<table>
<thead>
<tr>
<th>Title</th>
<th>FRM-based FIR filters with optimum finite word-length performance (Published version)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Lim, Yong Ching; Yu, Ya Jun; Teo, Kok Lay; Saramäki, Tapio</td>
</tr>
<tr>
<td>Date</td>
<td>2007</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10220/4662">http://hdl.handle.net/10220/4662</a></td>
</tr>
</tbody>
</table>

© 2007 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE. This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder. http://www.ieee.org/portal/site.
FRM-Based FIR Filters With Optimum Finite Word-Length Performance

Yong Ching Lim, Fellow, IEEE, Ya Jun Yu, Member, IEEE, Kok Lay Teo, Senior Member, IEEE, and Tapio Saramäki, Fellow, IEEE

Abstract—It is well known that filters designed using the frequency response masking (FRM) technique have very sparse coefficients. The number of nontrivial coefficients of a digital filter designed using the FRM technique is only a very small fraction of that of a minimax optimum design meeting the same set of specifications. A digital filter designed using FRM technique is a network of several subfilters. Several methods have been developed for optimizing the subfilters. The earliest method optimizes the subfilters separately and produces a network of subfilters with excellent finite word-length performance. Subsequent techniques optimize the subfilters jointly and produce filters with significantly smaller numbers of nontrivial coefficients. Unfortunately, these joint optimization techniques, that optimize only the overall frequency response characteristics, may produce filters with undesirable finite word-length properties. The design of FRM-based filters that simultaneously optimizes the frequency response and finite word-length properties had not been reported in the literatures. In this paper, we develop several new optimization approaches that include the finite word-length properties of the overall filter into the optimization process. These new approaches produce filters with excellent finite word-length performance with almost no degradation in frequency response performance.

Index Terms—Coefficient sensitivity, FIR digital filter, finite word-length effect, frequency response masking (FRM), high selectivity filter, low complexity filter, round off noise, sharp filter, signal word-length, sparse coefficient filter.

I. INTRODUCTION

THE FREQUENCY response masking (FRM) technique [1]–[20] was developed for the synthesis of very sharp digital filters with very sparse coefficients. Thus, a filter synthesized using the FRM technique has very low complexity even though the effective filter length is slightly longer than that of the minimax optimum design meeting the same set of frequency response specifications. The FRM technique has been extended to the synthesis of various types of filters such as half-band filters [21]–[23], 2-D filters [24], IIR filters [25]–[28], filter banks [29]–[34], decimators and interpolators [35], [36], and Hilbert transformers [37], [38]. Implementations on various platforms [39]–[41] such as field-programmable gate array (FPGA) have also been investigated. Its applications in transmultiplexer design [42], ECG signal processing [43], hearing aids [44], digital audio [45]–[49] application and analysis, speech recognition [50], array beamforming [51], software radio [52], and noise thermometer [53] have also been reported.

Fig. 1 shows the structure of an FIR filter synthesized using the FRM technique. A filter with $z$-transform transfer function $H(z)$ is synthesized using a network of subfilters $H_a(z^{M_a}), H_{Ma}(z), H_{Ma}(z), \text{ and } z^{-a}$, where $z^{-a}$ represents an appropriate negative integer power of $z$ and $M$ is an integer [1]; all the subfilters have very low arithmetic complexities. Many different optimization techniques have been developed for optimizing the subfilters of Fig. 1. For a given set of frequency response requirements imposed on $H(z)$, there is a wide range of subfilter frequency responses that can meet the requirement. The finite word-length properties [54] of $H(z)$ depend on the frequency responses of the subfilters. The coefficient sensitivity and round-off noise power may differ by many orders of magnitude for different choice of $H_a(z^{M_a}), H_{Ma}(z)$, and $H_{Ma}(z)$. Thus, it is important to steer the optimization algorithm during the course of optimization so as to produce a design with desirable finite word-length properties. This paper addresses the issue of designing FRM-based digital filters with optimum finite word-length properties.

The FRM technique produces a network of subfilters connected in parallel and in cascade. In [1], the frequency responses of the subfilters are optimized independently. It produces a final design with good finite word-length properties although its overall peak ripple magnitude is not as good as that where all the subfilters are optimized simultaneously using nonlinear optimization techniques. The coefficient sensitivity and round-off noise performance of the filter optimized using the technique presented in [1] is analyzed in Section II. Section III shows an example of an FRM-based filter optimized only for overall frequency response performance under infinite precision condition disregarding the finite word-length effect. The coefficient
sensitivity of an FRM-based filter is investigated in Section IV. The investigation leads to a new design approach where the coefficient sensitivity is used as the objective function for optimization. This leads to the design of FRM-based filters with low coefficient sensitivity. An example of a low coefficient sensitivity design is shown in Section V. Based on the principle of minimizing coefficient sensitivity, many approaches, each leading to a different objective function, may be developed. Two such approaches are presented in Section VI. The effects on the various design approaches on the signal word-length required to achieve a given signal-to-noise ratio is presented in Section VII.

II. FINITE WORD-LENGTH PERFORMANCE FOR THE FILTERS OPTIMIZED IN [1]

In general, the frequency responses of the subfilters optimized separately using the linear optimization technique such as that used in [1] will resemble that shown in Fig. 2. In Fig. 2, \( H(e^{j\omega}), H_a(e^{j\omega}M), H_{Ma}(e^{j\omega}), \) and \( H_{Mc}(e^{j\omega}) \) are the frequency responses of the filters whose \( z \)-transform transfer functions are \( H(z), H_a(z^M), H_{Ma}(z), \) and \( H_{Mc}(z) \), respectively. It can be seen from the frequency responses that, for most sinusoidal input frequencies within the pass-band of \( H(e^{j\omega}) \), the input signal flows through either the path \( H_a(z^M)H_{Ma}(z) \) or the path \( \{z^{-b} - H_a(z^M)\}H_{Mc}(z) \) since the pass-bands of \( H_a(e^{j\omega}M)H_{Ma}(e^{j\omega}) \) is the stop-bands of \( \{z^{-b} - H_a(e^{j\omega}M)\}H_{Mc}(e^{j\omega}) \) and vice versa. Input sinusoids with frequencies in the stop-band of \( H(e^{j\omega}) \) will be rejected by both \( H_a(e^{j\omega}M)H_{Ma}(e^{j\omega}) \) and \( \{z^{-b} - H_a(e^{j\omega}M)\}H_{Mc}(e^{j\omega}) \). Thus, for subfilters with frequency responses as shown in Fig. 2, the scenario that two large data are subtracted to form a small data (the scenario that will lead to a serious finite word-length problem) never occur.

Fig. 2. Frequency responses of the subfilters of Fig. 1.

Fig. 3. Frequency response plots for \( H_a(e^{j\omega}M)H_{Ma}(e^{j\omega}), \{z^{-b} - H_a(e^{j\omega}M)\}H_{Mc}(e^{j\omega}), \) and \( H(e^{j\omega}) \) for an example exhibiting a serious finite word-length problem.

III. FINITE WORD-LENGTH PERFORMANCE OF SUBFILTERS DESIGNED FOR OPTIMAL FREQUENCY RESPONSE PERFORMANCE

Many powerful nonlinear optimization techniques have been developed for the design of the subfilters. These advanced nonlinear optimization techniques jointly optimize all the subfilters for obtaining the optimum overall frequency response. The frequency response of the overall filter obtained using these nonlinear optimization techniques is significantly better than that obtained by optimizing the subfilters separately using the linear optimization technique. Unfortunately, even though the filter designed using these advanced techniques has good overall frequency response under infinite precision arithmetic condition; its finite word-length properties may be undesirable. The path \( H_a(e^{j\omega}M)H_{Ma}(z) \) and the path \( \{z^{-b} - H_a(z^M)\}H_{Mc}(z) \) may both have very high gain causing the outputs of \( H_{Ma}(z) \) and \( H_{Mc}(z) \) to be very large. Since the pass-band gain of the filter’s overall frequency response is unity, the very large output signals of \( H_{Ma}(z) \) and \( H_{Mc}(z) \) must have opposite signs so that the signals cancel each other to form the filter’s final output that has a comparable magnitude with the input. Since the output signal of the overall filter is obtained from the difference between two large signals, for filter designed using the nonlinear optimization technique, the filter exhibits serious finite word-length problem. We shall illustrate this problem by means of an example.

Consider the design of a low-pass filter with band edges at \( 0.3f_s \) and \( 0.3f_s \), respectively. The allowed peak ripple magnitude is 0.01. When the peak ripple magnitude is used as the objective function for minimization, there are a large number of minima with almost the same objective function values. The optimization algorithm may converge to any one of the minima if no further criterion is imposed. The frequency responses \( H_a(e^{j\omega}M)H_{Ma}(e^{j\omega}), \{z^{-b} - H_a(e^{j\omega}M)\}H_{Mc}(e^{j\omega}), \) and \( H(e^{j\omega}) \) (with the linear phase term removed), for a typical solution are shown in Fig. 3. The coefficient values are shown in Table I. The value of \( M \) in \( H_a(z^M) \) is 9. As can be seen from Table I, the coefficients of \( H_a(z) \) have very large magnitude.
and coefficient values must be made to the discussion of the coefficient sensitivity and signal round off noise. The investigation of the round off noise property is deferred to Section VII while this section is devoted to the discussion of the coefficient sensitivity. The frequency response of the overall filter \(H(\omega)\) is given by

\[
H(\omega) = H_a(\omega)H_{Ma}(\omega) + \{e^{-j\omega} - H_a(\omega)\}H_{Mc}(\omega),
\]

Let the \(\eta\)th coefficient values of \(H_a(z^{\omega}),H_{Ma}(z),\) and \(H_{Mc}(z)\) be \(h_a(n),h_{Ma}(n),\) and \(h_{Mc}(n),\) respectively. In actual implementation, all coefficient values must be made discrete since all implementation platforms are finite precision; this introduces round off errors into the coefficient values. The magnitudes of the errors depend on the multiplier word length. Let the errors introduced into \(h_a(n),h_{Ma}(n),\) and \(h_{Mc}(n)\) be \(\Delta h_a(n),\Delta h_{Ma}(n),\) and \(\Delta h_{Mc}(n),\) respectively, when the coefficient values are rounded. We shall investigate the change in \(H(\omega)\) caused by small values of \(\Delta h_a(n),\Delta h_{Ma}(n),\) and \(\Delta h_{Mc}(n),\) where \([x]\) denotes the “magnitude of \(x\)." Suppose that \(H(\omega),H_a(\omega),H_{Ma}(\omega),\) and \(H_{Mc}(\omega)\) become \(H(\omega) + \Delta H(\omega),H_a(\omega) + \Delta h_a(n),H_{Ma}(\omega) + \Delta h_{Ma}(n),\) and \(H_{Mc}(\omega) + \Delta h_{Mc}(n),\) respectively, when \(h_a(n),h_{Ma}(n),\) and \(h_{Mc}(n)\) become \(h_a(n) + \Delta h_a(n),h_{Ma}(n) + \Delta h_{Ma}(n),\) and \(h_{Mc}(n) + \Delta h_{Mc}(n).\) We shall assume that \(|\Delta h_a(n)|,|\Delta h_{Ma}(n)|,\) and \(|\Delta h_{Mc}(n)|\) are sufficiently small so that \(|\Delta H(\omega)|,|\Delta h_a(n)|,|\Delta H_{Ma}(\omega)|,\) and \(|\Delta H_{Mc}(\omega)|\) are small. Thus, we have

\[
\begin{align*}
\Delta H(\omega) &= \{H_a(\omega) - H_a(\omega)\} + \Delta h_{Ma}(n) - \Delta h_{Ma}(n)\{H_{Ma}(\omega) - H_{Ma}(\omega)\}

&+ \{e^{-j\omega} - H_a(\omega)\}\{H_{Mc}(\omega) - H_{Mc}(\omega)\}

&\times \{H_{Ma}(\omega) + \Delta H_{Ma}(\omega)\}.
\end{align*}
\]

Neglecting the second order error terms, we have

\[
\Delta H(\omega) = \{H_{Ma}(\omega) + \Delta H_{Ma}(\omega)\} - \Delta h_{Ma}(n)\{H_{Ma}(\omega) - H_{Ma}(\omega)\}

&+ \{e^{-j\omega} - H_a(\omega)\}\{H_{Mc}(\omega) - H_{Mc}(\omega)\}

&\times \{H_{Ma}(\omega) + \Delta H_{Ma}(\omega)\}.
\]

It can be seen from (3) that the magnitudes of the sensitivities of \(H(\omega)\) with respect to changes in \(H_a(\omega),H_{Ma}(\omega),\) and \(H_{Mc}(\omega)\) are \(|H_{Ma}(\omega) - H_{Ma}(\omega)|,|H_a(\omega)|,\) and \(|e^{-j\omega} - H_a(\omega)|,\) respectively. Thus, \(|H_{Ma}(\omega) - H_{Ma}(\omega)|,|H_a(\omega)|,\) and \(|e^{-j\omega} - H_a(\omega)|\) should be minimized for good robustness against changes in \(H_{Ma}(\omega),H_{Ma}(\omega),\) and \(H_{Mc}(\omega)\). Squaring both sides of (3) leads to

\[
[\Delta H(\omega)]^2 = \{H_{Ma}(\omega) - H_{Ma}(\omega)\}^2 + \{H_a(\omega)\}^2\{H_{Ma}(\omega)\}^2

&+ \{e^{-j\omega} - H_a(\omega)\}^2\{H_{Mc}(\omega)\}^2

&+ 2\{H_{Ma}(\omega) - H_{Ma}(\omega)\}\{H_{Mc}(\omega)\}^2

&\times \{H_{Ma}(\omega) + \Delta H_{Ma}(\omega)\}.
\]

Let \(E[x]\) denotes the expected value of \(x.\) Taking the expected values for both sides of (4) and assuming that \(E[H_a(\omega)H_{Ma}(\omega)\Delta H_{Ma}(\omega)] = E[H_{Ma}(\omega)\Delta H_{Ma}(\omega)] = 0\) (see Appendix 1), we have

\[
E[(\Delta H(\omega))^2] = \{H_{Ma}(\omega) - H_{Ma}(\omega)\}^2E[\{H_a(\omega)\}^2]

&+ \{H_a(\omega)\}^2E[(\Delta H_{Ma}(\omega))^2]

&+ \{e^{-j\omega} - H_a(\omega)\}^2E[(\Delta H_{Mc}(\omega))^2].
\]

The frequency response \(H_{c}(\omega)\) of a linear phase symmetrical impulse response FIR filter with length \(N_x\) and coefficient values \(h_x(n),n = 0,\ldots,N_x - 1\) is given by

\[
H_{c}(\omega) = e^{-j\omega}\sum_{n=0}^{N_x/2} 2h_x(n) \left(\frac{N_x}{2} - n\right) \cos\left(\omega\left(n - \frac{1}{2}\right)\right).
\]

(6a)
for \( N_x \) even, and given by

\[
H_x(e^{j\omega}) = e^{-j\frac{N_x-1}{2}\omega} \times \left\{ h_x \left( \frac{N_x - 1}{2} \right) + 2 \sum_{n=1}^{\frac{N_x-1}{2}} h_x \left( \frac{N_x - 1}{2} - n \right) \cos(\omega n) \right\}
\]  

(6b)

for \( N_x \) odd.

Suppose that changing \( h_x(n) \) to \( h_x(n) + \Delta h_x(n) \) causes \( H_x(e^{j\omega}) \) to change to \( H_x(e^{j\omega}) + \Delta H_x(e^{j\omega}) \). Thus

\[
\Delta H_x(e^{j\omega}) = e^{-j\frac{N_x-1}{2}\omega} \sum_{n=1}^{\frac{N_x}{2}} 2\Delta h_x \left( \frac{N_x}{2} - n \right) \cos \left( \omega \left( n - \frac{1}{2} \right) \right)
\]  

(7a)

for \( N_x \) even, and

\[
\Delta H_x(e^{j\omega}) = e^{-j\frac{N_x}{2}\omega} \sum_{n=1}^{\frac{N_x-1}{2}} \Delta h_x \left( \frac{N_x - 1}{2} - n \right) \cos(\omega n)
\]  

(7b)

for \( N_x \) odd.

Assume that for

\[
i \neq j, \quad E\{\Delta h_x(i)\Delta h_x(j)\} = 0.
\]  

(8a)

Define the quantity \( \varepsilon^2 \) as

\[
\varepsilon^2 = E\{(\Delta h_x(i))^2\}.
\]  

(8b)

Define

\[
||\Delta H_x(e^{j\omega})||^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} E\{(\Delta H_x(e^{j\omega}))^2\} d\omega.
\]  

(9)

From (7), (8), and (9), we have

\[
||\Delta H_x(e^{j\omega})||^2 = N_x E\{\Delta h_x(n)^2\} = N_x \varepsilon^2.
\]  

(10)

Although \( \Delta H_x(e^{j\omega}) \) is a function of \( \omega \) for a given filter, \( E\{(\Delta H_x(e^{j\omega}))^2\} \) for a large number of independent filters is a constant independent of \( \omega \) if \( \Delta h_x(i) \) has flat spectrum. Thus

\[
E\{(\Delta H_x(e^{j\omega}))^2\} = N_x \varepsilon^2.
\]  

(11)

Applying the result of (11), (5) becomes

\[
E\{(\Delta H(e^{j\omega}))^2\} = \varepsilon^2 \left\{ N_M [H_{Ma}(e^{j\omega}) - H_{Mc}(e^{j\omega})]^2 + N_M [H_a(e^{j\omega}M)]^2 \right\}
\]  

\[
+ N_M \{e^{-j\omega\bullet} - H_a(e^{j\omega}M)\}^2 \}
\]  

(12)

where \( N_M, N_{Ma}, \) and \( N_{Mc} \) are the numbers of coefficients of \( H_a(z), \) \( H_{Ma}(z), \) and \( H_{Mc}(z), \) respectively.

We have

\[
\frac{1}{2\pi} \int_{-\pi}^{\pi} (H_x(e^{j\omega}))^2 d\omega = \sum_{n=0}^{N_x-1} (h_x(n))^2.
\]  

(13)

Define

\[
||h_a||^2 = \sum_{n=0}^{N_x-1} (h_a(n))^2
\]  

(14a)

\[
||1 - h_a||^2 = \left( 1 - h_a \left( \frac{N_a - 1}{2} \right) \right)^2 + 2 \sum_{n=0}^{\frac{N_a-1}{2}} (h_a(n))^2
\]  

(14b)

\[
||h_{Ma} - h_{Mc}||^2 = \sum_{n=0}^{N_{Ma}-1} (h_{Ma}(n) - h_{Mc}(n))^2.
\]  

(14c)

For \( H_{Ma}(z) \) and \( H_{Mc}(z) \) to produce the same phase shifts so that their outputs can be summed correctly, \( H_{Ma}(z) \) and \( H_{Mc}(z) \) must have the same order [1], i.e., \( N_{Ma} = N_{Mc} \). If they do not have the same order, the lower order transfer function should be preceded and appended with zero valued coefficients so that \( H_{Ma}(z) \) and \( H_{Mc}(z) \) have the same order and that \( H_{Ma}(e^{j\omega}) \) and \( H_{Mc}(e^{j\omega}) \) have the same group delay [1].

From (12)–(14), we have

\[
||\Delta H(e^{j\omega})||^2 = \{N_a||h_{Ma} - h_{Mc}||^2
\]  

\[
+ N_M||h_a||^2 + N_{Mc}||1 - h_a||^2\} \varepsilon^2.
\]  

(15)

Let

\[
S^2 = N_a||h_{Ma} - h_{Mc}||^2 + N_M||h_a||^2 + N_{Mc}||1 - h_a||^2.
\]  

(16)

From (15) and (16), we have

\[
||\Delta H(e^{j\omega})||^2 = S^2 \varepsilon^2.
\]  

(17)

From (17), and since \( \varepsilon^2 = E\{\Delta h_x(i)^2\} \) by definition, it is clear that \( S^2 \) is a coefficient sensitivity measure. In order to minimize the coefficient sensitivity \( S^2 \), the peak ripple magnitude of \( H(e^{j\omega}) \), denoted as \( \delta \), may be set as a constraint and \( S^2 \) becomes the objective for minimization as in the following:

\[
\text{Minimize} \quad S^2 = N_a||h_{Ma} - h_{Mc}||^2 + N_M||h_a||^2
\]  

\[
+ N_{Mc}||1 - h_a||^2 \quad \text{subject to} \quad \delta \leq \delta_0.
\]  

(18a)

In (18b), \( \delta_0 \) is a predefined constant and, in (18a), the variables in the optimization are all the coefficient values.

V. EXAMPLE WITH MINIMUM COEFFICIENT SENSITIVITY

We choose the design of a low-pass filter with the same band edges, \( M \) value for \( H_a(z^M) \), and subfilter lengths as the filter whose frequency response is shown in Fig. 3 as an example to illustrate the superiority of this new design technique. The peak ripple magnitude is relaxed by 2% (the exact amount whether 1%, 2%, or 3% is not critical since there are a large number of minima with almost the same objective function value) to 0.0102 and (18a) is used as the objective function. The coefficient values are shown in Table II and the frequency response plots for \( H_a(e^{j\omega}M)H_{Ma}(e^{j\omega}), \) \( e^{-j\omega\bullet} - H_a(e^{j\omega}M)\) \( H_{Mc}(e^{j\omega}) \), and
TABLE II
COEFFICIENT VALUES FOR $H_a(z)$, $H_M(a-z)$, AND $H_M(z)$ FOR THE FILTER OBTAINED BY MINIMIZING $S_T^2$

| $a_{15}$ | $-0.03100$ | $-0.00(13)$ | $-20$ | $0.00(060)$. $a_{1,2}$ | $22$ |
| $a_{13}$ | $0.0163$ | $a_{12}$ | $21$ | $0.00(067)$. $a_{1,2}$ | $21$ |
| $a_{11}$ | $0.0333$ | $a_{11}$ | $20$ | $0.00(048)$. $a_{1,2}$ | $20$ |
| $a_{10}$ | $-0.03671$ | $a_{10}$ | $19$ | $0.00(762)$. $a_{1,2}$ | $19$ |
| $a_{9}$ | $-0.01144$ | $a_{9}$ | $18$ | $0.00(009)$. $a_{1,2}$ | $18$ |
| $a_{8}$ | $-0.06733$ | $a_{8}$ | $17$ | $-0.00(783)$. $a_{1,2}$ | $17$ |
| $a_{7}$ | $-0.02290$ | $a_{7}$ | $16$ | $-0.00(280)$. $a_{1,2}$ | $16$ |
| $a_{6}$ | $-0.07251$ | $a_{6}$ | $15$ | $0.00(962)$. $a_{1,2}$ | $15$ |
| $a_{5}$ | $-0.07350$ | $a_{5}$ | $14$ | $-0.00(874)$. $a_{1,2}$ | $16$ |
| $a_{4}$ | $-0.04921$ | $a_{4}$ | $13$ | $0.00(995)$. $a_{1,2}$ | $13$ |
| $a_{3}$ | $-0.01400$ | $a_{3}$ | $12$ | $0.01(540)$. $a_{1,2}$ | $12$ |
| $a_{2}$ | $-0.01813$ | $a_{2}$ | $11$ | $0.00(783)$. $a_{1,2}$ | $11$ |
| $a_{1}$ | $-0.04349$ | $a_{1}$ | $10$ | $0.00(215)$. $a_{1,2}$ | $10$ |
| $a_{0}$ | $0.00982$ | $a_{0}$ | $9$ | $0.00(000)$. $a_{1,2}$ | $9$ |
| $a_{-1}$ | $0.00775$ | $a_{-1}$ | $8$ | $0.03(195)$. $a_{1,2}$ | $8$ |
| $a_{-2}$ | $0.00000$ | $a_{-2}$ | $7$ | $0.01(304)$. $a_{1,2}$ | $7$ |
| $a_{-3}$ | $0.01675$ | $a_{-3}$ | $6$ | $0.02(394)$. $a_{1,2}$ | $6$ |
| $a_{-4}$ | $-0.02333$ | $a_{-4}$ | $5$ | $0.02(398)$. $a_{1,2}$ | $5$ |
| $a_{-5}$ | $0.01039$ | $a_{-5}$ | $4$ | $0.04(757)$. $a_{1,2}$ | $4$ |
| $a_{-6}$ | $0.06668$ | $a_{-6}$ | $3$ | $-0.00(629)$. $a_{1,2}$ | $3$ |
| $a_{-7}$ | $-0.01179$ | $a_{-7}$ | $2$ | $-0.05(590)$. $a_{1,2}$ | $2$ |
| $a_{-8}$ | $0.01089$ | $a_{-8}$ | $1$ | $0.21(111)$. $a_{1,2}$ | $1$ |
| $a_{-9}$ | $0.01963$ | $a_{-9}$ | $0$ | $0.07(998)$. $a_{1,2}$ | $0$ |

$r = 10$, the peak ripple magnitude is $0.01034$. The frequency response plot for the filter shown in Table II is shown in Fig. 5. Frequency response plot for the filter with coefficient quantization step size $= 2^{-14}$.

For the purpose of comparison, the value of $N_a||H_M(a-z) - h_{Ma}(z)||^2 + N_{Ma}||h_{Ma}(z)||^2 + N_{Ma}||1 - h_{Ma}(z)||^2$ (i.e., the equivalent $S_{T}^2$ value) for the filter whose coefficients are shown in Table I is $6.78 \times 10^9$. High coefficient sensitivity is expected. To achieve a peak ripple magnitude of about $0.0103$ the coefficient quantization step size for the filter shown in Table I should not be larger than $2^{-27}$. Specifically, if the quantization step size for the coefficients of $H_a(z)$ is $2^{-22}$ and that of $H_{Ma}(z)$ and $H_{M}(z)$ are $2^{-27}$, the peak ripple magnitude is $0.01034$. It is interesting to note that the requirement on the relative precision for the coefficients of $H_a(z)$ is much more stringent than that of $H_{Ma}(z)$, and $H_{M}(z)$ for the filter shown in Table I and that shown in Table II are roughly the same but the requirement on the relative precision for the coefficients of $H_{Ma}(z)$ and $H_{M}(z)$ for the filter shown in Table I and that shown in Table II differ by about $10^4$. This is not surprising since the ratio of their respective values of $S_{T}^2$ is about $10^{4}; \sqrt{6.78 \times 10^9/26.34} = 1.6 \times 10^4$. A summary of the comparisons between the filter of Table I and that of Table II is shown in Table III.

For the greatly improved coefficient sensitivity is achieved with almost no penalty in frequency response performance. This is because the objective function has many minima with insignificant difference in peak ripple magnitude. If the value of $\delta$ in (18b) is set close to (say within 2%) from the optimum solution obtained in minimizing $\delta$ without taking coefficient sensitivity into consideration, minimizing (18a) simply produces a solution with excellent coefficient sensitivity without noticeable degradation in frequency response performance.
VI. COEFFICIENT SENSITIVITY AS A FUNCTION OF FREQUENCY

If the coefficient quantization step size is $Q$, (17) may be rewritten as [54]

$$
\| \Delta H(\epsilon^{j\omega}) \|^2 = \frac{S^2 Q^2}{12}.
$$

From (9) and (19), we have

$$
\sqrt{\int_{-\pi}^{\pi} E[(\Delta H(\epsilon^{j\omega}))^2]} = S_1 Q \sqrt{6}.
$$

Equation (20) provides useful statistic for the frequency response due to coefficient quantization. Expression for $E[(\Delta H(\epsilon^{j\omega}))^2]$ is given in (12). In order to have a better understanding of the function $E[(\Delta H(\epsilon^{j\omega}))^2]$, we plot in Fig. 6 the various components in the right-hand side of (12) constituting $E[(\Delta H(\epsilon^{j\omega}))^2]$ for the filter of Table II.

It can be seen from Fig. 6 that $N_a \{ H_{Ma}(\epsilon^{j\omega}) - H_{Mc}(\epsilon^{j\omega}) \}^2 + N_{Ma}(H_a(\epsilon^{j\omega}))^2$ + $N_{Mc}(e^{-j\omega} - H_a(\epsilon^{j\omega}))^2$ peaks at around the transition band, i.e., the frequency response at the frequency band near the transition is most sensitive to coefficient quantization. In this particular example, the largest contributor is the $N_a \{ H_{Ma}(\epsilon^{j\omega}) - H_{Mc}(\epsilon^{j\omega}) \}^2$ term. However, the largest contributor depends on the specific example. For the filter shown in Table I, the largest contributors are the $N_{Ma}(H_a(\epsilon^{j\omega}))^2$ and $N_{Mc}(e^{-j\omega} - H_a(\epsilon^{j\omega}))^2$ terms.

The previous observations lead to the following new approaches in obtaining a low coefficient sensitivity design.

One of these approaches is to relax the peak frequency response ripple magnitude from its optimum value by a small amount (say 2%) and minimize the peak of

$$
N_a \{ H_{Ma}(\epsilon^{j\omega}) - H_{Mc}(\epsilon^{j\omega}) \}^2 + N_{Ma}(H_a(\epsilon^{j\omega}))^2 + N_{Mc}(e^{-j\omega} - H_a(\epsilon^{j\omega}))^2,
$$

i.e., let

$$
S_2(\omega) = N_a \{ H_{Ma}(\epsilon^{j\omega}) - H_{Mc}(\epsilon^{j\omega}) \}^2 + N_{Ma}(H_a(\epsilon^{j\omega}))^2 + N_{Mc}(e^{-j\omega} - H_a(\epsilon^{j\omega}))^2
$$

and the objective function is

$$
\text{minimize } ||S_2(\omega)||_{\infty}
$$

where $||S_2(\omega)||_{\infty}$ is the maximum value of $|S_2(\omega)|$ over all $\omega$ and the variables in the optimization are all the coefficient values. Depending on the optimization package used, minimizing $||S_2(\omega)||_{\infty}$ may be achieved by minimizing $\sum_i |S_2(\omega_i)|^p$ over a dense grid of $i$, where $p$ is a large positive integer.

There are other possibilities. It can be seen from (5) that $H_{Ma}(\epsilon^{j\omega}) - H_{Mc}(\epsilon^{j\omega})$, $(H_a(\epsilon^{j\omega}))$, and $(e^{-j\omega} - H_a(\epsilon^{j\omega}))$ are the sensitivity measures of $H(\epsilon^{j\omega})$ with respect to changes in $H_{Ma}(\epsilon^{j\omega})$, $H_{Mc}(\epsilon^{j\omega})$, and $H_a(\epsilon^{j\omega})$, respectively. The maximum of the peak values of $(H_{Ma}(\epsilon^{j\omega}) - H_{Mc}(\epsilon^{j\omega}))$, $(H_a(\epsilon^{j\omega}))$, and $(e^{-j\omega} - H_a(\epsilon^{j\omega}))$ may be minimized. Define

$$
S_{3,1}(\omega) = N_a \{ H_{Ma}(\epsilon^{j\omega}) - H_{Mc}(\epsilon^{j\omega}) \}^2
$$

$$
S_{3,2}(\omega) = N_{Ma}(H_a(\epsilon^{j\omega}))^2
$$

$$
S_{3,3}(\omega) = N_{Mc}(e^{-j\omega} - H_a(\epsilon^{j\omega}))^2.
$$

The objective function is

minimize maximum of

$$
\{|S_{3,1}(\omega)| \|^\infty, |S_{3,2}(\omega)| \|^\infty, |S_{3,3}(\omega)| \|^\infty\}. 
$$

The variables in the optimization are all the coefficient values. Depending on the optimization package used, minimizing the maximum of $\{|S_{3,1}(\omega)| \|^\infty, |S_{3,2}(\omega)| \|^\infty, |S_{3,3}(\omega)| \|^\infty\}$ may be achieved by minimizing $\sum |S_{3,1}(\omega_i)|^p + \sum |S_{3,2}(\omega_i)|^p + \sum |S_{3,3}(\omega_i)|^p$ over a dense grid of $i$ where $p$ is a large positive integer.

Our experience shows that the minimization of $S_{3,1}^p, S_{3,2}^p, S_{3,3}^p$, or the maximum of $\{|S_{3,1}(\omega)| \|^\infty, |S_{3,2}(\omega)| \|^\infty, |S_{3,3}(\omega)| \|^\infty\}$ all lead to filters with excellent coefficient sensitivities; their differences in coefficient sensitivity is insignificant. The actual approach that should be adopted depends on other factors such as availability and robustness of optimization packages, hardware implementation platform for the resulting filter, and etc.

The coefficient values for a design obtained by minimizing $||S_2(\omega)||_{\infty}$ are shown in Table IV and the frequency response plots for $H_{Ma}(\epsilon^{j\omega}), H_{Mc}(\epsilon^{j\omega}), H_a(\epsilon^{j\omega})$, and $H(\epsilon^{j\omega})$ are shown in Fig. 7. Its equivalent $S_1^2$ value (i.e., its $N_a|h_{Ma} - h_{Mc}|^2 + N_{Ma}|h_a|^2 + N_{Mc}|1 - h_a|^2$ value) is 29.02. The peak value for $N_a \{ H_{Ma}(\epsilon^{j\omega}) - H_{Mc}(\epsilon^{j\omega}) \}^2 + N_{Ma}(H_a(\epsilon^{j\omega}))^2 + N_{Mc}(e^{-j\omega} - H_a(\epsilon^{j\omega}))^2$ is 75.71. The various components in the right-hand side of (12) constituting $E[(\Delta H(\epsilon^{j\omega}))^2]$ are plotted in Fig. 8. If the quantization step size for the coefficients of $H_a(\epsilon^{j\omega})$ is $2^{-14}$ and that of $H_{Ma}(\epsilon^{j\omega})$ and $H_{Mc}(\epsilon^{j\omega})$ are $2^{-13}$, the peak ripple
magnitude is 0.01021; the frequency response plot for this case is shown in Fig. 9.

For comparison, the peak values for $N_M\{H_M(e^{j\omega}) - H_M(e^{j\omega M})\}^2 + N_M[H_M(e^{j\omega M})]^2 + H_M(e^{-j\omega M})\}^2$ for the filter in Table II and that in Table I are 86.3 and $1.34 \times 10^{10}$, respectively.

VII. SIGNAL WORD-LENGTH REQUIREMENT

If the filter is optimized for frequency response performance with infinite precision arithmetic disregarding finite word-length effect, the resulting filter may not only require very long word-length to represent the coefficient values in order to avoid deterioration in its frequency response but may also require very long word length to represent the signals at intermediate nodes in order to avoid signal overflow or excessive round off noise. For example, for the filter of Table I, the maximum gains of $H_a(z^M)H_{Ma}(z)$ and $\{z^M - H_a(z^M)\}H_{Ma}(z)$ are 17167.0568 and 17166.0486, respectively, at 0.19437 $f_s$, where $f_s$ is the sampling frequency, i.e., for a sinusoidal input at 0.19437 $f_s$ with unit magnitude, the outputs of $H_{Ma}(z)$ and $H_{Ma}(z)$ will have magnitudes 17167.0568 and 17166.0486, respectively. Since the overall filter gain in the pass-band is unity, the outputs of $H_{Ma}(z)$ and $H_{Ma}(z)$ cancel each other to produce a gain of 1.0082 at 0.19437 $f_s$, i.e., 15 binary digits of the most significant bits are lost in the cancellation. The peak frequency response magnitude of the subtransfer functions for the filters shown in Tables I, II, and IV are shown in Table V. A higher maximum gain implies more significant bits must be allocated to the filter to avoid overflow.

For any filter with input $x(n)$, output $y(n)$, and impulse response $h(n)$, if $\|x(n)\| \leq 1$, then $\|y(n)\| \leq \sum_{m=0}^{\infty} |h(m)|$. Thus, the sum of the magnitudes of the impulse response of a filter provides the absolute upper bound for the output of the filter. The sum of the magnitudes of the impulse responses of the subtransfer functions for the filters shown in Tables I, II, and Table IV are shown in Table VI. It can be seen from Table VI that, for the filter of Table I, for input signal with magnitude
bounded by unity, the outputs at $H_{Ma}(z)$ and $H_{Mc}(z)$ may have magnitudes as large as 72 123 and 72 121, respectively. This means that 17 integer bits (not including sign bit) are needed to avoid signal overflow. For the filter of Table II, only two integer bits are needed whereas, for the filter of Table IV, three integer bits are needed to avoid signal overflow if the input signal magnitude is bounded by unity.

Each multiplication produces a double word-length result. When the signal word length is shortened by rounding the signal, a rounding error is introduced. The rounding error may be represented as a round-off noise with noise power equal to $Q^2/12$ [54] injected into the system at the point of rounding, where $Q$ is the quantization step size. The noise model is shown in Fig. 10. If there is no word-length truncation at $z^{-1}$, $e_c(z) = 0$.

The output noise $\nu(z)$, due to $e_a(z), e_c(z), e_{Ma}(z)$, and $e_{Mc}(z)$ is given by (25)

$$\nu(z) = e_{Ma}(z) + e_{Mc}(z)H_{Ma}(z) + e_a(z)\{H_{Ma}(z) - H_{Mc}(z)\} + e_c(z)H_{Mc}(z).$$

Let the noise powers for $e_a(z), e_c(z), e_{Ma}(z)$, and $e_{Mc}(z)$ be $\sigma^2_a, \sigma^2_c, \sigma^2_{Ma}$, and $\sigma^2_{Mc}$, respectively. The total noise power at the output, denoted by $\sigma^2$, is given by (26)

$$\sigma^2 = \sigma^2_{Ma} + \sigma^2_{Mc} + \sigma^2_a[H_{Ma} - H_{Mc}]^2 + \sigma^2_c[H_{Mc}]^2.$$

TABLE V
PEAK FREQUENCY RESPONSE MAGNITUDES FOR FILTERS SHOWN IN TABLES I, II, AND IV

<table>
<thead>
<tr>
<th>Transfer function</th>
<th>Table I</th>
<th>Table II</th>
<th>Table IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_M(z)$</td>
<td>17.045.5363</td>
<td>0.5200</td>
<td>0.7317</td>
</tr>
<tr>
<td>$H_M(z)^{-1}$</td>
<td>17.044.5363</td>
<td>1.3807</td>
<td>1.5347</td>
</tr>
<tr>
<td>$H_M(z)^{-1}H_{Ma}(z)$</td>
<td>17.167.0568</td>
<td>0.5292</td>
<td>0.7412</td>
</tr>
<tr>
<td>$H_M(z)^{-1}H_{Mc}(z)$</td>
<td>17.166.0486</td>
<td>1.3892</td>
<td>1.5436</td>
</tr>
<tr>
<td>Overall $H(z)$</td>
<td>1.0100</td>
<td>1.0101</td>
<td>1.0100</td>
</tr>
</tbody>
</table>

TABLE VI
SUM OF IMPULSE RESPONSE MAGNITUDES FOR FILTERS SHOWN IN TABLES I, II, AND IV

<table>
<thead>
<tr>
<th>Transfer function</th>
<th>$\sum$ magnitudes of impulse response</th>
<th>Table I</th>
<th>Table II</th>
<th>Table IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_M(z)$</td>
<td>43.924.5023</td>
<td>1.3165</td>
<td>1.8559</td>
<td></td>
</tr>
<tr>
<td>$z^{-1}H_M(z)$</td>
<td>43.923.5023</td>
<td>2.1565</td>
<td>2.6362</td>
<td></td>
</tr>
<tr>
<td>$H_M(z)H_{Ma}(z)$</td>
<td>72.123.4078</td>
<td>2.4075</td>
<td>3.1829</td>
<td></td>
</tr>
<tr>
<td>$z^{-1}H_M(z)H_{Ma}(z)$</td>
<td>72.121.8457</td>
<td>3.6068</td>
<td>4.3811</td>
<td></td>
</tr>
<tr>
<td>$H(z)$</td>
<td>2.9947</td>
<td>2.9916</td>
<td>2.9947</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 10. Noisemodel for FRM-based filter. If there is no word-length truncation at $z^{-1}, e_c(z) = 0$.

VIII. CONCLUSION

For a given frequency response specification, a filter synthesized using the FRM technique will have very much smaller number of non-zero coefficients than the minimax optimum design. The original FRM technique [1] produces filters with excellent finite word-length property (significantly better than that of the minimax optimum design) but the number of coefficients can be further reduced by using advanced nonlinear optimization technique. In this paper, we show by means of an example that advanced nonlinear optimization technique that optimizes
the overall frequency response disregarding finite word-length properties may produce filters that have very high coefficient sensitivity and require very long signal word length. In order to overcome this shortcoming, in this paper, we present several techniques in (18a), (22), and (24), respectively, that include coefficient sensitivity measures into the objective function for optimization. These techniques produce filters with excellent finite word-length properties.

The computing resources required and convergent properties of the new techniques do not differ significantly from the conventional technique. The actual computing resources required and convergent properties depend on the optimization package used. The optimization package we used always converges but, unfortunately, it is easily trapped in local optimum solution. We overcome the problem by initiating the optimization process at different initial solutions. The development of an optimization package that is not easily trapped in local optimum solution will be the next challenge.

The optimal low sensitivity design will be a good initial solution for further research in low complexity designs such as minimum shift-and-add design.

APPENDIX I

We have

\[ \Delta H_{Ma}(e^{j\omega}) = \sum_{n} \Delta h_{Ma}(n)e^{-jn\omega} \]

and

\[ \Delta H_{Me}(e^{j\omega}) = \sum_{n} \Delta h_{Me}(n)e^{-jn\omega}. \]

Thus

\[ E\{\Delta H_{Ma}(e^{j\omega})\Delta H_{M}(e^{j\omega})\} = E\{\sum_{n} \Delta h_{Ma}(n)e^{-jn\omega} \sum_{n} \Delta h_{Me}(n)e^{-jn\omega}\} = \sum_{m} \sum_{n} E\{\Delta h_{Ma}(n)\Delta h_{Me}(m)\}e^{-jn(m+n)} = 0 \]

since

\[ E\{\Delta h_{Ma}(n)\Delta h_{Me}(m)\} = 0. \]

Similarly, it can be shown that

\[ E\{\Delta h_{a}(e^{j\omega})\Delta H_{Ma}(e^{j\omega})\} = E\{\Delta h_{a}(e^{j\omega})\Delta H_{Me}(e^{j\omega})\} = 0. \]

REFERENCES


Yong Ching Lim (S’80­M’80–SM’92–F’00) received the A.C.G.I., B.Sc., D.I.C., and Ph.D. degrees, all in electrical engineering, from Imperial College, University of London, U.K., in 1977, 1977, 1980, and 1980, respectively.

Since 2003, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he is currently a Professor. From 1980 to 1982, he was a National Research Council Research Associate in the Naval Postgraduate School, Monterey, CA. From 1982 to 2003, he was with the Department of Electrical Engineering, National University of Singapore. His research interests include digital signal processing and VLSI circuits and systems design.

Dr. Lim was a recipient of the 1996 IEEE Circuits and Systems Society’s Guillemin­Cauer Award, the 1990 IRE (Australia) Norman Hayes Memorial Award, the 1977 IEEE (UK) Prize, and the 1974–1977 Siemens (Imperial College) Award. He served as a lecturer for the IEEE Circuits and Systems Society under the distinguished lecturer program from 2001 to 2002 and as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS from 1991 to 1993 and from 1999 to 2001. He has also served as an Associate Editor for Circuits, Systems and Signal Processing from 1993 to 2000. He served as the Chairman of the DSP Technical Committee of the IEEE Circuits and Systems Society from 1998 to 2000. He served in the Technical Program Committee’s DSP Track as Track Chairman in IEEE ISCAS ’97 and IEEE ISCAS ’00 and as a Track Co­chairman in IEEE ISCAS ’99. He is the General Chairman for IEEE APCCAS’06. He is a member ofEta Kappa Nu.

Ya Jun Yu (S’99–M’05) received the B.Sc. and M.Eng. degrees in biomedical engineering from Zhejiang University, Hangzhou, China, in 1994 and 1997, respectively, and the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2004.

Since 2005, she has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where she is currently an Assistant Professor. From 1997 to 1998, she was a Teaching Assistant at Zhejiang University. She joined the Department of Electrical and Computer Engineering, National University of Singapore as a Post Master Fellow in 1998 and remained in the same department as a Research Engineer until 2004. In 2002, she was a visiting Researcher at the Tampere University of Technology, Tampere, Finland, and The Hong Kong Polytechnic University, Hong Kong. Since 2007, she has been with the Temasek Laboratories at Nanyang Technological University as a Research Fellow in 2004. Her research interests include digital signal processing and VLSI circuits and systems design.

Kok Lay Teo (M’74–SM’87) received the B.Sc. degree in telecommunications engineering from Ngee Ann Technical College, Singapore, and the M.A.Sc and Ph.D. degrees in electrical engineering from the University of Ottawa, Ottawa, ON, Canada.

He is currently Chair of Applied Mathematics and Head of the Department of Mathematics and Statistics, Curtin University of Technology, Australia. He was with the Department of Applied Mathematics, University of New South Wales, Australia, the Department of Industrial and Systems Engineering, Na-
Tapi Saramäki (M’98–SM’01–F’02) was born in Orivesi, Finland, on June 12, 1953. He received the Diploma Engineer (with honors) and the Doctor of Technology (with honors) degrees in electrical engineering from the Tampere University of Technology (TUT), Tampere, Finland, in 1978 and 1981, respectively.

Since 1977, he has held various research and teaching positions at TUT, where he is currently a Professor of Signal Processing and a Docent of Telecommunications. He is also a co-founder and a system-level designer of VLSI Solution Oy, Tampere, Finland, specializing in efficient VLSI implementations of both analog and digital signal processing algorithms for various applications. He is also the President of Aragit Oy Ltd., Tampere, Finland, which was founded by four TUT professors and concentrates on spreading worldwide their know-how on information technology to the industry. In 1982, 1985, 1986, 1990, and 1998 he was a Visiting Research Fellow with the University of California, Santa Barbara, and in 1987 with the California Institute of Technology, Pasadena, and in 2001, with the National University of Singapore. He has written more than 250 international journal and conference articles, various international book chapters, and holds three worldwide-used patents. His research interests are in digital signal processing, especially filter and filter bank design, VLSI implementations, and communications applications, as well as approximation and optimization theories.

Dr. Saramäki was a recipient of the 1987 and 2007 IEEE Circuits and Systems Society’s Guillemin-Cauer Award as well as two other Best Paper Awards. In 2004, he was also awarded the honorary membership (Fellow) of the A. S. Popov Society for Radio-Engineering, Electronics, and Communications (the highest membership grade in the society and the 80th honorary member since 1945) for “great contributions to the development of DSP theory and methods and great contributions to the consolidation of relationships between Russian and Finnish organizations.” He is also a founding member of the Median-Free Group International. He was an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS: ANALOG AND DIGITAL SIGNAL PROCESSING from 2000 to 2001, and is currently an Associate Editor for Circuits, Systems, and Signal Processing. He has been actively taking part in many duties in the IEEE Circuits and Systems Society’s DSP Committee, namely by being a Chairman (2002–2004), a Distinguished Lecturer (2002–2003), a Tract or a Co-Track Chair for many ISCAS symposiums (2003–2005). In addition, he has been one of the three chairmen of the annual workshop on Spectral Methods and Multirate Signal Processing (SMMSP), started in 2001.