DESIGN OF LINEAR PHASE FIR FILTERS WITH RECURSIVE STRUCTURE AND DISCRETE COEFFICIENTS

H. H. Dam, S. Nordebo, K. L. Teo, A. Cantoni

Australian Telecommunications Research Institute (ATRI),
Curtin University of Technology, Kent St., Bentley, WA 6102, Australia

ABSTRACT

In this paper, we consider a class of FIR filters defined by the first order difference routing digital filter (DRDF) structure and sums of two powers–of–two coefficients. A novel design method is developed for constructing high quality filters with reference to the min–max error criterion. This method is highly efficient in terms of computational time.

Simulation studies show a large improvement over existing methods such as quantization [6]. In some cases, the peak ripple magnitude over the stop and pass bands is reduced up to 13 dB over the quantization method. These results are achieved even for cases involving small number of delays.

1. INTRODUCTION

Most of the research on digital filter design has been concerned with problems related to general purpose digital signal processors. This includes finite word length effects, tradeoffs between filter–length and word–length [1] and design by discrete optimization methods [2].

In recent years, attention has also been given to problems related to actual hardware implementation issues. Filters with sums of two powers–of–two coefficients are of particular significance. For such a filter, multiplication is converted into simple operations involving shift and add, (see e.g. [3], [4] for more details), and hence the hardware implementations become simple with low costs.

The most commonly used criteria in filter designs are the weighted least squares and the weighted min–max [3]. For Finite Impulse Response (FIR) filters with infinite precision coefficients, these optimization problems can be efficiently solved by using appropriate optimization techniques such as least squares approximation methods, linear programming methods and the Remez exchange algorithm [5]. However, for filter designs with discrete coefficients, the corresponding integer programming problems are combinatoric optimization problems and hence are much more difficult to solve than their infinite precision counterparts.

The branch and bound technique is perhaps the most well–known and straightforward integer programming method. It has been successfully used for the design of FIR digital filters, see e.g. [2], [3]. This type of technique is, however, notorious in its computational requirement for problems with large filter lengths. Unless a very good lower bound is available together with other useful forcing and pruning rules, this method is only applicable for small size filter design problems.

In this paper, we consider a specific FIR filter structure based on sums of two powers–of–two FIR coefficients and a digital integrator. This FIR–integrator structure is particularly well suited for low–pass filters with some degree of oversampling. It is attractive for hardware implementations due to its simplicity, stability and cost–effectiveness. This filter structure was proposed in [6] for the DRDF structure which is basically a hardware structure for serial implementation. However, the structure is applicable to a much wider class of implementations, such as with parallel arithmetic, etc. To ensure that the FIR–integrator filter is of linear phase, a simple symmetry constraint is imposed.

The contribution of this paper is a novel computational method based on the branch and bound technique in conjunction with an optimized quantization procedure for solving the linear integer programming problem.

2. FIRST ORDER DRDF WITH LINEAR PHASE

The FIR–integrator filter structure is given in Fig. 1. The

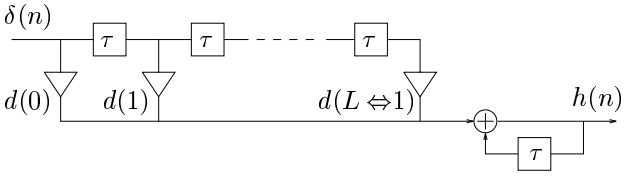

Figure 1: FIR filter structure based on sum of two powers–of–two coefficients $d(n)$ and a digital integrator.

filter consists of a transversal filter with tap–weights $d(n)$ at every τ seconds and cascaded with a first order digital integrator. The overall impulse response $h(n)$ is given by

the difference relation

$$h(n) = h(n \Leftrightarrow 1) + d(n) \tag{1}$$

where $d(n)$ are the DRDF coefficients. It is assumed that $h(n) = d(n) = 0$ for $n < 0$ and $n \geq L$ where L denotes the filter length. The transversal filter coefficients $d(n)$ are expressed as the sum or difference of two powers–of–two terms [3]. Thus $d(n) \in \mathcal{A}_b$ where $\mathcal{A}_b = \left\{ \sum_{i=1}^{2} S_i 2^{g_i} \right\}$ and $0 \leq g_i \leq b \Leftrightarrow 1$. The value of b is the maximum shift in bits, or "shift–range" and S_i takes one of the values $1, 0, \Leftrightarrow 1$.

There are generally four cases for the design of linear–phase FIR filters. Without loss of generality, we will only consider the case where L is odd and the coefficients $h(n)$ are symmetric. Other cases can be reformulated in a similar manner.

It follows from (1) that a necessary and sufficient condition for $h(n)$ to be a FIR filter is that the tap–weights $d(n)$ satisfy the condition

$$h(L \Leftrightarrow 1) = \sum_{k=0}^{L-1} d(k) = 0. \tag{2}$$

This condition also ensures that the transfer function $D(z)$ (the $\mathcal{Z}$–transform of $d(n)$) has a zero at unity, thus the corresponding pole of the integrator is cancelled.

Next, we impose the linear phase constraint to the first order DRDF, considered in [6]

$$h(n) = h(L \Leftrightarrow 1 \Leftrightarrow n) \tag{3}$$

Initially, for $n = 0$ and by (2) we have $h(0) = h(L \Leftrightarrow 1) \Leftrightarrow d(0) = h(0) = 0$. Using induction with respect to n, it can be shown that (3) holds iff the tap–weights $d(n)$ satisfy the following anti–symmetry condition

$$d(n) = \Leftrightarrow d(L \Leftrightarrow n) \tag{4}$$

for all n. The frequency response $H(f)$ of the linear–phase FIR filter $h(n)$ can be written as

$$H(f) = \sum_{n=0}^{L-1} h(n) e^{-j2\pi f \tau n} = e^{-j2\pi f \tau (\frac{L-1}{2})} A(f) \tag{5}$$

where $A(f)$ is the real amplitude response, cf. [5].

For filters with odd length L, the real amplitude is given by $A(f) = \phi(f)^T \mathbf{h}$, where

$$\mathbf{h} = \begin{pmatrix} h(1) \\ \vdots \\ h(M \Leftrightarrow 1) \\ h(M) \end{pmatrix} \text{ and } \phi(f) = \begin{pmatrix} 2cos(2\pi f \tau (M \Leftrightarrow 1)) \\ \vdots \\ 2cos(2\pi f \tau) \\ 1 \end{pmatrix} \tag{6}$$

and $M = (L \Leftrightarrow 1)/2$.

Let $\mathbf{d} = [d(1) \cdots d(M)]^T$ be a $M \times 1$ vector containing the transversal filter coefficients and let $\mathbf{T}$ denote a $M \times M$ lower triangular matrix of all ones. The FIR filter coefficients is then given by $\mathbf{h} = \mathbf{T}\mathbf{d}$ and the real amplitude response is

$$A(f) = \phi(f)^T \mathbf{T}\mathbf{d} \tag{7}$$

3. PROBLEM FORMULATION AND SOLUTION

The filter design problem is to find a coefficient vector $\mathbf{d}$ satisfying the feasibility constraint

$$|A(f) \Leftrightarrow A_d(f)| \leq \varepsilon(f) \tag{8}$$

where $A_d(f)$ is the real amplitude of the desired response, $\varepsilon(f)$ is the specified (strictly positive) design tolerance and f belongs to the interval $[0, .5]$.

For the infinite precision solution, a feasible solution to (8) (if any), can be found by solving the min–max problem

$$\min_{\mathbf{d} \in R^M} \max_{f \in [0,.5]} v(f) |A(f) \Leftrightarrow A_d(f)| \tag{9}$$

where $v(f) = 1/\varepsilon(f)$. The design specification (8) has a feasible solution iff the optimum objective value in (9) is less than or equal to one.

To obtain a quality filter with discrete coefficients, a scaling factor is included in the design procedure. The new mixed integer programming problem is to obtain a set of coefficients $d(n) \in \mathcal{A}_b$, and the scaling factor δ such that the following feasibility conditions are satisfied.

$$\begin{cases} |\delta A(f) \Leftrightarrow A_d(f)| \leq \varepsilon(f) \\ 0 \leq \delta \leq \delta_{max}, \end{cases} \tag{10}$$

where δ_{max} is the maximum value of δ. Note that (10) has M discrete variables and one continuous variable δ.

Let $\mathbf{h}_o$ denote the infinite precision solution to the integer programming problem. The design by quantization starts by fixing the value of δ as

$$\delta = \delta_q = \max_{1 \leq n \leq M} \frac{|h_o(n) \Leftrightarrow h_o(n \Leftrightarrow 1)|}{2^b} \tag{11}$$

followed by a recursive quatization procedure to obtain the discrete coefficients [6] (cf. δ–modulation). Let $Q(x)$ denote a non–uniform quantizer defined for all x. For $|x| \leq 2^b$, the quantized value $Q(x)$ is determined by rounding x to the closest value in $\mathcal{A}_b$, and for $|x| > 2^b$ the quantizer is saturated at levels $Q(x) = \pm 2^b$.

The quantized solution $\mathbf{h}_q$ is then calculated according to the following recursion:

- Initial value:

$$h_q(1) = d_q(1) = Q(h_o(1)/\delta_q) \tag{12}$$

- Repeat for $n = 2, \ldots, M$

$$d_q(n) = Q(h_o(n)/\delta_q \Leftrightarrow h_q(n \Leftrightarrow 1)) \quad (13)$$
$$h_q(n) = h_q(n \Leftrightarrow 1) + d_q(n) \quad (14)$$

An optimal quantized solution is obtained as follows. We observe that for quantization, the scaling value δ contains the most crucial information. The main idea of this step is to obtain an optimum quantized solution by searching for the best value of δ. The maximum ripple in the pass and stop bands is plotted as a function of δ. The best value δ_{q1} with the corresponding coefficients $\mathbf{d}_{q1}$ and $\mathbf{h}_{q1}$ are chosen. This procedure can be done very fast (using e.g. Matlab) and the best solution has much smaller maximum ripple in both pass and stop bands when compared to the quantized solution.

To overcome the non–linear formulation in (10), divide both sides of (10) by δ and introduce a new variable $\lambda = 1/\delta$. Hence, (10) becomes

$$\begin{cases} |A(f) \Leftrightarrow \lambda A_d(f)| \le \lambda \varepsilon(f) \\ \lambda \ge 1/\delta_{max} \end{cases} \quad (15)$$

To obtain an integer solution satisfying (15), a new positive variable γ is introduced. The solution of the following integer optimization problem (if any) is a feasible solution to (15).

$$\begin{cases} \max \gamma \\ |\phi(f)^T \mathbf{T}\mathbf{d} \Leftrightarrow \lambda A_d(f)| + \gamma \le \lambda \varepsilon(f) \\ d(n) \in \mathcal{A}_b \text{ and } \gamma \ge 0 \\ \lambda \ge 1/\delta_{max} \end{cases} \quad (16)$$

or

$$\begin{cases} \max \gamma \\ \phi(f)^T \mathbf{T}\mathbf{d} \Leftrightarrow \lambda(A_d(f) + \varepsilon(f)) + \gamma \le 0 \\ \Leftrightarrow \phi(f)^T \mathbf{T}\mathbf{d} + \lambda(A_d(f) \Leftrightarrow \varepsilon(f)) + \gamma \le 0 \\ d(n) \in \mathcal{A}_b \text{ and } \gamma \ge 0 \\ \lambda \ge 1/\delta_{max} \end{cases} \quad (17)$$

The optimization problem (17) is a linear mixed integer programming problem, and hence the methods for MILP such as "branch and bound" can be applied. Unfortunately, the discrete coefficients are allowed to take values from a non–uniform space and hence not all the commercial software packages can be used to solve the problem directly. However, the problem can be converted into a linear mixed integer problem with binary variables as follows: Introduce $2b$ zeros-ones variables $x_i(n)$ and $y_i(n)$, $0 \le i \le b \Leftrightarrow 1$ for each coefficients $d(n)$, where

$$d(n) = \sum_{i=0}^{b-1} x_i(n) 2^i \Leftrightarrow \sum_{i=0}^{b-1} y_i(n) 2^i \quad (18)$$

and impose the following inequality constraint for $d(n) \in \mathcal{A}_b$

$$\sum_{i=0}^{b-1} x_i(n) + \sum_{i=0}^{b-1} y_i(n) \le 2. \quad (19)$$

This mixed zeros–ones integer programming problem can, in principle, be solved by many MILP software packages such as CPLEX to obtain the global solution. The reality is, however, the exponential increase in the computational requirement when the number of filter coefficients is increased.

In this paper, we introduce an efficient optimization method, based on the above ideas combined with the branch and bound technique to get a further improvement of the optimal quantized solution by searching around its neighbourhood.

For each value of n, $1 \le n \le M$, let $d_l(n)$ and $d_u(n) \in \mathcal{A}_b$ denote the greatest lower bound and the least upper bound of the sequence $h_o(n)/\delta_{q1} \Leftrightarrow h_{q1}(n \Leftrightarrow 1)$, respectively. Obviously, each value of $d_{q1}(n)$ is either $d_l(n)$ or $d_u(n)$, cf. (13). The range of each $d(n)$ is therefore restricted to the smaller set $\{d_l(n), d_u(n)\} \subset \mathcal{A}_b$.

The solution obtained by the quantization procedure is often a good estimate of the optimum solution to (15). The optimization method is to further improve the quantized solution $\mathbf{h}_{q1}$ by optimizing over the reduced region in which $\mathbf{h}_{q1}$ is contained.

The main reason for employing the reduced region as defined above is that this region contains very few variables, and the computational burden is significantly reduced.

New variables $v_1(n)$ and $v_2(n)$, $n = 1, \ldots, M$, are introduced by the transformation

$$d(n) = d_l(n) \cdot v_1(n) + d_u(n) \cdot v_2(n) \quad (20)$$

where $v_1(n), v_2(n) \in \{0, 1\}$ and $v_1(n) + v_2(n) = 1$.

Let $\mathbf{v}$ be the $2M \times 1$ vector: $\mathbf{v} = [v_1(1), v_2(1), \cdots, v_2(M)]^T$. Substituting (20) into the optimization problem (17) yields a new optimization problem with $2*M+2$ variables $\left[\mathbf{v}^T \ \lambda \ \gamma\right]$. This linear mixed optimization problem can be solved efficiently using CPLEX.

4. DESIGN EXAMPLES

We consider a symmetric FIR low–pass filter of length $L = 35$. The filter has pass band $[0 \ .1]$ and stop band $[.2 \ .5]$ in the normalized frequency with $\varepsilon(f) = .004$ for both pass and stop bands. The number of bits for the coefficients is $b = 9$.

Fig. 2 shows the resulting frequency responses where the quantization (a) and the optimum quantization (b) plots are indicated by the dashed and solid lines, respectively. The solution improved by the optimization (c) and the infinite precision solution (d) are represented by the dashed

and solid lines, respectively. The plots show a large improvement, 10 dB for the optimal quantized solution. Further improvement by 3 dB is obtained by searching around the neighbourhood of the optimum quantized solution using CPLEX.

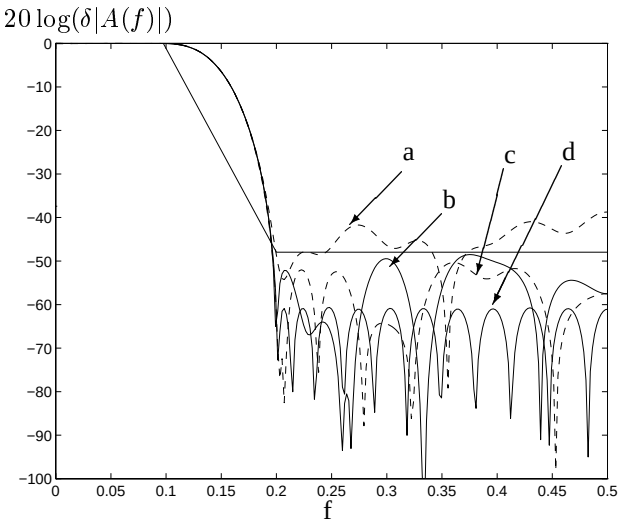

Figure 2: Frequency response for the first order DRDF

Fig. 3 shows the maximum deviation for both pass and stop bands in dB as a function of δ by varying δ up to 20%, either to the left or right of δ_q with small deviation $.000005$. The ring around the point in the middle of the graph stands for the position of the quantized solution. From the plot, it can be seen that a much better solution with smaller ripple can be obtained by changing the value of δ.

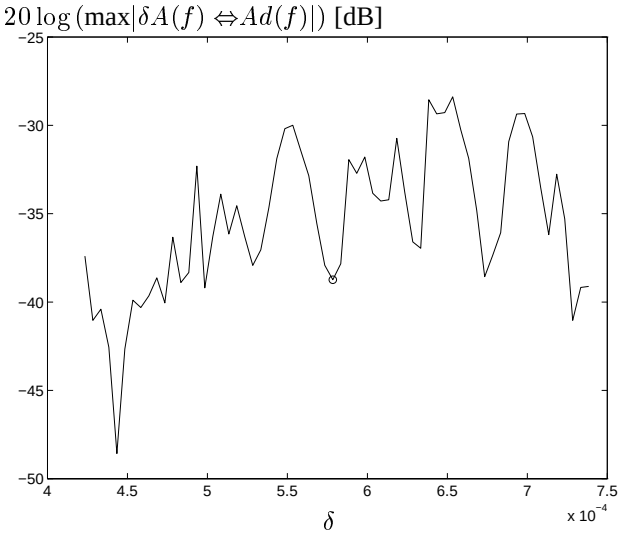

Figure 3: Maximum deviation versus scaling factor δ

5. CONCLUSION

In this paper, we introduced a novel design method for obtaining a near optimum solution to the min–max problem for the first order different routing digital filter (DRDF) structure. The method gives a good solution in a short amount of time. A large improvement over the quantization method was obtained, in some cases up to 13 dB. An important observation was that the traditional quantization procedure was highly sensitive to the choice of δ, a property which was exploited in this contribution. Further research will be performed in order to quantify and further exploit this property.

6. REFERENCES

[1] D. M. Kodek, K. Steiglitz, "Filter–Length Word–Length Tradeoffs in FIR Digital Filter Design", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP–28, no. 6, pp. 739–744, December 1980.

[2] D. M. Kodek, "Design of Optimal Finite Wordlength FIR Digital Filters Using Integer Programming Techniques", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP–28, no. 3, pp. 304–308, June 1980.

[3] Y. C. Lim, S. R. Parker, "FIR Filter Design Over a Discrete Powers– of–Two Coefficient Space", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-31, no. 3, pp. 583–591, June 1983.

[4] S. Mohanakrishnan, J. B. Evans, "Automatic Implementation for FIR Filters on Field Programming Gate Arrays", *IEEE Signal Processing Letters*, vol. 2, pp. 51–53, March 1995.

[5] T. W. Parks, C. S. Burrus, *Digital Filter Design*, John Wiley & Sons, Inc., 1987.

[6] P. J. Van Gerwen, W. F. G. Mecklenbraucker, N. A. M. Verhoeckx, F. A. M. Snijders, H. A. Van Essen, "A New Type of Digital Filter for Data Transmission" *IEEE Transactions on Communications*, vol. COM–23, no. 2, pp. 222–232, February 1975.