An MILP Approach for the Design of Linear-Phase FIR Filters with Minimum Number of Signed-Power-of-Two Terms

Oscar Gustafsson, Håkan Johansson, and Lars Wanhammar*

Abstract — **In this paper a new approach for the design of linear phase FIR filters with signed-powers-of-two (SPT) coefficients is proposed. A mixed integer linear programming (MILP) problem is formulated that minimizes the number of SPT terms for a given filter specification. The method is applicable both for filters with a specified passband gain and filters where the normalized peak ripple magnitude is of interest. In both cases the globally minimal solution is found subject to the filter specification, filter order, and number of coefficient bits. A preprocessing method that removes 30-50% of the variables is also proposed and it is shown by an example that this method speeds up the optimization process significantly.**

1 Introduction

Finding good fixed-point coefficients for FIR filters has received considerable attention during the last 20 years. Many different techniques have been proposed. For a comprehensive list of references we refer to [1], but in this paper we utilize mixed-integer linear programming (MILP). This technique has been used earlier in [2]–[4], but there the objective was to minimize the ripple subject to hardware specifications (wordlength and filter order). In this paper we aim at minimizing the hardware cost (or more specifically, the number of signed-power-of-two terms) given a filter specification.

The measure of good coefficients depends on the filter implementation, but in this paper we focus on a minimum number of signed-power-of-two (SPT) terms for the coefficients [5]. This is motivated by that a multiplication can be separated into a number of shift-and-add operations. Each shift-and-add operation corresponds to a SPT term which makes this an interesting variable to minimize. However, as it is possible to use subexpression sharing [6], [7] or multiplier blocks [8], the resulting solution may not necessarily be optimal in terms of hardware complexity.

In this paper we formulate a MILP problem that minimizes the number of SPT terms for a given filter specification subject to the filter order and coefficient wordlength. A preprocessing technique that reduces the number of 0/1-variables in the optimization with 30-50% is also proposed. The proposed method has been modelled using AMPL [9] and solved using the general purpose optimization package CPLEX [10]. An example shows the usefulness of the proposed technique.

* Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, SWEDEN.
E-mail: {oscarg, hakanj, larsw}@isy.liu.se, Tel: +46-13-28{4059, 1676, 1344}, Fax: +46-13-139282.

2 Linear-Phase FIR Filters

The frequency response of a linear-phase FIR filter can be separated into a real-valued function $H_R(\omega T)$ and a real-valued phase function $\Theta(\omega T)$ as

$$H(e^{j\omega T}) = H_R(\omega T)e^{j\Theta(\omega T)} \tag{1}$$

where $H_R(\omega T)$ is the *zero-phase frequency response*. Clearly, we have $|H(e^{j\omega T})| = |H_R(\omega T)|$, as $|e^{jx}| = 1$ for real-valued x.

For an Nth-order linear-phase FIR filter the zero-phase frequency response can be written as

$$H_R(\omega T) = \sum_{m=1}^{M} h_m c(m, \omega T) \tag{2}$$

where for symmetric impulse response and even N we have

$$c(m, \omega T) = x_m \cos(\omega T[m-1]) \tag{3}$$

where

$$x_m = \begin{cases} 1, & m = 1 \\ 2, & m = 2, 3, \ldots, M \end{cases} \tag{4}$$

and

$$M = N/2 + 1 \tag{5}$$

while for odd N we have

$$c(m, \omega T) = 2\cos\left(\omega T\left[m - \frac{1}{2}\right]\right) \tag{6}$$

and

$$M = (N+1)/2 \tag{7}$$

Similar expressions can also be derived for antisymmetric impulse responses but will be left out here.

Let the specifications of the filter be

$$\begin{aligned} 1-\delta_c \le H_R(\omega T) \le 1+\delta_c, &\quad \omega T \in \Omega_c \\ -\delta_s \le H_R(\omega T) \le \delta_s, &\quad \omega T \in \Omega_s \end{aligned} \tag{8}$$

where δ_c and δ_s are the allowed ripples in the passband and stopband, respectively, and Ω_c and Ω_s are the passband and stopband, respectively.

3 Proposed Design Method

3.1 Signed-Powers-of-Two Coefficients

A fixed-point coefficient of wordlength B can be represented as a sum of SPT terms in the general form

$$h_m = \sum_{i=1}^{B} s_i 2^{-i} \tag{9}$$

where $s_i \in \{-1, 0, 1\}$. Here we assume $-1 < h_m < 1$.

A minimum representation refers to the representation with the minimum required number of SPT terms. One minimum representation is the canonic signed digit code (CSDC) representation. Here, no two SPT terms can be adjacent.

It is possible to transform the SPT representation in (9) to 0/1-variables as

$$h_m = \sum_{i=1}^{B} (a^+_{m,i} - a^-_{m,i}) 2^{-i} \tag{10}$$

where $a^+_{m,i} \in \{0, 1\}$ and $a^-_{m,i} \in \{0, 1\}$. This is advantageous as the formulation of the optimization goal function will be linear, otherwise it would have to be formulated as a non-linear programming problem. Furthermore, it will make it possible to have linear constraints on the number of SPT terms per coefficient.

3.2 Normalized Peak Ripple Magnitude

For many filter implementations the absolute value of the passband gain is of less importance. Instead the relative attenuation between the passband and stopband is of interest. This is referred to as normalized peak ripple magnitude (NPRM) [4]. Allowing an arbitrary passband gain will also reduce the required number of SPT-terms [4].

By introducing a passband gain scaling variable s and utilizing (2)-(7) we can rewrite the constraints in (8) as

$$\begin{aligned}
&\sum_{m=1}^{M} h_m c(m, \omega T) \le s(1 + \delta_c), \quad \omega T \in \Omega_c \\
&-\sum_{m=1}^{M} h_m c(m, \omega T) \le s(\delta_c - 1), \quad \omega T \in \Omega_c \\
&\sum_{m=1}^{M} h_m c(m, \omega T) \le s\delta_s, \quad \omega T \in \Omega_s \\
&-\sum_{m=1}^{M} h_m c(m, \omega T) \le s\delta_s, \quad \omega T \in \Omega_s
\end{aligned} \tag{11}$$

3.3 Problem Formulation

The resulting optimization problem is formulated in (15). As the constraints are continuous in ωT it is necessary formulate the constraint for discrete values of ωT. This is easily done by selecting a number of values in Ω_c and Ω_s to obtain a grid of values which each leads to one constraint. As the specifications are checked only in these points it is necessary to check the resulting transfer function with a much finer grid to verify that a valid coefficient set is obtained. If not, additional points must be added to the constraints.

The solution to this optimization problem yields a coefficient set that is minimal in terms of SPT terms when the NPRM is considered. The solution may not be in CSDC form, but it will be minimal. To obtain a CSDC solution the following constraint can be added

$$a^+_{m,i} + a^-_{m,i} + a^+_{m,i+1} + a^-_{m,i+1} \le 1 \tag{12}$$

for $m = 1, 2, ..., M$ and $i = 1, 2, ..., B - 1$. Adding this constraint yields a significant decrease in solution time, as will be seen in the example.

If a prescribed passband gain is required the value of s can be fixed before the optimization starts. Note also that s must be larger than 0 as this otherwise would yield an optimal value with all variables equal to 0. In general it is possible to constrain s to be larger than 0.1, say, without losing optimality in the solution. This is due to the fact that scaling s with 2 is equal to shifting each coefficient one step.

3.4 Limiting the Number of SPT Terms per Coefficients

In some applications it may be of interest to limit the number of SPT terms for each coefficients. This may be the case for a transposed form FIR filter for which the critical path without pipelining is determined by the coefficient with most SPT terms. Of course, the filter can be pipelined, but the number of pipeline stages required for a given sample rate is also dependent on the maximum number of SPT terms for a coefficient. This can be handled in the optimization by adding the following constraint.

$$\sum_{i=1}^{B} (a^+_{m,i} + a^-_{m,i}) \le L_{max} \tag{13}$$

where $m = 1, 2, \ldots, M$ and L_{max} is the maximum number of SPT terms per coefficient.

4 Reducing the Number of Variables

In the optimization problem in (15) the number of 0/1-variables is $2MB$. This number will be high for filters with stringent specifications and the execution time will be long. It is therefore of interest to find ways to decrease the number of variables before the actual optimization starts.

To obtain the ranges for the coefficients it is possible to use linear programming. The problem can be stated as first a separate maximization of each variable h_m, subject to (8). These values are assigned to the variables $h_m^{(ub)}$. Then a minimization of each variable h_m is performed and the result is assigned the variables $h_m^{(lb)}$. However, these values are for $s = 1$, so the possible range of s must be found to obtain the true bounds for h_m.

For a B-bit CSDC number the maximal possible value, when $B \to \infty$, is

$$\sum_{j=0}^{\lfloor B/2 \rfloor - 1} 2^{-2j-1} = \frac{2}{3} \tag{14}$$

so the maximal value for any positive coefficient can be bounded by 2/3. A negative coefficient is bounded by –2/3.

$$
\begin{aligned}
\text{minimize} \quad & \sum_{m=1}^{M} \sum_{i=1}^{B} (a_{m,i}^{+} + a_{m,i}^{-}) \\
\text{subject to} \quad & \sum_{m=1}^{M} c(m, \omega T) \sum_{i=1}^{B} (a_{m,i}^{+} - a_{m,i}^{-}) 2^{-i} - s(1+\delta_c) \le 0 \qquad \omega T \in \Omega_c \\
& -\sum_{m=1}^{M} c(m, \omega T) \sum_{i=1}^{B} (a_{m,i}^{+} - a_{m,i}^{-}) 2^{-i} + s(1-\delta_c) \le 0 \qquad \omega T \in \Omega_c \\
& \sum_{m=1}^{M} c(m, \omega T) \sum_{i=1}^{B} (a_{m,i}^{+} - a_{m,i}^{-}) 2^{-i} - s\delta_s \le 0 \qquad \omega T \in \Omega_s \\
& -\sum_{m=1}^{M} c(m, \omega T) \sum_{i=1}^{B} (a_{m,i}^{+} - a_{m,i}^{-}) 2^{-i} - s\delta_s \le 0 \qquad \omega T \in \Omega_s
\end{aligned}
\tag{15}
$$

On the other hand, to use the complete range of the CSDC coefficients the value for the largest positive coefficient should be larger than 1/3. The largest negative coefficient should be less than –1/3.

The minimal value for the scaling factor is obtained when the maximal absolute value for the upper or lower bound times the scaling factor is 1/3. Hence

$$
s^{(lb)} = \frac{1}{3\max\{\max\{\left|h_m^{(ub)}\right|, \left|h_m^{(lb)}\right|\}\}} \tag{16}
$$

The maximal value for the scaling factor is when the maximal value of the coefficients minimum absolute bounds are 2/3. This is for coefficients that do not change sign. The upper bound is then

$$
s^{(ub)} = \frac{2}{3\max\{\min\{\left|h_m^{(ub)}\right|, \left|h_m^{(lb)}\right|\}\}} \tag{17}
$$

where the coefficients concerned are those that do not change sign. This reasoning is only valid when at least one coefficient have the same sign on its maximum and minimum value. However, this is the case for most filters unless the design margin is very large.

Considering (14), new bounds on the coefficients can now be derived as

$$
\hat{h}_m^{(ub)} = \begin{cases} \min(2/3, h_m^{(ub)} s^{(lb)}), & h_m^{(ub)} \ge 0 \\ \max(-2/3, h_m^{(ub)} s^{(ub)}), & h_m^{(ub)} < 0 \end{cases} \tag{18}
$$

and

$$
\hat{h}_{m,\mathrm{lb}} = \begin{cases} \min(2/3, h_m^{(lb)} s^{(ub)}), & h_m^{(lb)} \ge 0 \\ \max(-2/3, h_m^{(lb)} s^{(lb)}), & h_m^{(lb)} < 0 \end{cases} \tag{19}
$$

If a the passband gain is specified we have $s^{(ub)} = s^{(lb)} = s_s$, where s_s is the specified passband gain.

Finally, by using (10) the following inequality is obtained

$$
\hat{h}_m^{(ub)} \ge \sum_{i=1}^{B} (a_{m,i}^{+} - a_{m,i}^{-}) 2^{-i} \ge \hat{h}_m^{(lb)} \tag{20}
$$

This inequality can either be added to the constraints of the optimization or used to preevaluate possible values of the variables. Here, (20) is utilized to determine which variables have a fixed value and remove these from the optimization problem. This can be done by searching all possible values for the filter coefficient and keep track of which variables assumes one. All variables that do not assume the value of 1 at any time can be removed. There may also be variables that are 1 for all possible values of the filter coefficient. These variables can also be removed by moving them to the right hand side of the constraint inequalities. If the number of SPT terms per coefficient is limited only the possible values should be searched.

As will be shown in the example below, the number of variables that can be removed depends on the available design margin. The number of variables that can be removed are in general not dependent on the coefficient wordlength. However, the relative savings will be smaller when the wordlength is increased.

5 Examples

In this example a filter with passband edge at 0.2π rad and stopband edge at 0.5π rad is considered. The maximum ripple is 0.01 in both the passband and stopband. The minimum order for this filter is 14. Filters with order 14 to 21 are optimized with a coefficient wordlength of 7 bits. Both unlimited number of SPT terms per coefficients and two SPT terms per coefficient are considered.

The filters were optimized on a Sun Ultra 5 333 MHz with 384 MB of memory. The execution times using the proposed method with and without preprocessing is shown in Fig. 1 (a). However, for both cases s were bounded according to (16) and (17). For reference, the speed-up of constraining the coefficients to be in CSDC representation is shown in Fig. 1 (b).

The speed-up using the proposed preprocessing technique is shown in Fig. 2 (a). Figure 2 (b) shows the number of 0/1-variables with and without preprocessing. The savings are in this case between 37 and 47%. It is clear that significant savings are obtained using the proposed preprocessing technique.

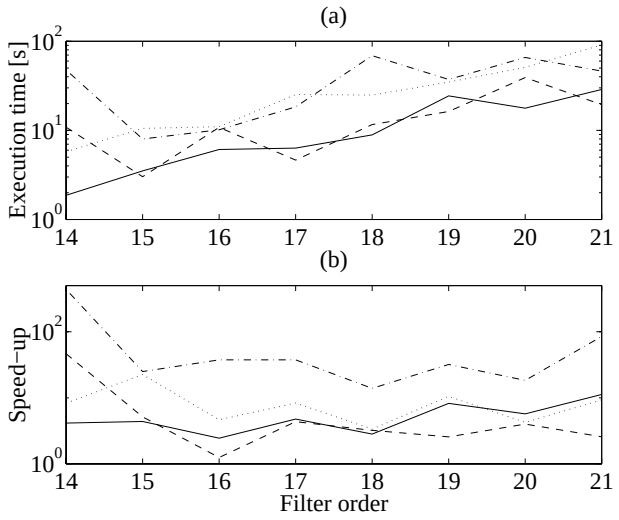

Figure 1: (a) Execution times and (b) speed-up using CSDC in the example, without preprocessing (dash-dotted), without preprocessing with L_{max} = 2 (dotted), with preprocessing (dashed), and with preprocessing and L_{max} = 2 (solid).

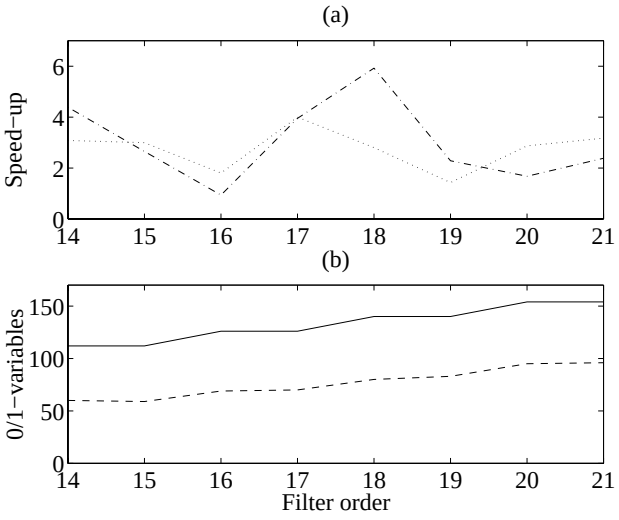

Figure 2: (a) Speed-up using the proposed preprocessing technique in the example for with (dotted) and without (dash-dotted) L_{max} = 2. (b) Number of 0/1-variables with (dashed) and without (solid) the proposed preprocessing technique.

For a filter order of 14 the best solution requires 16 SPT terms for the 8 filter coefficients. A solution with at most 2 SPT terms per coefficient was not possible. For all the odd order filters the best solution requires 13 SPT terms with at most 2 SPT terms per coefficients and 10 SPT terms without this requirement. The other even order filters require 11 SPT terms both with and without the requirement of a maximum of 2 SPT terms per coefficient. The magnitude functions for the resulting filters are shown in Fig. 3.

6 Conclusions

In this paper a mixed integer linear programming problem for design of linear-phase FIR filters with a minimum number of signed-powers-of-two terms was formulated. A preprocessing technique to remove 0/1-variables was proposed and an example showed that the technique improved the execution time significantly.

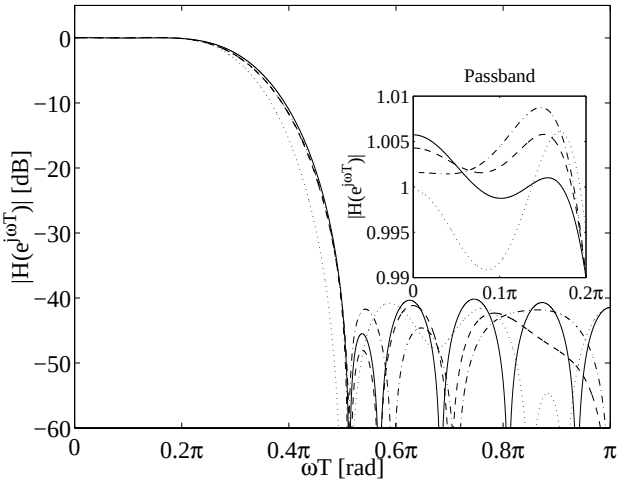

Figure 3: Magnitude functions for the optimized filters in the example, order 14 (solid), odd order (dash-dotted), odd order with Lmax = 2 (dashed), and even order 16 and above (dotted).

References

[1] T. Saramäki and J. Yli-Kaakinen, “Design of digital filters and filter banks by optimization: applications,” *Proc. X European Signal Processing Conf.*, Tampere, Finland, Sept. 4–8, 2000.

[2] Y. C. Lim, S. R. Parker, and A. G. Constantinides, “Finite word length FIR filter design using integer programming over a discrete coefficient space,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, pp. 661–664, Aug. 1982.

[3] Y. C. Lim and S. R. Parker, “FIR filter design over a discrete powers-of-two coefficient space,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 31, pp. 583–591, June 1983.

[4] Y. C. Lim, “Design of discrete-coefficient-value linear phase FIR filters with optimum normalized peak ripple magnitude,” *IEEE Trans. Circuits Syst.*, vol. 37, pp. 1480–1486, Dec. 1990.

[5] H. Samuli, “An improved search algorithm for the design of multiplierless FIR filters with powers-of-two coefficients,” *IEEE Trans. Circuits Syst.*, vol. 36, pp. 1044–1047, July 1989.

[6] R. I. Hartley, “Subexpression sharing in filters using canonic signed digit multipliers,” *IEEE Trans. Circuits Syst.–II*, vol. 43, pp. 677–688, Oct. 1996.

[7] R. Pasko, P. Schaumont, V. Derudder, S. Vernalde, and D. Durackova, “A new algorithm for elimination of common subexpressions,” *IEEE Trans. Computer-Aided Design Integrated Circuits Syst.*, vol. 18, pp 58–68, Jan. 1999.

[8] A. G. Dempster and M. D. Macleod, “Use of minimum-adder multiplier blocks in FIR digital filters,” *IEEE Trans. Circuits Syst.–II*, vol. 42, pp. 569–577, Sep. 1995.

[9] R. Fourer, D. M. Gay, and B. W. Kernighan, *AMPL: A Modeling Languague for Mathematical Programming*, Scientific Press, 1993.

[10] ILOG CPLEX, http://www.ilog.com/products/cplex/, Jan. 2001.