN_1\omega}A_d(\omega)$. Consequently, the weighted least-square design of $H(z)$ with SP2 coefficients can be formulated as the $\{-1,\ 1\}$-optimization problem

$$\underset{b_i\in\{-1,1\}}{\text{minimize}} \quad \boldsymbol{b}^T\boldsymbol{Q}\boldsymbol{b} + 2\boldsymbol{b}^T\boldsymbol{q} \tag{9}$$

C. A Sequential Convex Relaxation of Problem (9)

If vector $\boldsymbol{b}$ in (9) is treated as a *continuous* variable vector, then the objective function in (9) is strictly convex because matrix $\boldsymbol{Q}$ defined by (8a) is positive definite. Therefore, if the binary constraints $b_i \in \{-1,\ 1\}$ are *relaxed* to $-1 \le b_i \le 1$ then the problem in (9) is relaxed to

$$\text{minimize} \qquad \boldsymbol{b}^T\boldsymbol{Q}\boldsymbol{b} + 2\boldsymbol{b}^T\boldsymbol{q} \tag{10a}$$

$$\text{subject to:} \quad -1 \le b_i \le 1 \qquad \text{for } i = 1, \ldots,\ N_1 \tag{10b}$$

which is obviously a *convex* QP problem that admits a unique global solution [10]. The solution of (10) is however not binary in general. At this point, an intuitively meaningful step toward a binary solution is to use a multistage strategy that sets the components of $\boldsymbol{b}$, whose counterparts in the corresponding continuous-valued solution are sufficiently close to either 1 or -1, to 1 or -1. Here the closeness may be measured by a pre-specified threshold $\alpha > 0$, i.e.

$$b_i = \begin{cases} 1 \text{ if } b_i > \alpha \\ -1 \text{ if } b_i < -\alpha \\ \text{to be determined in subsequent iterations} \end{cases}$$

The determined components of $\boldsymbol{b}$ are then substituted into (10), yielding a convex QP problem of reduced size and the step described above applies again to its solution. This process continues until it reaches the last stage of the design where all remaining components of $\boldsymbol{b}$ are set to 1 or -1 based on their signs. The value of threshold α may vary from stage to stage, but in many cases a constant α in the vicinity of 0.5 yields satisfactory designs.

D. Solution Enhancement by Low-Bit Descent Search

The solution obtained can be enhanced by low-bit descent search where each time only a small number of components in vector $\boldsymbol{b}$ are switched and the performance of the modified $\boldsymbol{b}$ is then evaluated. In this section we show that one-bit and two-bit descent search can be accomplished with a very small amount of computation.

1) A One-Bit Descent Search

Denote $f(\boldsymbol{b}) = \boldsymbol{b}^T\boldsymbol{Q}\boldsymbol{b} + 2\boldsymbol{q}^T\boldsymbol{b}$. Switching the sign for the ith component of $\boldsymbol{b}$ can be described as using $\boldsymbol{b}_i = \boldsymbol{b} - 2b_i\boldsymbol{e}_i$ to replace $\boldsymbol{b}$, where $\boldsymbol{e}_i$ is the ith column of the identity matrix. Using the fact that b_i^2 is always equal to one, the change in the objective function due to the above one-bit switch is found to be

$$\begin{aligned}\delta_i &= f(\boldsymbol{b}_i) - f(\boldsymbol{b}) \\ &= 4[q_{ii} - b_i(\boldsymbol{q}_i^T\boldsymbol{b} + q_i)] \\ &= -4b_i\left(\sum_{j=1,\, j\neq i}^{N_1+1} q_{ij}b_j + q_i\right)\end{aligned}$$

where $\boldsymbol{q}_i$ is the ith column of $\boldsymbol{Q}$, q_{ij} is the $(i,\, j)$-component of $\boldsymbol{Q}$ and q_i is the ith component of $\boldsymbol{q}$. Let $\hat{\boldsymbol{Q}}$ be the matrix obtained from $\boldsymbol{Q}$ by setting its diagonal components to zero and define vector $\boldsymbol{v} = \hat{\boldsymbol{Q}}\boldsymbol{b} + \boldsymbol{q}$, then we have $\delta_i = -4b_iv_i$ where v_i denotes the ith compnent of $\boldsymbol{v}$. Consequently, the one-bit descent search can be carried out by evaluating vector $\boldsymbol{b}\odot\boldsymbol{v}$ (here $\odot$ denotes component-wise multiplication), then identifying its index i^* where the associated component assumes the maximum value, and switching the sign of b_{i^*}.

2) A Two-Bit Descent Search

Switching the sign of two components in vector $\boldsymbol{b}$ can be described as replacing $\boldsymbol{b}$ with $\boldsymbol{b}_{ij} = \boldsymbol{b} - 2b_i\boldsymbol{e}_i - 2b_j\boldsymbol{e}_j$ where i and j are distinct. It can be readily verified that the change in the objective function due to this two-bit switch is given by

$$\begin{aligned}\delta_{ij} &= f(\boldsymbol{b}_{ij}) - f(\boldsymbol{b}) \\ &= -4[(b_iv_i + b_jv_j) - 2b_ib_jq_{ij}]\end{aligned}$$

where $i \neq j$ is assumed. If we define matrix $\boldsymbol{D} = \{\delta_{ij}\}$ whose diagonal is set to zero, then we can write $\boldsymbol{D} = -4\boldsymbol{P}$ with

$$\boldsymbol{P} = (\boldsymbol{b}\odot\boldsymbol{v})\boldsymbol{e}^T + \boldsymbol{e}(\boldsymbol{b}\odot\boldsymbol{v})^T - 2\boldsymbol{Q}\odot(\boldsymbol{b}\boldsymbol{b}^T)$$

where $\boldsymbol{e}$ denotes the all-one vector and the diagonal of $\boldsymbol{P}$ is set to zero. The two-bit descent search can be carried out by evaluating matrix $\boldsymbol{P}$, then identifying its indices $(i^*,\ j^*)$ where the associated component reaches the maximum value, and switching each of the signs of b_{i^*} and b_{j^*} to its opposite.

Because both one-bit and two-bit search are descent algorithms, each algorithm can be used alone. Better still, they can be used in parallel to yield an improved search result at a cost of increased complexity.

III. Design Examples

The QP relaxation (QPR) method described in Sec. 2 was applied to design a variety of linear phase FIR filters with SP2 coefficients. Here we present ten designs of lowpass FIR filters with nomalized passband edge $\omega_p = 0.2$ and stopband edge $\omega_a = 0.25$ of lengths $N = 7 + 8i$ for $i = 0,\ 1,\ \ldots,\ 9$. The wordlength L was 8 for the first five designs and L was 12 for the last five designs. The weighting function $W(\omega) \equiv 1$ in both passband and stopband and $W(\omega) \equiv 0$ elsewhere. In all designs, the proposed method was used with 2 stages and threshold $\alpha = 0.5$, and the two-bit descent search was employed. The weighted least squares error e in (1) was used to evaluate the filter performance, and the computational complexity was evaluated in terms of the CPU time consumed. The design results are shown in Table I where for a given N, each case has shown two numbers with the first being the WLS error e and second the CPU time in seconds. For comparison purposes, Table I also includes design results obtained using the SDPR method and the optimal filters obtained by exhaustive search (we were unable to perform the optimal designs for $N = 63,\ 71$, and 79 because the amount of computation required in those 3 cases was not affordable for a Pentium 4 PC). It is observed that the designs obtained by the proposed QPR method and by the SDPR method [7] are both near optimal in terms of WLS error, but the QPR algorithm required significantly less computation relative to the SDPR algorithm.

TABLE I
Comparison of Various Designs

N	QPR	SDPR	Optimal
7	0.0311	0.0311	0.0311
	0.0231	0.0469	0.0103
15	0.0060	0.0060	0.0060
	0.0235	0.3438	0.0156
23	0.0012	0.0012	0.0012
	0.0239	1.3281	1.2656
31	0.2291e−3	0.2291e-3	0.2291e−3
	0.0242	9.2500	11.7813
39	0.0735e−3	0.0735e-3	0.0735e−3
	0.0253	18.2188	26.8750
47	0.1122e−4	0.1120e−4	0.1119e−4
	0.0264	41.7969	231.1875
55	0.3456e−5	0.3481e−5	0.3456e−5
	0.0273	138.30	3799.80
63	0.1039e−5	0.1021e−5	–
	0.0284	210.22	–
71	0.0430e−5	0.0410e−5	–
	0.0301	416.80	–
79	0.2375e−6	0.2375e−6	–
	0.0326	723.67	–

The amplitude responses of the WLS optimal filter with continuous coefficients compared with QPR-based FIR filter with SP2 coefficients are depicted in Fig. 2a, while the comparison of the optimal filter with the SDPR-based filter are shown in Fig. 2b where, in both cases, $N = 55$ and $L = 12$.

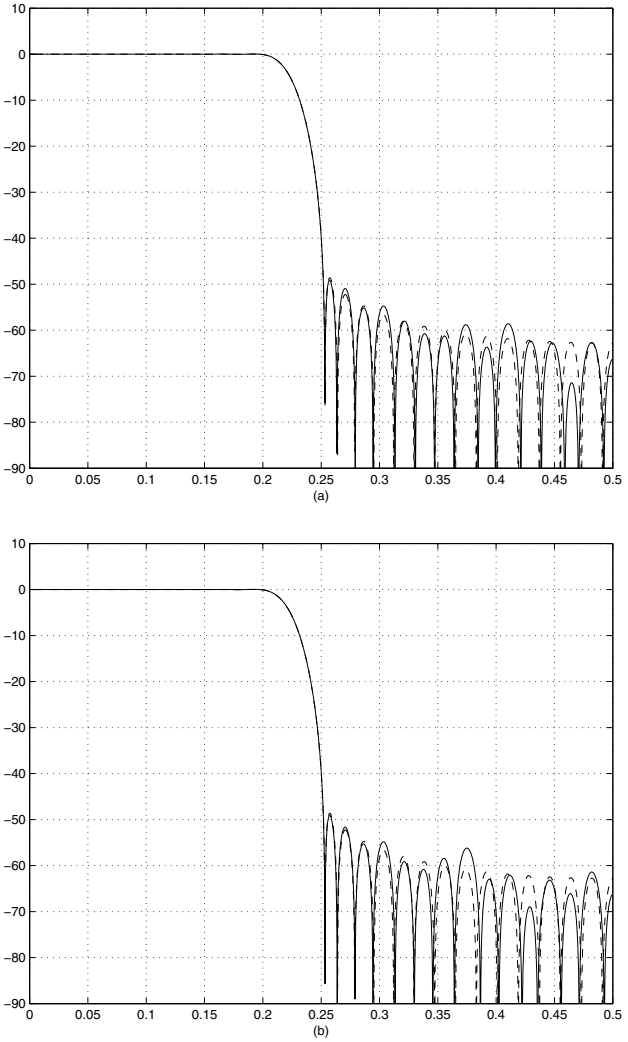

Fig. 2 (a) Amplitude responses of optimal FIR filter with continuous coefficients (dashed line) and QPR-based FIR filter with SP2 coefficients, (b) Amplitude responses of optimal filter with continuous coefficients (dashed line) and SDPR-based FIR filter with SP2 coefficients. In both cases $N = 55,\ L = 12$.

IV. Conclusion

We have described a new method for the design of FIR digital filters with SP2 coefficients using sequential convex QP relaxation in conjunction with low-bit descent search. Simulations have demonstrated that the proposed algorithm offers near optimal designs with a computatial complexity that is only a small fraction of the previously proposed design method based on SDPR.

Acknowledgement

The author is grateful to the Natural Sciences and Engineering Research Council of Canada (NSERC) for supporting this work.

References

[1] Y. C. Lim and S. R. Parker, "FIR filter design over a disrete power-or-two coefficient space," *IEEE Trans. ASSP*, vol. 31, pp. 588–591, June 1983.

[2] P. P. Vaidyanathan, "Efficient and multiplierless design of FIR filters with very sharp cutoff via maximally flat building blocks," *IEEE Trans. CAS*, vol. 32, pp. 236–244, March 1985.

[3] Y. C. Lim, "Design of discrete-coefficient-value linear phase FIR filters with optimum normalized peak ripple magnitude," *IEEE Trans. CAS*, vol. 37, pp. 1480-1486, Dec. 1990.

[4] Y. C. Lim, R. Yang, D. Li, and J. Song, "Signed power-or-two (SPT) term allocation scheme for the design of digital filters," *Proc. ISCAS*, Monterey, CA., June 1998.

[5] J. T. Yli-Kaakinen and T. A. Saramäki, "An algorithm for the design of multiplierless approximately linear-phase lattice wave digital filters," *Proc. ISCAS*, vol. II, pp. 77-80, Geneva, June 2000.

[6] Y. C. Lim and Y. J. Yu, "A successive reoptimization approach for the design of discrete coefficient perfect reconstruction lattice filter bank", *Proc. ISCAS*, vol. II, pp. 69–72, Geneva, June 2000.

[7] W.-S. Lu, "Design of FIR filters with discrete coefficients: A semidefinite programming relaxation approach," *Proc. ISCAS*, vol. 2, pp. 297–300, Sydney, Australia, May 2001.

[8] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization*, Prentice-Hall, 1982.

[9] A. Antoniou, *Digital Filters: Analysis, Design, and Applications*, 2nd ed., McGraw-Hill, 1993.

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.