An Improved Approach for the Synthesis of Multiplication-Free Highly-Selective FIR Half-Band Decimators and Interpolators

(Invited Paper)

Tapio Saramäki and Juha Yli-Kaakinen
Department of Signal Processing
Tampere University of Technology
Tampere, Finland
{tapio.saramaki, juha.yli-kaakinen}@tut.fi

***Abstract*—A substantial improvement is provided in comparison with the systematic approach recently presented by the authors of this paper to generate multiplication-free decimators and interpolators. As the earlier one, the proposed approach is founded, first, on expressing the transfer function of a linear-phase finite-impulse response (FIR) half-band filters as a sum of the terms $(1/2)z^{-M}$ and $G(z^2)$, where the odd integer M is the order of $G(z)$, and, then, on constructing $G(z)$ as a special tapped cascaded interconnection of identical subfilters. The essential improvement is in a quite optimized way of generating multiplication-free tap coefficients such that the resulting passband alteration of the subfilter required by meeting the overall filter criterion becomes significantly larger than in the earlier approach, where this variation is already huge against what is required by the overall filter. This increased passband alteration results in a considerably lowered subfilter order, which, in turn, even further eases finding multiplication-free representations for its coefficient values and decreases the overall filter order. Examples are included illustrating the superiority of the resulting multiplication-free filters compared with those achievable by the earlier approach and their direct-form FIR equivalents.**

***Index Terms*—Multiplication-free filters, optimization, half-band filters, decimators and interpolators.**

I. Introduction

Linear-phase finite-impulse response (FIR) half-band filters play a very important role, due to their many attractive properties, in various systems, where sampling rate conversion is used. First of all, a low-pass half-band filter transfer function of order $2M$ with M odd is expressible as (see, e.g., [1])

$$H(z) = \sum_{n=0}^{2M} h[n]z^{-n} = \frac{1}{2}z^{-M} + G(z^2), \qquad (1a)$$

where

$$G(z) = \sum_{n=0}^{M} g[n]z^{-n} \qquad (1b)$$

with

$$g[M-n] = g[n] \quad \text{for} \quad n = 0, 1, \ldots, M. \qquad (1c)$$

Hence, when assuming that $h[M] = 2^{-1}$ is realized by using only a single shift operation and the coefficient symmetry in $G(z)$ is exploited, only $(M+1)/2$ multipliers are required in the overall implementation, as is really desired. Second, constructing $H(z)$ as shown above guarantees that the resulting filter has an exactly linear phase response, which is of great importance in many multirate filtering applications. Third, most importantly, when using commutative structures [2], the above half-band filters acting as the 2-to-1 decimator and the 1-to-2 interpolator can be efficiently implemented using the structures of Figs. 1(a) and 1(b), respectively. It is worth pointing out that in most overall systems using sampling rate conversion, these filters serve as sub-blocks in order to make the entire system computationally efficient (see, e.g., [3]).

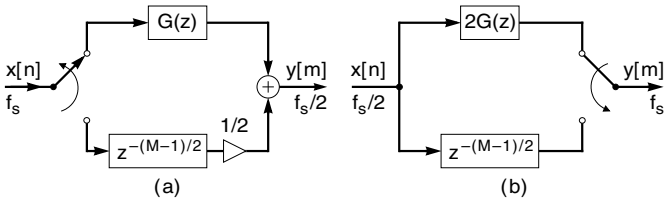

Fig. 1. Efficient implementations of the half-band linear-phase FIR filter for (a) the 2-to-1 decimator and (b) the 1-to-2 interpolator.

When the multirate systems of Fig. 1 are intended to be implemented as a very large-scale integration (VLSI) circuit, it is crucial to avoid the use of the costly multiplier element by representing all the coefficient values as a few signed powers-of-two terms. This is because for these representation forms, the coefficients values can be realized as a sequence of shifts and adds and/or subtracts. These shifts are often hardwired and, therefore, essentially free. Hence, only a few adders and/or subtracters remain in the overall implementation.

However, finding out such few signed power-of-two representations for all the coefficient values of $G(z)$ in Fig. 1 is not at all a very trivial task. This is especially true in applications, where both a high attenuation and a narrow transition bandwidth are required and, consequently, a large number of fractional bits and high-order filters are demanded. The above-mentioned difficulties arise because the absolute values of the impulse-response samples of $G(z)$ around the center, taking place between the impulse-response samples $g[(M-1)/2]$ and $g[(M+1)/2]$, are relative high with a very few zero-

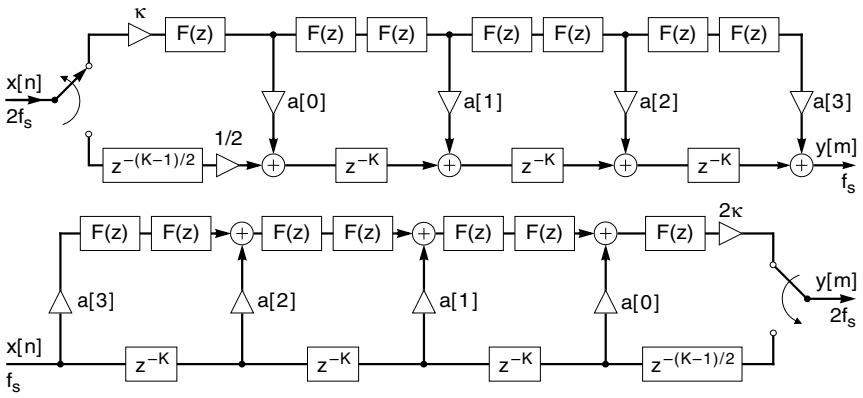

Fig. 2. Implementation of the proposed half-band filters for the sampling rate alteration by a factor of two in the $L = 3$ case.

valued most significant fractional bits, whereas the absolute values of the remaining samples decrease quite quickly when approaching the beginning and the end of the impulse response. In addition of having large absolute values, the central impulse-response values demand a relatively high accuracy, due to their important role in establishing the desired magnitude response. Therefore, in the above-mentioned demanding applications, these central coefficient values necessarily require significantly more signed powers-of-two terms in comparison with the remaining impulse-response sample values.

In order to get around this problem, it has been shown by Saramäki, Karema, Ritoniemi, and Tenhunen in [3] that it is beneficial to construct $G(z)$ as a tapped cascaded interconnection of identical subfilter transfer functions $F(z)$ as follows:

$$G(z) = \sum_{\ell=0}^{L} \alpha[\ell] z^{-(L-\ell)K} [F(z)]^{2\ell+1}, \tag{2a}$$

where

$$F(z) = \sum_{n=0}^{K} f[n] z^{-n} \tag{2b}$$

with

$$f[K-n] = f[n] \quad \text{for} \quad n = 0, 1, \ldots, K \tag{2c}$$

and

$$\alpha[\ell] = \kappa a[\ell] \quad \text{for} \quad \ell = 0, 1, \ldots, L. \tag{2d}$$

Here, K, the order of $F(z)$ with symmetric impulse response is odd. Most importantly, for each term in the summation of (2a) both the impulse response is symmetric and the delay is equal to $(2L+1)K/2$. Therefore, $(2L+1)K$, the order of the overall $G(z)$ with symmetric impulse response is odd, which is required to end up with the transfer function $H(z)$ of a linear-phase half-band FIR filter to be expressible according to (1a), (1b), and (1c). Hence, the order of the $H(z)$ using $G(z)$, as given by (2a), (2b), (2c), and (2d), is $2(2L+1)K$ and its zero-phase frequency response can be written as

$$H(\omega) = 1/2 + G(2\omega), \tag{3a}$$

where

$$G(\omega) = \sum_{l=0}^{L} \alpha[\ell] [F(\omega)]^{2\ell+1} \tag{3b}$$

with

$$F(\omega) = \sum_{n=0}^{(K-1)/2} f\left[\frac{K-1}{2} - n\right] \cos\left[\left(n + \frac{1}{2}\right)\omega\right] \tag{3c}$$

and $\alpha[\ell] = \kappa a[\ell]$, according to (2d).

The first and second subfigures in Fig. 2 show the efficient decimator and interpolator structures for the resulting overall half-band filter in the $L = 3$ case. These are commutative structures [2], where the delay terms are shared for reducing their use in the implementation. In this figure, the "effective" tap coefficients used for interconnecting the identical subfilter transfer functions $F(z)$ together are given by $\alpha[\ell] = \kappa a[\ell]$ for $\ell = 0, 1, \ldots, L$, that is, all the tap coefficients are multiplied by the same constant κ. As will be seen in Section III in this contribution, the use of the additional constant κ in the proposed improved approach plays a very crucial role in finding out the desired few signed powers-of-two terms for the tap coefficient values. The structures corresponding to the approach described in [3] and [10] is obtained setting $\kappa \equiv 1$ and $a[\ell] = \alpha_\ell$. Similarly, in (2a) and (3b), $\alpha[\ell] = \kappa a[\ell]$ for $\ell = 0, 1, \ldots L$ do not have the common constant κ and are simply replaced by α_ℓ.

Several multiplication-free designs have been given in [3] without describing the quantization scheme behind them at all. Since the appearance of the article, these results have achieved a great interest in literature (see, e.g., [4]–[9]). That was the reason why the authors of this paper have disclosed in [10] the quantization scheme used in [3]. This scheme is very systematic and is based on the reasoning which enables one to find the tap coefficients using simple formulas up to $L = 3$.

When considering the allowable passband ripple required by the subfilter to meet the given overall filter specifications, the above technique closely resembles, from the approximation theory point of view, a maximally flat approach in the determining of the tap coefficients. More clearly speaking, this ripple, even though it is huge against what is demanded by the overall filter, is considerably smaller than the maximum achievable, as has been pointed out in the Conclusion in [10].

The purpose of this contribution is to describe an improved approach to determine multiplication-free tap coefficients such the passband variation of the subfilter is considerably increased. As a matter of fact, for the infinite-precision designs,

this improved design scheme resembles the minimax approach because it enables one to maximize the passband variation for the subfilter. When regarding the use of the quantization scheme to be described in Section III, the following three issues should be pointed out. First, as in the previous approach, $L = 3$ is the highest values under consideration due to the facts that $L = 3$ gives a wide enough passband variation for the subfilter even in very stringent applications and in the $L > 3$ case the search for proper multiplication-free tap coefficients becomes difficult. Second, the highest number of fractional bits worth using is the one with which 95 percent of the maximum achievable passband variation is obtained. In this case, the increased variation in the passband ripple of the subfilter results in significantly milder criteria for the subfilter in comparison with the earlier approach described in [10] and [3]. This enlarged variation both reduces the order of the subfilter required to meet the stated overall specifications and makes the criteria for the minimum-order subfilter the mildest possible. This fact, in turn, both decreases the overall filter order and eases finding multiplication-free coefficient representations for the subfilter. Third, the quantization scheme enables one to find solutions such that the smaller is the number of fractional bits in use, the less is the reduction in the passband variation in comparison to the maximized one. This gives an opportunity to make proper compromises between the multiplication-free tap coefficients and both the corresponding coefficient values and the order of the subfilter. In some cases, it is beneficial even to slightly increase the subfilter order in order to arrive at very simple coefficient representation forms at the expense of a minor increase in the overall filter order.

The outline of the rest of this paper is as follows. Section II concentrates on those fact, tools, and key observations on which the quantization scheme to be described in Section III heavily relies. Section III present in detail the technique to find multiplication-free tap coefficients. Section IV exemplifies how to appropriately achieve the above-mentioned compromises between multiplication-free tap coefficient and both the multiplication-free coefficients and the order of subfilter in one practical filter design problem. Finally, Section V gives concluding remarks.

II. Preliminary Facts, Tools, and Key Observation Behind the Proposed Synthesis Quantization Scheme

This section concentrates on those facts, tools, and key observations on which the quantization scheme to be described in Section III relies. The organization of this section is as follows. Subsections II-A, II-B, and II-C present the facts referred to as Facts I, II and III, Subsections II-D and II-E present tools referred to as Tools I and II, whereas Subsection II-F presents the observations referred to as Key Observation I.

A. Fact I: Transfer of the Synthesis of $H(z)$ to that of $G(z)$

It is well-known (see, e.g., [1]) that if $G(\omega)$, as given by (3b) and (3c), oscillates within the limits $1/2 \pm \delta$ on $[0, 2\omega_p]$, then $H(\omega)$, as given by (3a), (3b), and (3c), automatically oscillates within $1 \pm \delta$ ($\pm\delta$) on $[0, \omega_p]$ ($[\pi - \omega_p, \pi]$). Hence, the rest of this contribution concentrates on the synthesis of $G(z)$, as given by (2a), (2b), (2c), and (2d), such that the above-mentioned specifications are met.

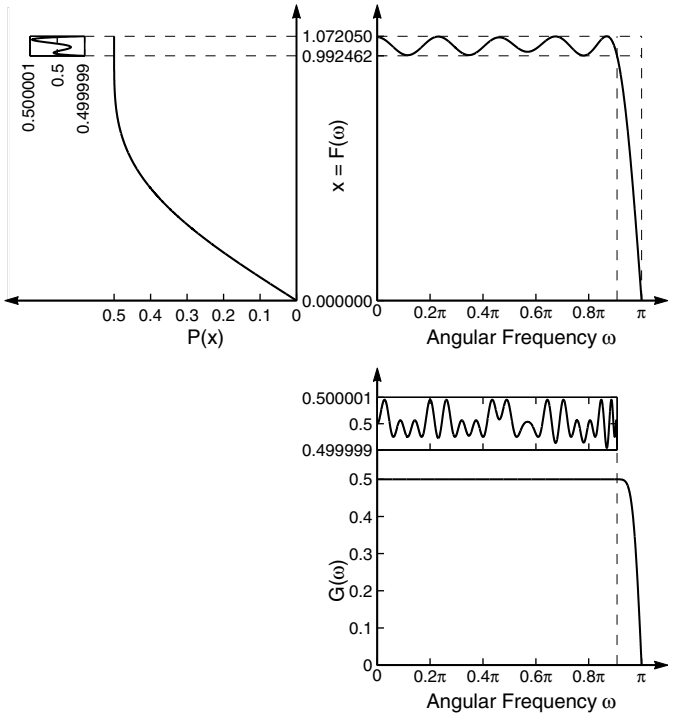

Fig. 3. Mapping of $P(x)$ with the aid of the substitution $x = F(\omega)$ to $G(\omega)$.

B. Fact II: General Simultaneous Criteria for the Tap Coefficients and the Subfilter

The condition for $G(\omega)$, as given by (3b) and (3c), to stay within the limits $1/2 \pm \delta$ on $[0, 2\omega_p]$ can be combined with the aid of the "effective" tap coefficient $\alpha[\ell] = \kappa a[\ell]$ for $\ell = 0, 1, \ldots, L$ and the zero-phase frequency response $F(\omega)$, as given by (3c), as follows:

$$1/2 - \delta \le G(\omega) = P(x)_{|x\,=\,F(\omega)} \le 1/2 + \delta \qquad (4a)$$

for $\omega \in [0, 2\omega_p]$, where

$$P(x) = \sum_{\ell=0}^{L} \alpha[\ell] x^{2\ell+1}. \qquad (4b)$$

The above criterion is met by requiring that the polynomial $P(x)$ and the zero-phase frequency response $F(\omega)$ simultaneously satisfy the following relations:

$$1/2 - \delta \le P(x) \le 1/2 + \delta \quad \text{for } x^{(\text{low})} \le x \le x^{(\text{up})} \qquad (5a)$$

and

$$x^{(\text{low})} \le F(\omega) \le x^{(\text{up})} \quad \text{for } \omega \in [0, 2\omega_p]. \qquad (5b)$$

Figure 3 exemplifies these relations. As seen from this figure, the substitution $x = F(\omega)$ can be regarded as a transformation that maps the passband region $x^{(\text{low})} \le x \le x^{(\text{up})}$ of $P(x)$ to the passband region $0 \le \omega \le 2\omega_p$ of $G(\omega)$. Hence, the amplitude values are preserved and only the argument axis is changed. Alternatively, $P(x)$ can be interpreted as an

amplitude change function, which tells that if the subfilter response $F(\omega)$ achieves the value x_0, then the overall response $G(\omega)$ achieves the value $P(x_0)$ without regard of the angular frequency ω. Hence, the passband regions of $F(\omega)$ and $G(\omega)$ are the same and all that happens is that the proper multiple use of the same subfilter reduces the large passband variation in $F(\omega)$ to a small passband variation in $G(\omega)$.

C. Fact III: Recreation of the Same Infinite-Precision $G(\omega)$ by an Extremely Many Different Simultaneous Choices of the Tap Coefficients and the Subfilter

In the above, $x^{(\text{low})}$ and $x^{(\text{up})}$ are by no means fixed for the infinite-precision design as the same $G(\omega)$ can be generated by using $\widetilde{x}^{(\text{low})} = \gamma x^{(\text{low})}$ and $\widetilde{x}^{(\text{up})} = \gamma x^{(\text{up})}$ as the lower and upper limits for the interval, respectively, instead of $x^{(\text{low})}$ and $x^{(\text{up})}$. This is achieved by simultaneously changing the "effective" tap coefficient values and the subfilter coefficients to become $\widetilde{\alpha}[\ell] = \alpha[\ell]/\gamma^{2\ell+1}$ for $\ell = 1, 2, \ldots, L$ and $\widetilde{f}[n] = \gamma f[n]$ for $n = 1, 2, \ldots, K$, respectively. Simultaneously, the original interval center $x_{\text{ave}} = (x^{(\text{low})} + x^{(\text{up})})/2$ becomes $\widetilde{x}_{\text{ave}} = \gamma x_{\text{ave}}$. The explanation why it is very essential to take x_{ave} as a continuous unknown in the overall design technique of Section III will be postponed to Subsection II-E after describing in the following subsection the algorithm to generate the optimized infinite-precision polynomial for any given value of x_{ave}.

D. Tool I: Generation of Optimum Infinite-Precision Polynomials

The first tool on which the quantization scheme to be described in Section III relays is to optimize, after giving δ, x_{ave}, and L, the infinite-precision polynomial, denoted for the later use by [1]

$$R(x_{\text{ave}}, \delta, L, x) = \sum_{\ell=0}^{L} \alpha(x_{\text{ave}}, \delta, L)[\ell] x^{2\ell+1}, \tag{6}$$

such it stays within $1/2 - \delta$ and $1/2 + \delta$ on $[x^{(\text{low})}(x_{\text{ave}}, \delta, L), x^{(\text{up})}(x_{\text{ave}}, \delta, L)]$, where

$$x^{(\text{low})}(x_{\text{ave}}, \delta, L) = x_{\text{ave}} - \Delta(x_{\text{ave}}, \delta, L) \tag{7a}$$

$$x^{(\text{up})}(x_{\text{ave}}, \delta, L) = x_{\text{ave}} + \Delta(x_{\text{ave}}, \delta, L) \tag{7b}$$

so that $\Delta(x_{\text{ave}}, \delta, L)$ is maximized. This goal is reached by determining the polynomial coefficients such that $\Delta(x_{\text{ave}}, \delta, L)$ is maximized such that the polynomial $R(x_{\text{ave}}, \delta, L, x)$ oscillates on $[x^{(\text{low})}(x_{\text{ave}}, \delta, L), x^{(\text{up})}(x_{\text{ave}}, \delta, L)]$ within $1/2 - \delta$ and $1/2 + \delta$ achieving these values at $L + 2$ points so that $R(x_{\text{ave}}, \delta, L, x^{(\text{low})}(x_{\text{ave}}, \delta, L)) = 1/2 - \delta$. Because there exist $L + 2$ points on $[x^{(\text{low})}(x_{\text{ave}}, \delta, L), x^{(\text{up})}(x_{\text{ave}}, \delta, L)]$, where the polynomial alternatively achieves the values $1/2 - \delta$ and $1/2 + \delta$, this polynomial is the optimum in the sense that the length of the interval is the maximum achievable. This is because the number of extremal points is $L + 2$, which is one more than the number of unknowns in the polynomial.

[1]From now on in this contribution, the dependences of the polynomials, coefficients, and other quantities on the design parameters are emphasized whenever appropriate.

The above polynomial can be conveniently generated, by means of a Remez multiple exchange algorithm (for instance, the function `firpm` or `firgr` in the Filter Design Toolbox [11] provided by MathWorks, Inc.) for the design of linear-phase FIR filters and proper manipulations, using the following steps:

Step 1: Consider the transfer function [2]

$$E(Z) = \sum_{n=0}^{2L+1} e[n] Z^{-n}, \tag{8a}$$

where

$$e[2L + 1 - n] = e[n] \quad \text{for} \quad n = 0, 1, \ldots, 2L + 1. \tag{8b}$$

Determine the $L+1$ unknowns $e[n]$ of its zero-phase frequency response as given by

$$E(\Omega) = \sum_{n=0}^{L} 2e[L - n] \cos[(n + 1/2)\Omega] \tag{9}$$

as well as $\Omega_p(\delta, L)$ such that $\Omega_p(\delta, L)$ is maximized subject to the condition that $E(\Omega)$ oscillates on $[0, \Omega_p(\delta, L)]$ within $1/2 - \delta$ and $1/2 + \delta$ achieving these values at $L + 2$ angular frequencies so that $E(\Omega_p(\delta, L)) = 1/2 - \delta$. Because $\Omega_p(\delta, L)$ is in regard to δ the function that becomes larger, the desired value of $\Omega_p(\delta, L)$ can be determined quite quickly by using an efficient line search technique together with the Remez multiple exchange algorithm.

Step 2: Because $E(Z)$ has a single fixed zero at $Z = -1$, $E(\Omega)$ can be factorized as (see, e.g., [1])

$$E(\Omega) = \cos(\Omega/2) \left[\widehat{e}[L] + 2 \sum_{n=1}^{L} 2\widehat{e}[L - n] \cos(n\Omega) \right], \tag{10a}$$

where

$$\widehat{e}[0] = 2e[0] \quad \text{and} \quad \widehat{e}[n] = 2e[n] - \widehat{e}[n] \text{ for } n = 1, 2, \ldots, L. \tag{10b}$$

Step 3: Use the identities $\cos(3\Omega) = 32\cos^6(\Omega/2) - 48\cos^4(\Omega/2) + 18\cos^2(\Omega/2) - 1$, $\cos(2\Omega) = 8\cos^4(\Omega/2) - 8\cos^2(\Omega/2) + 1$, and $\cos(\Omega) = 2\cos^2(\Omega/2) - 1$ to rewrite $E(\Omega)$ in the $L = 3$ case as

$$E(\Omega) = \cos(\Omega/2) \times \left\{ \widehat{\alpha}[0] + \widehat{\alpha}[1]\cos^2(\Omega/2) + \widehat{\alpha}[2]\cos^4(\Omega/2) + \widehat{\alpha}[3]\cos^6(\Omega/2) \right\}, \tag{11a}$$

where, after some manipulations,

$$\begin{aligned} \widehat{\alpha}[0] &= \widehat{e}[3] - 2\widehat{e}[2] + 2\widehat{e}[1] - 2\widehat{e}[0], \\ \widehat{\alpha}[1] &= 4\widehat{e}[2] - 16\widehat{e}[1] + 36\widehat{e}[0], \\ \widehat{\alpha}[2] &= 16\widehat{e}[1] - 96\widehat{e}[0], \\ \widehat{\alpha}[3] &= 64\widehat{e}[0]. \end{aligned} \tag{11b}$$

[2]For the transfer function $E(Z)$ at Steps 1 and 2, Z, instead of z, is used in order to clearly distinguish it from the transfer function $H(z)$, $G(z)$, and $F(z)$ which are synthesized in the contribution. Similarly, for the zero-phase frequency response of $E(Z)$, Ω, instead of ω, is used for the same reasons.

For $L = 1$ and $L = 2$, the corresponding values of $\widehat{\alpha}[\ell]$ for $\ell = 1, 2, 3, 4$ are

$$\begin{aligned}\widehat{\alpha}[0] &= \widehat{e}[1] - 2\widehat{e}[0],\\ \widehat{\alpha}[1] &= 4\widehat{e}[0],\\ \widehat{\alpha}[2] &= \widehat{\alpha}[3] = 0,\end{aligned} \tag{12a}$$

and

$$\begin{aligned}\widehat{\alpha}[0] &= \widehat{e}[2] - 2\widehat{e}[1] + 2\widehat{e}[0],\\ \widehat{\alpha}[1] &= 4\widehat{e}[1] - 16\widehat{e}[0],\\ \widehat{\alpha}[2] &= 16\widehat{e}[0],\\ \widehat{\alpha}[3] &= 0,\end{aligned} \tag{12b}$$

respectively.

Step 4: Form the polynomial $\widehat{R}(\delta, L, x)$, which oscillates on $[\cos(\Omega_p(\delta, L)/2), 1]$ within $1/2 - \delta$ and $1/2 + \delta$ achieving these values at $L+2$ points so that $\widehat{R}(\delta, L, \cos(\Omega_p(\delta, L)/2)) = 1/2 - \delta$, as

$$\begin{aligned}\widehat{R}(\delta, L, x) &= E(\Omega)_{|\cos(\Omega/2) = x}\\ &= x\big\{\widehat{\alpha}(\delta, L)[0] + \widehat{\alpha}(\delta, L)[1]x^2 +\\ &\quad \widehat{\alpha}(\delta, L)[2]x^4 + \widehat{\alpha}(\delta, L)[3]x^6\big\},\end{aligned} \tag{13a}$$

where the following notations:

$$\widehat{\alpha}(\delta, L)[\ell] \equiv \widehat{\alpha}[\ell] \quad \text{for} \quad \ell = 0, 1, \ldots, L \tag{13b}$$

are used for both emphasizing the dependence of the coefficient values on δ and L and clarifying the description of the proposed overall quantization scheme in Section III.

Step 5: The above polynomial is the desired one with the exception that the interval average is given by

$$\widehat{x}_{\text{ave}}(\delta, L) = [1 + \cos(\Omega_p(\delta, L)/2)]/2. \tag{14}$$

Hence, the desired polynomial $R(x_{\text{ave}}, \delta, L, x)$, as given by (6), is obtained by replacing x in $\widehat{R}(\delta, L, x)$ by $x = \gamma x$, where

$$\gamma = x_{\text{ave}}/\widehat{x}_{\text{ave}}. \tag{15}$$

The resulting $R(x_{\text{ave}}, \delta, L, x)$ has thus the following coefficient values:

$$\alpha(x_{\text{ave}}, \delta, L)[\ell] = \widehat{\alpha}(\delta, L)[\ell]/\gamma^{2\ell+1} \tag{16}$$

for $\ell = 1, 2, \ldots, L$. Furthermore, the resulting lower edge $x^{(\text{low})}(x_{\text{ave}}, \delta, L)$ and the upper edge $x^{(\text{up})}(x_{\text{ave}}, \delta, L)$ of the above-mentioned maximized interval for $R(x_{\text{ave}}, \delta, L, x)$ are given by (7a) and (7b), respectively, where

$$\Delta(x_{\text{ave}}, \delta, L) = \gamma[1 - \cos(\Omega_p(\delta, L)/2)]/2. \tag{17}$$

E. Tool II: Use of the Center of the Polynomial Interval as the Second Degree of Freedom

Fact III of Subsection II-C leads to the insight, where x_{ave} is used as the second degree of freedom, in addition to the first degree of freedom proposed in the Introduction, which is to split the "effective" tap coefficients as $\alpha[\ell] = \kappa a[\ell]$ when searching for the desired signed powers-of-two terms for κ and $a[\ell]$ for $\ell = 0, 1, \ldots, L$. Using x_{ave} as the second degree of freedom is extremely crucial in the overall quantization scheme to be described in Section III as for each value of x_{ave}, the infinite-precision start-up values of $\alpha(x_{\text{ave}}, \delta, L)[\ell] = \kappa a[\ell]$ for $\ell = 0, 1, \ldots, L$ have their own specific values. Consequently, when varying the value of x_{ave} between $2/3$ and $4/3$ [3] with a small increment, there is a high probability to find values of x_{ave} for which the following goal is achieved. The desired multiplication-free values for κ and the $a[\ell]$'s can be found such that there exist an interval, denoted by $[x^{(\text{low})}_{\text{quan}}(x_{\text{ave}}, \delta, L), x^{(\text{up})}_{\text{quan}}(x_{\text{ave}}, \delta, L)]$, where the resulting polynomial with quantized coefficient values stays within the limits $1/2 \pm \delta$, so that the lower limit $x^{(\text{low})}_{\text{quan}}(x_{\text{ave}}, \delta, L)$ and the upper limit $x^{(\text{up})}_{\text{quan}}(x_{\text{ave}}, \delta, L)$ are simultaneously very close to the minimum achievable limit $x^{(\text{low})}(x_{\text{ave}}, \delta, L)$ and the maximum achievable limit $x^{(\text{up})}(x_{\text{ave}}, \delta, L)$. Of course, this implies that the number of fractional bits and the number of powers-of-two terms is high enough for representing κ and the $a[\ell]$'s in the desired forms. In these cases, the demands of the subfilter are attained well near mildest. As a matter of fact, the quantity

$$\beta = \epsilon_{\text{quan}}(x_{\text{ave}}, \delta, L)/\epsilon_{\text{ideal}}(x_{\text{ave}}, \delta, L), \tag{18a}$$

where

$$\epsilon_{\text{ideal}}(x_{\text{ave}}, \delta, L) = \frac{x^{(\text{up})}(x_{\text{ave}}, \delta, L) - x^{(\text{low})}(x_{\text{ave}}, \delta, L)}{x_{\text{ave}}} \tag{18b}$$

$$\epsilon_{\text{quan}}(x_{\text{ave}}, \delta, L) = \frac{x^{(\text{up})}_{\text{quan}}(x_{\text{ave}}, \delta, L) - x^{(\text{low})}_{\text{quan}}(x_{\text{ave}}, \delta, L)}{\widetilde{x}_{\text{ave}}(x_{\text{ave}}, \delta, L)} \tag{18c}$$

with

$$\widetilde{x}_{\text{ave}}(x_{\text{ave}}, \delta, L) = [x^{(\text{up})}_{\text{quan}}(x_{\text{ave}}, \delta, L) + x^{(\text{low})}_{\text{quan}}(x_{\text{ave}}, \delta, L)]/2 \tag{18d}$$

can be considered as a measure which determines the demands of the subfilter in the two following ways. First, due to Fact III of Subsection II-C, β directly sets for the infinite-precision subfilter its minimum order as well as determines how well the minimum-order subfilter exceeds its specifications. This quantity is independent of the value of x_{ave} because the multiplication of x_{ave} with the constant ξ also signifies the multiplication of subfilter coefficients with ξ. Second, when finite-precision solutions are examined, the bigger value of β increases the possibility to find the subfilter which meets its specifications with few signed powers-of-two coefficient values. For finite-precision designs, the number of fractional bits becomes smaller (larger) with the larger (smaller) value of x_{ave} roughly so that the enlarging (reducing) of x_{ave} by a factor of two will reduce (increase) the number of the bits by one. In the cases, where x_{ave} is divided (multiplied) by two, both the infinite-precision and finite-precision filters remain the same with the exception that all the subfilter coefficient values are expanded (reduced) by a factor of two.

F. Key Observation I: Dependence of the Tap Coefficient Values on the Allowable Passband Variation

When examining possibilities to develop a well-functioning algorithm in Section III, the following observation is

[3] Only one octave of the values of x_{ave} needs to be seached for because multiplying or dividing the values of x_{ave} leads to the same few signed powers-of-two representations with the exception that for the multiplication and division the number of fractional bits is decreased and increased by one, respectively.

very crucial. When comparing the coefficient values of the following three optimized infinite-precision polynomials, namely, $R(x_{\text{ave}}, \gamma_1\delta, L, x)$ with $\gamma_1 < 1$, $R(x_{\text{ave}}, \delta, L, x)$, and $R(x_{\text{ave}}, \gamma_2\delta, L, x)$ with $\gamma_2 > 1$, it has turned out that among their coefficient values the following fact

$$\alpha(x_{\text{ave}}, \gamma_1\delta, L)[\ell] < \alpha(x_{\text{ave}}, \delta, L)[\ell] < \alpha(x_{\text{ave}}, \gamma_2\delta, L)[\ell] \quad (19)$$

for $\ell = 0, 1, \ldots, L$ is valid at least up to $L = 3$. It should be pointed out that without this observation the problem of generating signed powers-of-two values for κ and the $a[\ell]$'s such that the resulting "effective" tap coefficient values $\widehat{\alpha}(x_{\text{ave}}, \delta, L)[\ell]$ for $\ell = 0, 1, \ldots, L$ are in the vicinity of their infinite-precision values $\alpha(x_{\text{ave}}, \delta, L)[\ell]$ would not have been solved at all. How to utilize this fact will be explained in more detail in the next section, where the proposed quantization scheme is described.

III. Proposed Overall Quantization Scheme

This section describes the quantization scheme for finding out a few signed powers-of-two terms for κ as well as for $a[\ell]$ for $\ell = 0, 1, \ldots, L$

Given L, the passband variation δ, γ_1, and γ_2, as well as all the desired coefficient representation forms, that is, $B_{\text{frac}}^{(a)}$ and $B_{\text{frac}}^{(\kappa)}$, the number of fractional bits for the $a[\ell]$'s and κ, respectively, $P^{(a)}$ and $P^{(\kappa)}$, the number of powers-of-two terms for the $a[\ell]$'s and κ, respectively, along with $x_{\text{ave}}^{(\text{low})}$, the start-up value for x_{ave}, Δx_{ave}, the increment of x_{ave} used in the procedure, and τ as a fraction of β, the optimized finite-wordlength solutions can be found efficiently using the following procedure:

Step 1: Determine the vector $\Gamma_{\text{CSD}}^{(\kappa)}$ containing all the signed powers-of-two values with $B_{\text{frac}}^{(\kappa)}$ fractional bits and $P^{(\kappa)}$ powers-of-two terms between $1/2$ and 1. Denote by S the number of elements in $\Gamma_{\text{CSD}}^{(\kappa)}$. Furthermore, denote by $\Gamma_{\text{CSD}}^{(\kappa)}(s)$ the sth existing discrete value in $\Gamma_{\text{CSD}}^{(\kappa)}$.

Step 2: Perform Steps 1–4 of Subsection II-D for the given value of L and for the passband variations of $\gamma_1\delta$, δ, and $\gamma_2\delta$. Store the coefficient values for these two variations for later use as $\widehat{\alpha}(\gamma_1\delta, L)[\ell]$ and $\widehat{\alpha}(\gamma_2\delta, L)[\ell]$ for $\ell = 0, 1, \ldots, L$. In addition, store $\Omega_p(\gamma_1\delta, L)$, $\Omega_p(\delta, L)$, and $\Omega_p(\gamma_2\delta, L)$ as well as

$$\widehat{x}_{\text{ave}}(\gamma_1\delta, L) = \frac{1}{2}\left(1 + \cos[\Omega_p(\gamma_1\delta, L)]\right), \quad (20a)$$

$$\widehat{x}_{\text{ave}}(\delta, L) = \frac{1}{2}(1 + \cos[\Omega_p(\delta, L)]), \quad (20b)$$

and

$$\widehat{x}_{\text{ave}}(\gamma_2\delta, L) = \frac{1}{2}(1 + \cos[\Omega_p(\gamma_2\delta, L)]). \quad (20c)$$

Furthermore, set $x_{\text{ave}} = x_{\text{ave}}^{(\text{low})} - \Delta x_{\text{ave}}$.

Step 3: Set $s = 0$ and update $x_{\text{ave}} = x_{\text{ave}} + \Delta x_{\text{ave}}$. If $x_{\text{ave}} > 2x_{\text{ave}}^{(\text{low})} - \Delta x_{\text{ave}}$, then stop. Otherwise, go to the next step.

Step 4: Calculate

$$\alpha(x_{\text{ave}}, \gamma_1\delta, L)[\ell] = \widehat{\alpha}(\gamma_1\delta, L)[\ell]\left[\frac{\widehat{x}_{\text{ave}}(\gamma_1\delta, L)}{x_{\text{ave}}}\right]^{(2\ell+1)} \quad (21a)$$

and

$$\alpha(x_{\text{ave}}, \gamma_2\delta, L)[\ell] = \widehat{\alpha}(\gamma_2\delta, L)[\ell]\left[\frac{\widehat{x}_{\text{ave}}(\gamma_2\delta, L)}{x_{\text{ave}}}\right]^{(2\ell+1)} \quad (21b)$$

for $\ell = 0, 1, \ldots, L$.

Step 5: Determine

$$\Psi^{(\text{low})}(x_{\text{ave}}) = x_{\text{ave}} - \psi(x_{\text{ave}}, \delta, L) \quad (22a)$$

and

$$\Psi^{(\text{up})}(x_{\text{ave}}) = x_{\text{ave}} + \psi(x_{\text{ave}}, \delta, L), \quad (22b)$$

where

$$\psi(x_{\text{ave}}, \delta, L) = \frac{(1 - \cos[\Omega_p(\delta, L)])x_{\text{ave}}}{2\widehat{x}_{\text{ave}}(\delta, L)} \quad (22c)$$

as the lower and upper limits for the maximum achievable interval for the given x_{ave}.

Step 6: Determine

$$B_{\text{int}} = \max\left(1, \max_{\ell=0,1,\ldots,L}\left\lceil |\widehat{\alpha}(\gamma_2\delta, L)[\ell]| \right\rceil\right). \quad (23)$$

Step 7: Determine the vector $\Gamma_{\text{CSD}}^{(a)}$ containing all the signed powers-of-two values with $B_{\text{frac}}^{(a)}$ fractional bits and $P^{(a)}$ powers-of-two terms between $-2^{B_{\text{int}}}$ and $2^{B_{\text{int}}}$. Denote by $\Gamma_{\text{CSD}}^{(a)}(i)$ the ith existing discrete value in $\Gamma_{\text{CSD}}^{(a)}$.

Step 8: Update $s = s + 1$. If $s > S$, then go to Step 3. Otherwise, go to the next step.

Step 9: Select $\kappa = \Gamma_{\text{CSD}}^{(\kappa)}(s)$ and determine

$$a(x_{\text{ave}}, \gamma_1)[\ell] = \frac{1}{\kappa}\alpha(x_{\text{ave}}, \gamma_1\delta, L)[\ell] \quad (24a)$$

and

$$a(x_{\text{ave}}, \gamma_2)[\ell] = \frac{1}{\kappa}\alpha(x_{\text{ave}}, \gamma_2\delta, L)[\ell]. \quad (24b)$$

Step 10: Denote by $i^{(\text{low})}(\ell)$ and $i^{(\text{up})}(\ell)$ the index of the smallest and largest signed power-of-two values in $\Gamma_{\text{CSD}}^{(a)}$ that exist between $a(x_{\text{ave}}, \gamma_1)[\ell]$ and $a(x_{\text{ave}}, \gamma_2)[\ell]$ for $\ell = 0, 1, \ldots, L$. If for one or more values of ℓ such indices do not exit, then go to Step 8. Otherwise, go to the next step.

Step 11: Check whether there exist a coefficient value combination or combinations between $\kappa\Gamma_{\text{CSD}}^{(a)}(i^{(\text{low})}(\ell))$ and $\kappa\Gamma_{\text{CSD}}^{(a)}(i^{(\text{up})}(\ell))$ for $\ell = 0, 1, \ldots, L$ such that the corresponding polynomial stays within $1/2 \pm \delta$ on the subinterval $[x_{\text{ave}} - \tau\psi(x_{\text{ave}}, \delta, L), x_{\text{ave}} + \tau\psi(x_{\text{ave}}, \delta, L)]$ of $[\Psi^{(\text{low})}(x_{\text{ave}}), \Psi^{(\text{up})}(x_{\text{ave}})]$. If there is not such a combination, then go to Step 3. Otherwise, estimate for each successful coefficient value combination β by determining the portion of $[\Psi^{(\text{low})}(x_{\text{ave}}), \Psi^{(\text{up})}(x_{\text{ave}})]$, where the polynomial stays within $1/2 \pm \delta$. For each combination, store the coefficient value combination as well as β and x_{ave} into the memory for further study. Go to Step 3.

The above algorithm has been used as follows. First, x_{ave} varies from $x_{\text{ave}} = x_{\text{ave}}^{(\text{low})} = 2/3$ with the increment $\Delta x_{\text{ave}} = 0.0001$ up to $x_{\text{ave}} = 2x_{\text{ave}}^{(\text{low})} - \Delta x_{\text{ave}} = 4/3 - 0.0001$. For γ_1 and γ_2, proper values have turned out to be 10^{-6} and 10^{6}; 10^{-3} and 10^{3}; and 10^{-1} and 10^{1} for $L = 1$, $L = 2$,

and $L = 3$, respectively. A good value for τ is 0.2 in order to collect enough data for making compromises between the multiplication-free tap coefficients and both the multiplication-free coefficients and the order of the subfilter.

IV. Use of the Proposed Quantization Scheme in Practical Filter Design and an Illustrative Example

The purpose of this section is two-fold. First, it is shown how to properly use the proposed quantization algorithm to form tabulated data in order to make proper compromises between the multiplication-free coefficient representations of the tap coefficients and those of the subfilter. Second, an example is included to illustrate the superiority of the multiplication-free overall filters generated based on the tabulated data over those ones achievable using the earlier approach [3], [10] and their direct-form FIR equivalents.

A. Tabulated Data for a 120-dB Attenuation and $L = 3$

Table I shows the maximum achievable value of β for polynomials with $L = 3$ and providing at least a 120-dB attenuation, that is, $\delta \leq 10^{-6}$, in the three powers-of-two representations of tap coefficients for the number of fractional bits ranging from 4 to 19. In addition, x_{ave}, the center of the interval, as well as $\Delta_{\text{ave}}^{(x)}$, the maximum absolute deviation from x_{ave}, are given for these coefficient representation forms. For these criteria, the maximized infinite-precision value of β is $0.043\,722\,83 x_{\text{ave}}$ and 19 fractional bits is required to achieve at least 95 percent of this value. In addition, the table contains for each number of fractional bits the minimum subfilter order along with the number of fractional bits in the three powers-of-two representation required to meet the overall criteria to be considered in the following subsection. Furthermore, the overall order of $G(z)$ is included. Figure 4 shows the polynomials for the optimized infinite-precision design as well as finite-precision designs for 4, 13, 17, and 19 fractional bits.

B. Illustrative Example

In [3], an efficient multiplication-free decimator has been generated without any general multipliers after a sigma-delta modulator in such a manner that the overall analog-to-digital converter has a 20-bit resolution. In this decimator, there are altogether three linear-phase half-band filters. This example concentrates on the design of the one having the most stringent criteria.

It is desired to design a half-band decimator in such a way that the sampling rate reduction ratio is two, the output sampling rate is 44.1 kHz, and the components aliasing into the band from 0 Hz to 20 kHz are attenuated at least 120 dB. In this case, the problem is to design $G(z)$ such that the deviation of its zero-phase frequency response $G(\omega)$ from $1/2$ is at most 10^{-6} in the passband with edge angle being $2\omega_p = [20/(44.1/2)]\pi = 0.907\,029\,48\pi$.

According to Table I, the design with 14 fractional bits for the tap coefficients is the best among filters for which the order of $G(z)$ is 119, in terms of implementation complexity of these coefficients. Also the design with 13 fractional bits is worth studying because the number of fractional bits for the subfilter coefficients is only 7, at the expense of a minor increase in the order of $G(z)$. Tables II and III show the optimized finite-precision coefficients for these two cases.

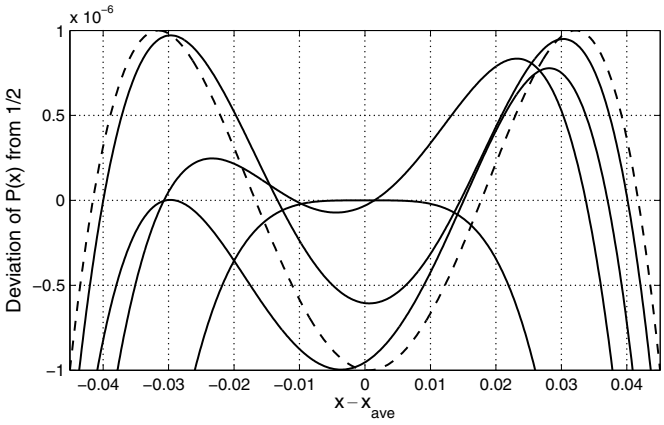

Fig. 4. The relations of the deviation of $P(x)$ from $1/2$ to the deviation of x from x_{ave} for $L = 3$ and a 120-dB attenuation. The dashed line shows the response for the infinite-precision design whereas the solid lines show the responses for the 4, 13, 17, and 19 bit designs.

TABLE I
Summary of Filter Designs with Various Word-Lengths

$B_{\text{frac}}^{(\kappa)}$	x_{ave}	$\Delta_{\text{ave}}^{(x)}$	β	K	$B_{\text{frac}}^{(a)}$	$(2L+1)K$
$4-11$	0.999 797	0.026 006	0.594 903	21	8	147
12	1.027 723	0.027 420	0.610 214	21	8	147
13	1.033 424	0.037 815	0.836 911	19	7	133
14	1.032 256	0.039 794	0.881 695	17	10	119
$15-16$	1.034 291	0.040 415	0.893 698	17	9	119
$17-18$	1.034 537	0.040 899	0.904 183	17	9	119
19	1.031 257	0.043 666	0.968 425	17	8	119

TABLE II
Optimized Coefficient Values for the Multiplication-Free Design in the Illustrative Example with $\kappa = 2^{-1} + 2^{-14}$

$a[0] = \ \ 2 + 2^{-3} - 2^{-12}$	$a[2] = \ \ 1 + 2^{-3} + 2^{-7}$
$a[1] = -2 - 2^{-8} + 2^{-13}$	$a[3] = -2^{-2} - 2^{-8}$
$f[0] = f[17] = \ \ 2^{-5} - 2^{-8} - 2^{-10}$	$f[5] = f[12] = -2^{-4} - 2^{-6} - 2^{-9}$
$f[1] = f[16] = -2^{-6} - 2^{-8} - 2^{-10}$	$f[6] = f[11] = \ \ 2^{-3} - 2^{-8} + 2^{-10}$
$f[2] = f[15] = \ \ 2^{-5}$	$f[7] = f[10] = -2^{-2} + 2^{-5} + 2^{-8}$
$f[3] = f[14] = -2^{-4} + 2^{-6} + 2^{-8}$	$f[8] = f[9] \ \ = \ \ 2^{-1} + 2^{-3} + 2^{-5}$
$f[4] = f[13] = \ \ 2^{-4} - 2^{-7} + 2^{-9}$	

TABLE III
Optimized Coefficient Values for the Multiplication-Free Design in the Illustrative Example with $\kappa = 1 - 2^{-10} + 2^{-12}$

$a[0] = \ \ 1 + 2^{-4} - 2^{-10}$	$a[2] = \ \ 2^{-1} + 2^{-4}$
$a[1] = -1 + 2^{-9} + 2^{-13}$	$a[3] = -2^{-3} - 2^{-10} + 2^{-12}$
$f[0] = f[19] = -2^{-6}$	$f[5] = f[14] = \ \ 2^{-4} - 2^{-7}$
$f[1] = f[18] = \ \ 2^{-6}$	$f[6] = f[13] = -2^{-4} - 2^{-6}$
$f[2] = f[17] = -2^{-5} + 2^{-7}$	$f[7] = f[12] = \ \ 2^{-3}$
$f[3] = f[16] = \ \ 2^{-5}$	$f[8] = f[11] = -2^{-2} + 2^{-5}$
$f[4] = f[15] = -2^{-4} + 2^{-6}$	$f[9] = f[10] = \ \ 2^{-1} + 2^{-3} + 2^{-5}$

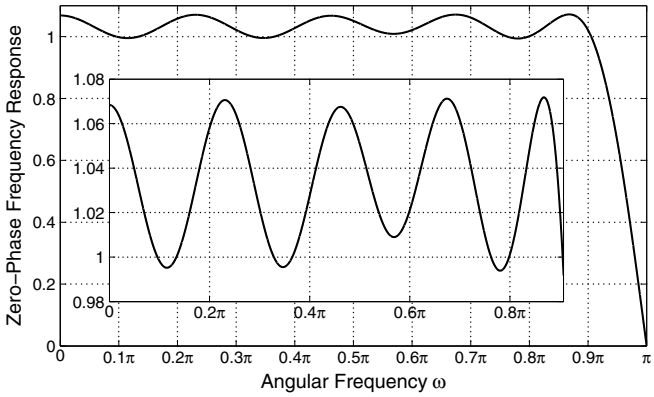

Fig. 5. Zero-phase frequency response as well as the passband details for the optimized finite-precision $F(z)$ in Illustrative Example.

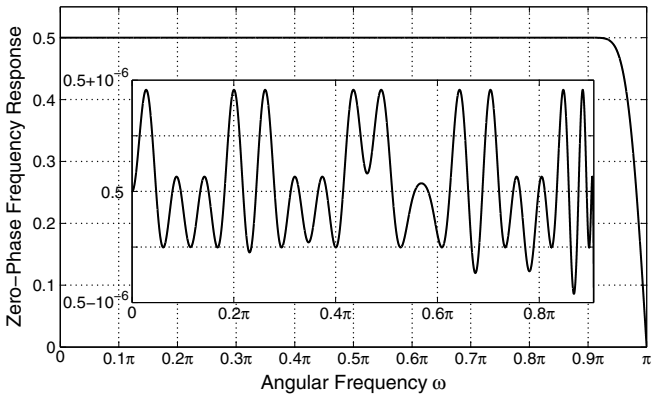

Fig. 6. Zero-phase frequency response as well as the passband details for the optimized finite-precision $G(z)$ in the Illustrative Example.

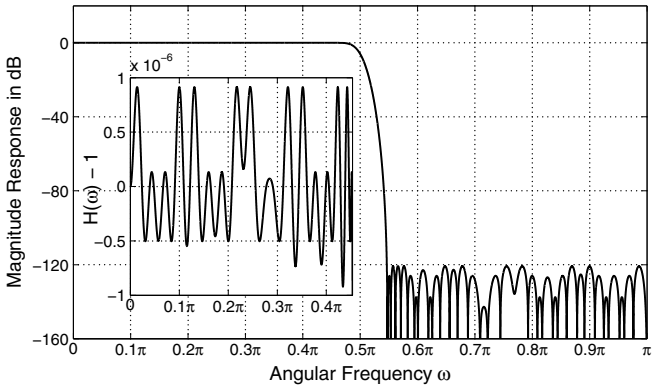

Fig. 7. Magnitude response as well as the passband details for the optimized finite-precision half-band FIR filter $H(z)$ in the Illustrative Example.

For the Table II design, Figs. 5, 6, and 7 show the characteristics of $F(z)$, $G(z)$, and the resulting overall half-band FIR filter $H(z)$ with passband and stopband edges at $\omega_p = 0.453515\pi$ and $\omega_s = 0.546485\pi$, respectively, and a 120-dB stopband attenuation. In Table I, the $4 - 11$ fractional bits case corresponds to the design achievable using the approach proposed in [10].

For comparing the proposed design with the direct-form half-band filters, the coefficient values of these filters with various orders were rounded to the minimum number of fractional bits to meet the criteria. Table IV summarizes the results by showing the minimum value of M, the order of $G(z)$ [cf. (1b) and (1c)], required for 23, 24, 25, and 26 fractional bit representations. When using rounding, 23 factional bits is the limit for achieving the given criteria.

TABLE IV
MINIMUM VALUE OF M, THE ORDER OF $G(z)$, TO MEET THE STATED CRITERIA WITH VARIOUS NUMBERS OF FRACTIONAL BITS

Number of fractional bits	22	23	24	25	26
M	–	91	87	87	85

V. CONCLUSION

This paper provided a substantial improved technique in comparison to the approach described in [10] for the design of highly selective multiplication-free FIR half-band decimators and interpolators. As in the earlier approach, the overall transfer function is first split into the sum of two terms $(1/2)z^{-M}$ and $G(z^2)$, where the odd integer M is the order of $G(z)$, and, then, $G(z)$ is constructed as a special tapped cascaded interconnection of identical subfilters. The proposed scheme to generate multiplication-free tap coefficients and the multiplication-free subfilters was shown to be more efficient, in terms of reduced subfilter orders. In addition it turned out to be very flexible in providing several compromises between the multiplication-free tap coefficients and multiplication-free subfilters, among which the user can select the proper one.

REFERENCES

[1] T. Saramäki, "Finite impulse response filter design," in *Handbook for Digital Signal Processing*, S. K. Mitra and J. F. Kaiser, Eds. New York: John Wiley and Sons, 1993, ch. 4, pp. 155–277.

[2] L. Rabiner and D. Crochiere, *Multirate Digital Signal Processing*. Prentice-Hall, 1983.

[3] T. Saramäki, T. Karema, T. Ritoniemi, and H. Tenhunen, "Multiplier-free decimator algorithms for superresolution oversampled converters," in *Proc. IEEE Int. Symp. Circuits Syst.*, New Orleans, LA, May 1–3 1990, pp. 3275–3278.

[4] B. Leung, "The oversampling technique for analog to digital conversion: A tutorial overview," *J. Analog Integrated Circuits, Signal Process.*, vol. 1, no. 1, Mar. 1991.

[5] L. Lin and T. Aboulnasr, "Adaptive signal processing in subbands using sigma-delta modulation technique," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 3, Adelaide, SA, Australia, Apr. 19–22 1994, pp. 169–172.

[6] V. Liberali, S. Brigati, F. Francesconi, and F. Maloberti, "Progress in high-speed and high-resolution CMOS data converters," in *Proc. 1995 20th Int. Conf. Microelectron.*, vol. 1, Nis, Serbia, Sept. 12–14 1995, pp. 19–28.

[7] K. Ichige and M. Kamada, "An approximation for discrete B-splines in time domain," *IEEE Signal Processing Lett.*, vol. 4, no. 3, pp. 82–84, Mar. 1997.

[8] L. Presti, "Efficient modified-sinc filters for sigma-delta A/D converters," *IEEE Trans. Circuits Syst. II*, vol. 47, no. 11, pp. 1204–1213, Nov. 2000.

[9] T. Asahi, K. Ichige, and R. Ishii, "A new formulation for discrete box splines reducing computational cost and its evaluation," *IEICE Trans. Fund.*, vol. E84-A, no. 3, pp. 884–892, Mar. 2001.

[10] T. Saramäki and J. Yli-Kaakinen, "A novel systematic approach for synthesizing multiplication-free highly-selective fir half-band decimators and interpolators," in *in Proc. IEEE Acia Pacific Conf. Circuits Syst.*, Singapore, Dec. 4–7 2006, pp. 922–925.

[11] The MathWorks, *Filter Design Toolbox User's Guide*, The MathWorks, Inc., Sept. 2009, Version 4.6.