PART II: Design of digital filters using identical subfilters as basic building blocks

- This part is very long since the lecturer has been studied intensively these filter structures.
- It is not worth trying to remember all the details by heart. For the examination, it is important to understand the basic ideas.
- The key idea behind the filter structures to be considered lies in the fact that we are able to design filters in such a way that no general multipliers are needed.
- Filters of this kind are very attractive for VLSI implementations where a general multiplier element is very costly.
- We start by considering results. Then, we concentrate on how to generate these results.

VLSI Realizable Digital Filter Structures

Tapio Saramäki and Tapani Ritoniemi

VLSI Solution Oy

Kanslerinkatu 6

SF–33720 Tampere

Finland

1. FIR Filters

2. IIR Filters

3. Decimators and Interpolators

BASIC QUESTION

QUESTION: How to drastically decrease the silicon area in VLSI implementations of digital filters?

ANSWER: Design the filters such that their implementation requires no general multipliers.

Why Multiplier-Free Filters?

- In VLSI-implementation, a general multiplier element is very costly:
  - Large silicon area
  - The delay of the multiplier sets an upper limit for the sampling rate.
  - For selective digital filters, several multipliers are required to achieve high sampling rate.
  - High power consumption

$\Rightarrow$

- It is advantageous to design the filter such that all the coefficient values are representable in the form

$$\pm 2^{-P_1} \pm 2^{-P_2}(\pm 2^{-P_3})$$

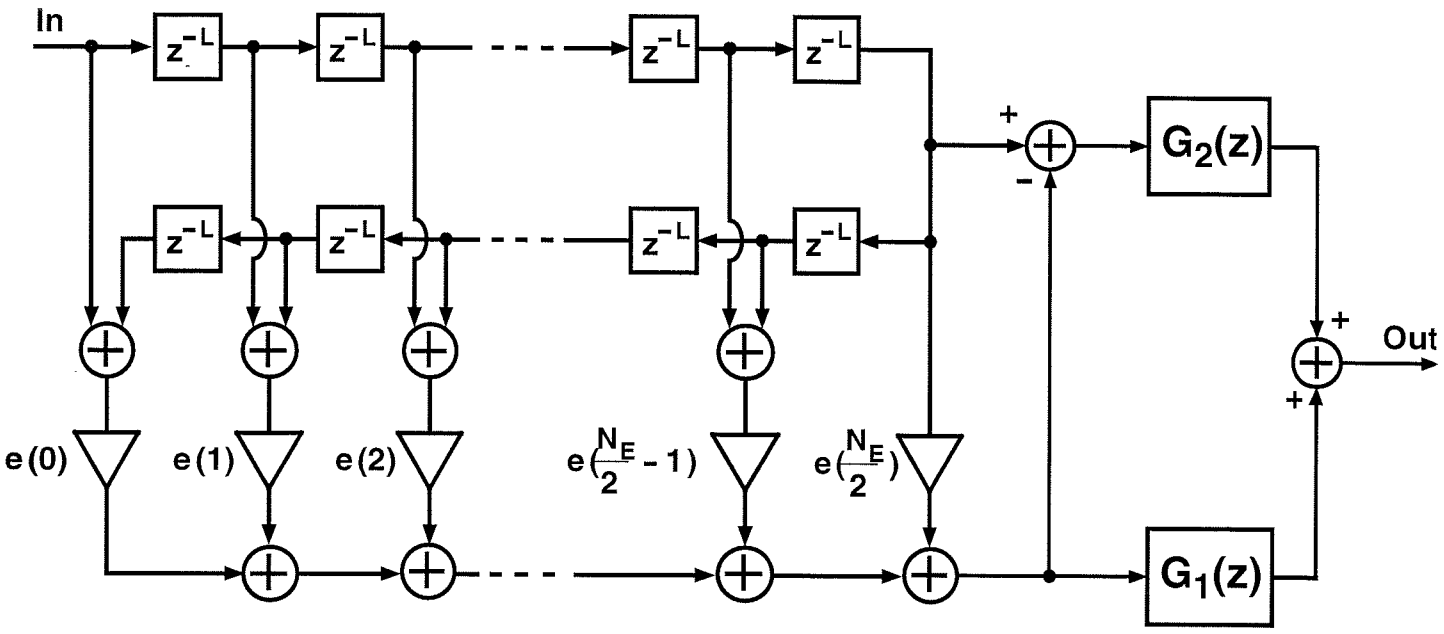
  - Only shifts and additions are required.

Starting point to achieve the desired result

- For specifications with large passband and stopband ripples, the design of desired filters is trivial.

- Examples
  - FIR filter: $\omega_p = 0.4\pi$, $\omega_s = 0.402\pi$, $\delta_p = 0.18$, $\delta_s = 0.12$ ($A_p = 3.2$ dB, $A_s = 18$ dB)
  - IIR filter: $\omega_p = 0.4\pi$, $\omega_s = 0.42\pi$, $A_p = 0.2$ dB, $A_s = 32$ dB
  - Half-band FIR filter: $\omega_p = 0.454\pi$, $\omega_s = 0.546\pi$, $\delta_p = \delta_s = 0.016$ ($A_s = 35$ dB)

FIR filter: $\omega_p = 0.4\pi$, $\omega_s = 0.402\pi$, $\delta_p = 0.18$, $\delta_s = 0.12$ ($A_p = 3.2$ dB, $A_s = 18$ dB)



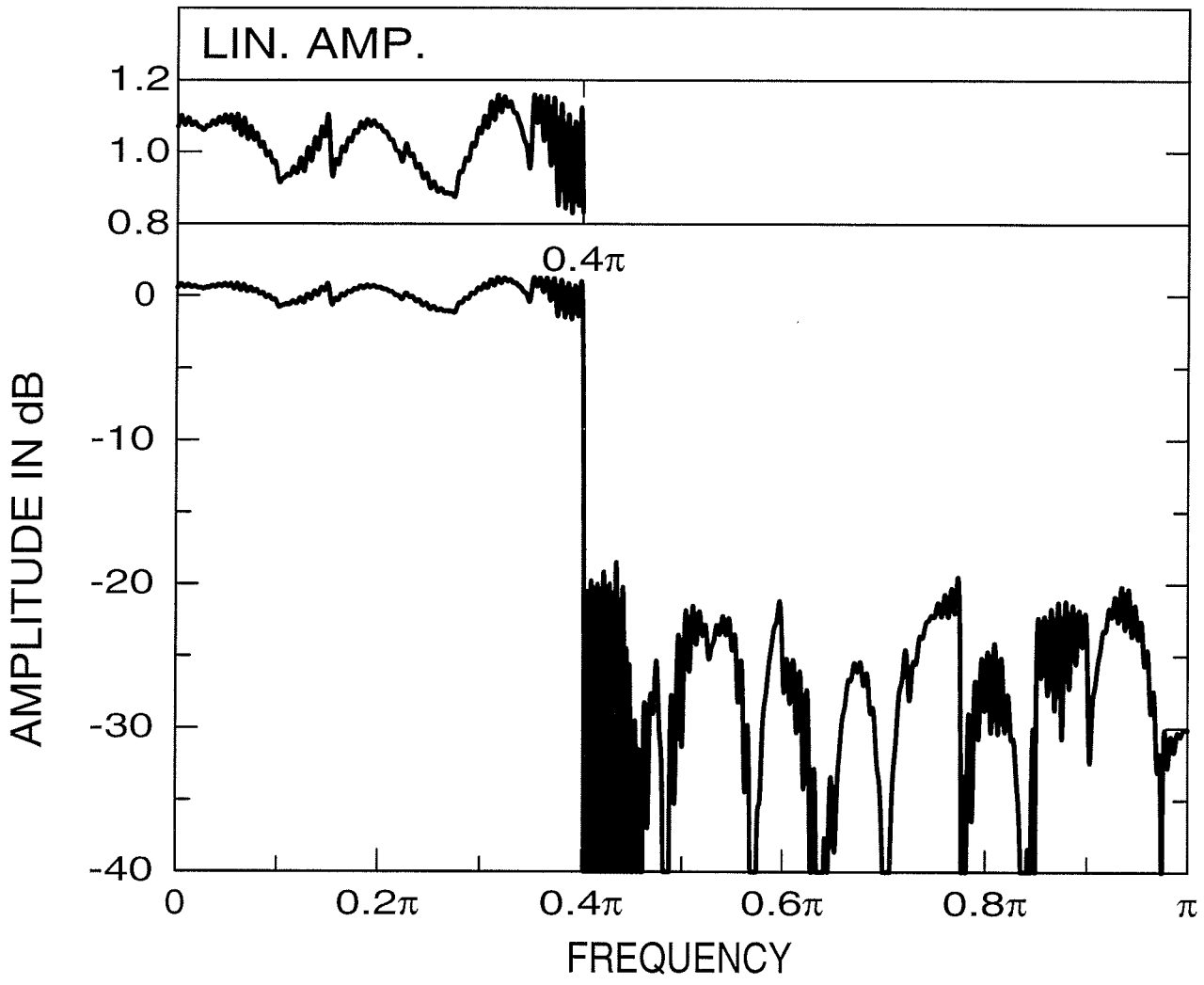
- $L = 16$, $N_E = 40$, $N_{G_1} = 22$, $N_{G_2} = 32$

$e(0) = 2 \cdot 2^{-6}$ $e(4) = 1 \cdot 2^{-6}$ $e(8) = 0$ $e(12) = -3 \cdot 2^{-6}$ $e(16) = -5 \cdot 2^{-6}$ $e(20) = 28 \cdot 2^{-6}$	$e(1) = -4 \cdot 2^{-6}$ $e(5) = 0$ $e(9) = 2 \cdot 2^{-6}$ $e(13) = 1 \cdot 2^{-6}$ $e(17) = -6 \cdot 2^{-6}$	$e(2) = -3 \cdot 2^{-6}$ $e(6) = -1 \cdot 2^{-6}$ $e(10) = 1 \cdot 2^{-6}$ $e(14) = 4 \cdot 2^{-6}$ $e(18) = 5 \cdot 2^{-6}$	$e(3) = -2 \cdot 2^{-6}$ $e(7) = -2 \cdot 2^{-6}$ $e(11) = -2 \cdot 2^{-6}$ $e(15) = 1 \cdot 2^{-6}$ $e(19) = 17 \cdot 2^{-6}$

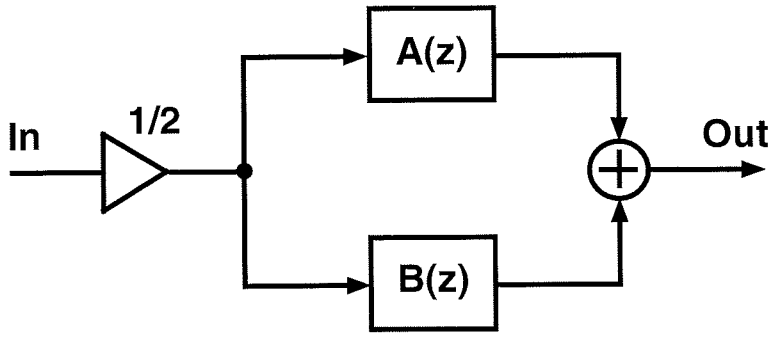
$g_1(0) = 2 \cdot 2^{-6}$ $g_1(4) = -1 \cdot 2^{-6}$ $g_1(8) = -5 \cdot 2^{-6}$	$g_1(1) = 3 \cdot 2^{-6}$ $g_1(5) = 3 \cdot 2^{-6}$ $g_1(9) = 4 \cdot 2^{-6}$	$g_1(2) = -1 \cdot 2^{-6}$ $g_1(6) = 2 \cdot 2^{-6}$ $g_1(10) = 20 \cdot 2^{-6}$	$g_1(3) = -2 \cdot 2^{-6}$ $g_1(7) = -4 \cdot 2^{-6}$ $g_1(11) = 28 \cdot 2^{-6}$

$g_2(0) = -3 \cdot 2^{-6}$ $g_2(4) = 1 \cdot 2^{-6}$ $g_2(8) = 3 \cdot 2^{-6}$ $g_2(12) = -2 \cdot 2^{-6}$	$g_2(1) = -1 \cdot 2^{-6}$ $g_2(5) = -1 \cdot 2^{-6}$ $g_2(9) = 3 \cdot 2^{-6}$ $g_2(13) = 7 \cdot 2^{-6}$	$g_2(2) = 1 \cdot 2^{-6}$ $g_2(6) = -2 \cdot 2^{-6}$ $g_2(10) = -1 \cdot 2^{-6}$ $g_2(14) = 19 \cdot 2^{-6}$	$g_2(3) = 2 \cdot 2^{-6}$ $g_2(7) = 0$ $g_2(11) = -5 \cdot 2^{-6}$ $g_2(15) = 24 \cdot 2^{-6}$

FIR filter: $\omega_p = 0.4\pi$, $\omega_s = 0.402\pi$, $\delta_p = 0.18$, $\delta_s = 0.12$ ($A_p = 3.2$ dB, $A_s = 18$ dB)



IIR filter: $\omega_p = 0.4\pi$, $\omega_s = 0.42\pi$, $A_p = 0.2$ dB, $A_s = 32$ dB



- $A(z)$ contains two second-order **allpass** sections with adaptor coefficients

$$\gamma_1 = -2^{-1} + 2^{-5} + 2^{-6}, \quad \gamma_2 = 2^{-1} - 2^{-6} - 2^{-7}$$

$$\gamma_1 = -2^{0} + 2^{-5} + 2^{-7}, \quad \gamma_2 = 2^{-2} + 2^{-5} + 2^{-6}.$$

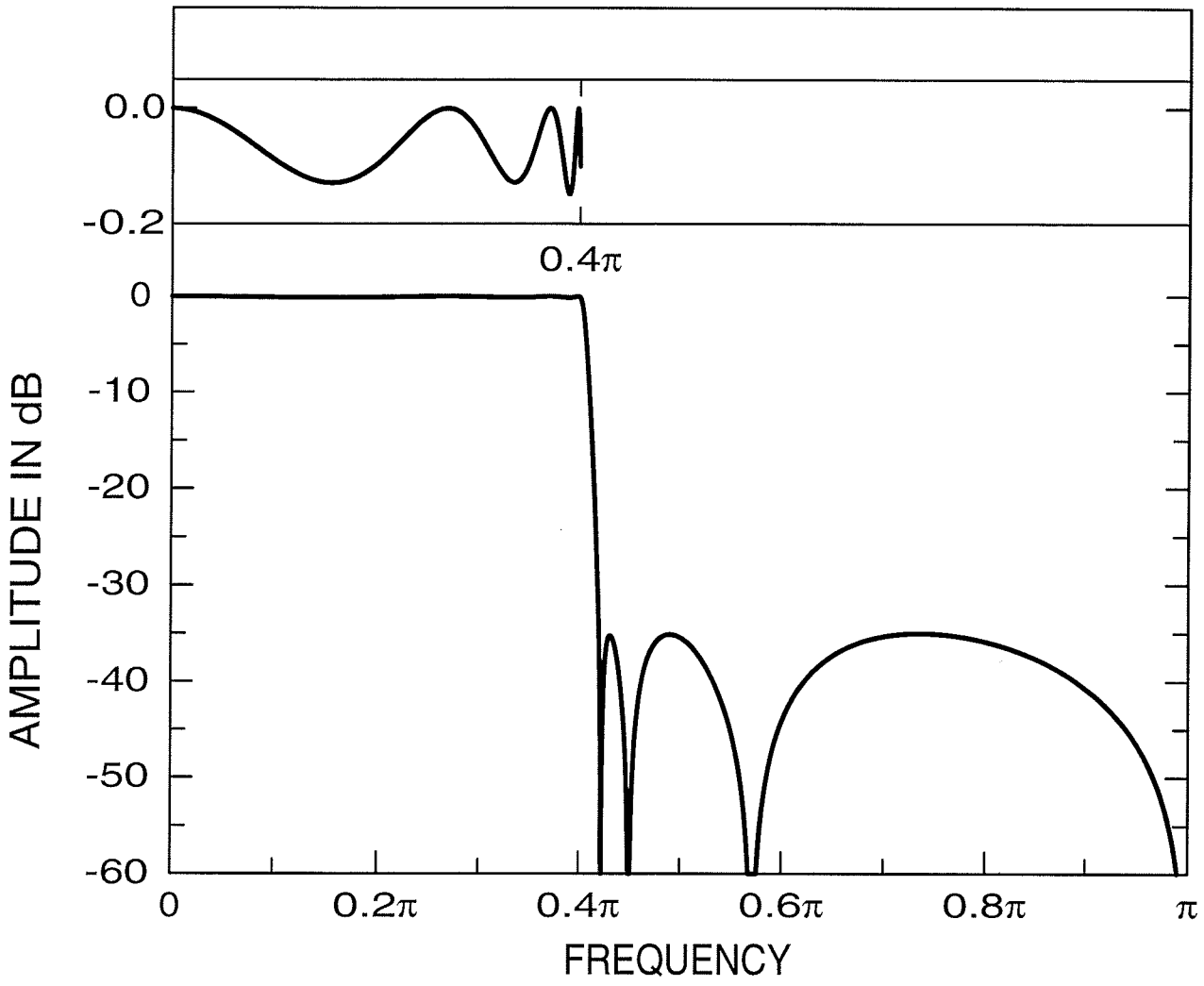
- $B(z)$ consists of one first-order section with

$$\gamma = 2^{-2} + 2^{-3} + 2^{-7}$$

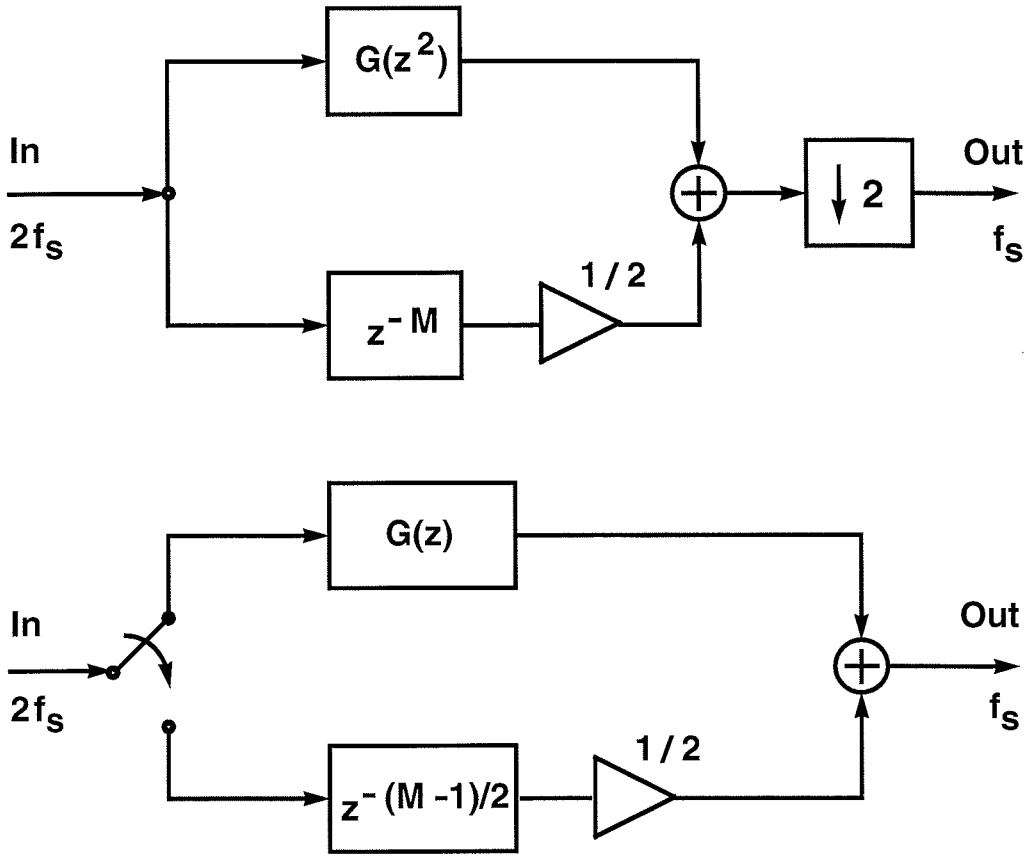
and one second-order section with

$$\gamma_1 = -2^{-1} - 2^{-2} - 2^{-4}, \quad \gamma_2 = 2^{-2} + 2^{-4} + 2^{-6} + 2^{-7}.$$

IIR filter: $\omega_p = 0.4\pi$, $\omega_s = 0.42\pi$, $A_p = 0.2$ dB, $A_s = 32$ dB

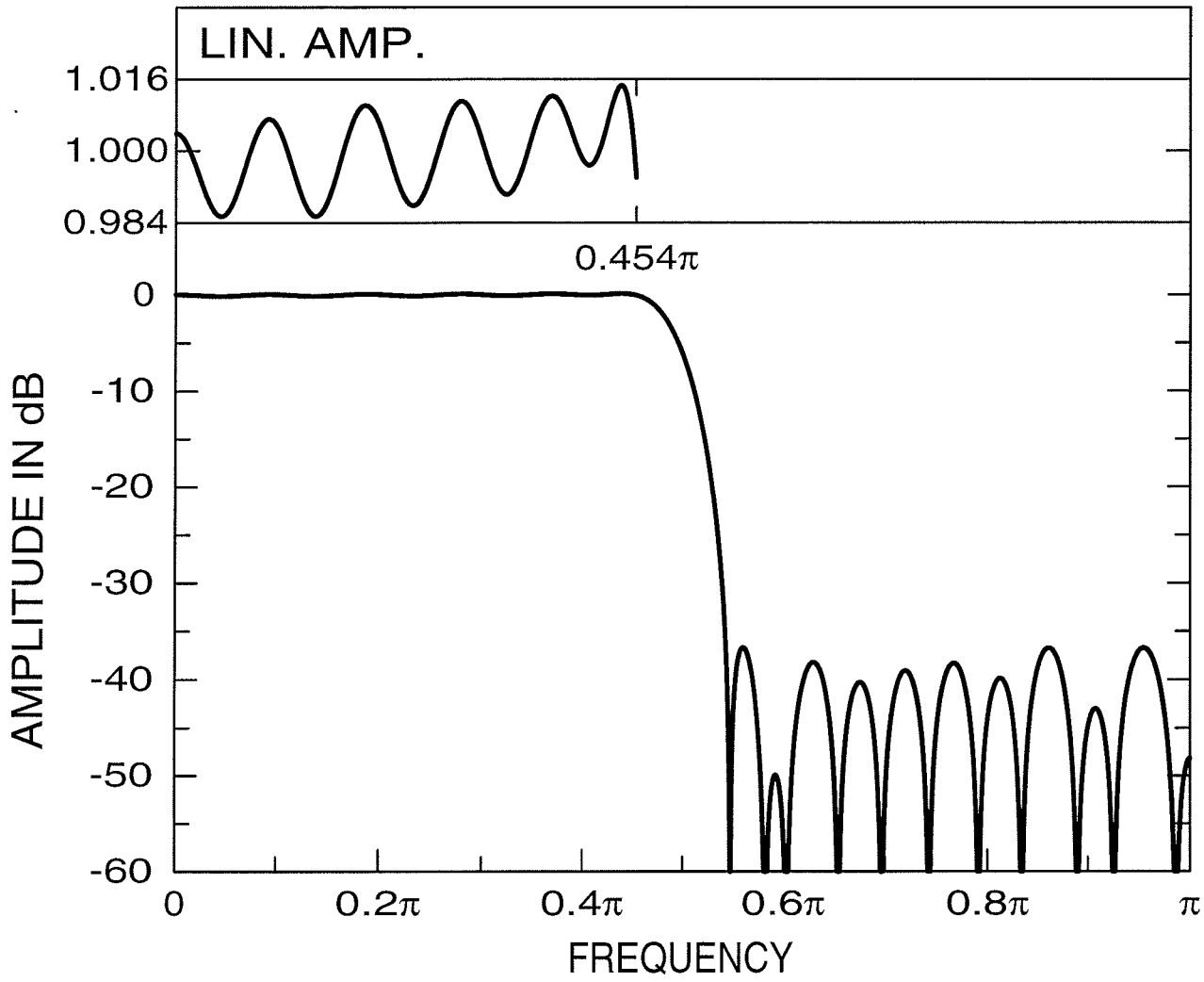


Half-band FIR filter: $\omega_p = 0.454\pi$, $\omega_s = 0.546\pi$, $\delta_p = \delta_s = 0.016$ $(A_s = 35$ dB$)$



$$G(z) = 2^{-6}(1 + z^{-21}) + (-2^{-7} - 2^{-8})(z^{-1} + z^{-20})$$
$$+ 2^{-6}(z^{-2} + z^{-19}) + (-2^{-6} - 2^{-7})(z^{-3} + z^{-18})$$
$$+ 2^{-5}(z^{-4} + z^{-17}) + (-2^{-5} - 2^{-7} - 2^{-8})(z^{-5} + z^{-16})$$
$$+ (2^{-4} - 2^{-8})(z^{-6} + z^{-15}) + (-2^{-4} - 2^{-6} - 2^{-8})(z^{-7} + z^{-14})$$
$$+ (2^{-3} - 2^{-8})(z^{-8} + z^{-13}) + (-2^{-2} + 2^{-5} + 2^{-7})(z^{-9} + z^{-12})$$
$$+ (2^{-1} + 2^{-3} + z^{-7})(z^{-10} + z^{-11})$$

Half-band FIR filter: $\omega_p = 0.454\pi$, $\omega_s = 0.546\pi$, $\delta_p = \delta_s = 0.016$ $(A_s = 35$ dB$)$

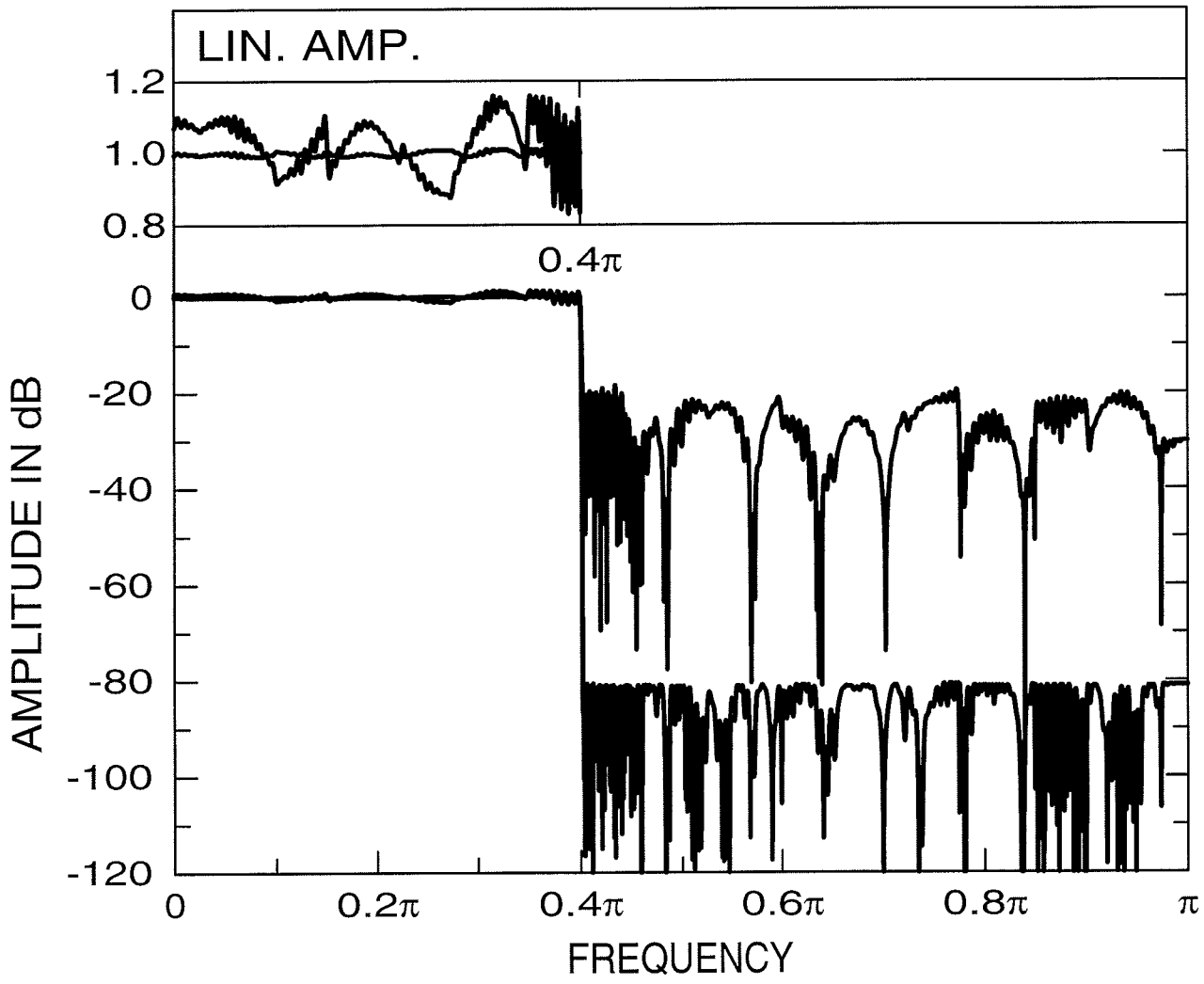


How to design highly selective filters without multipliers ?

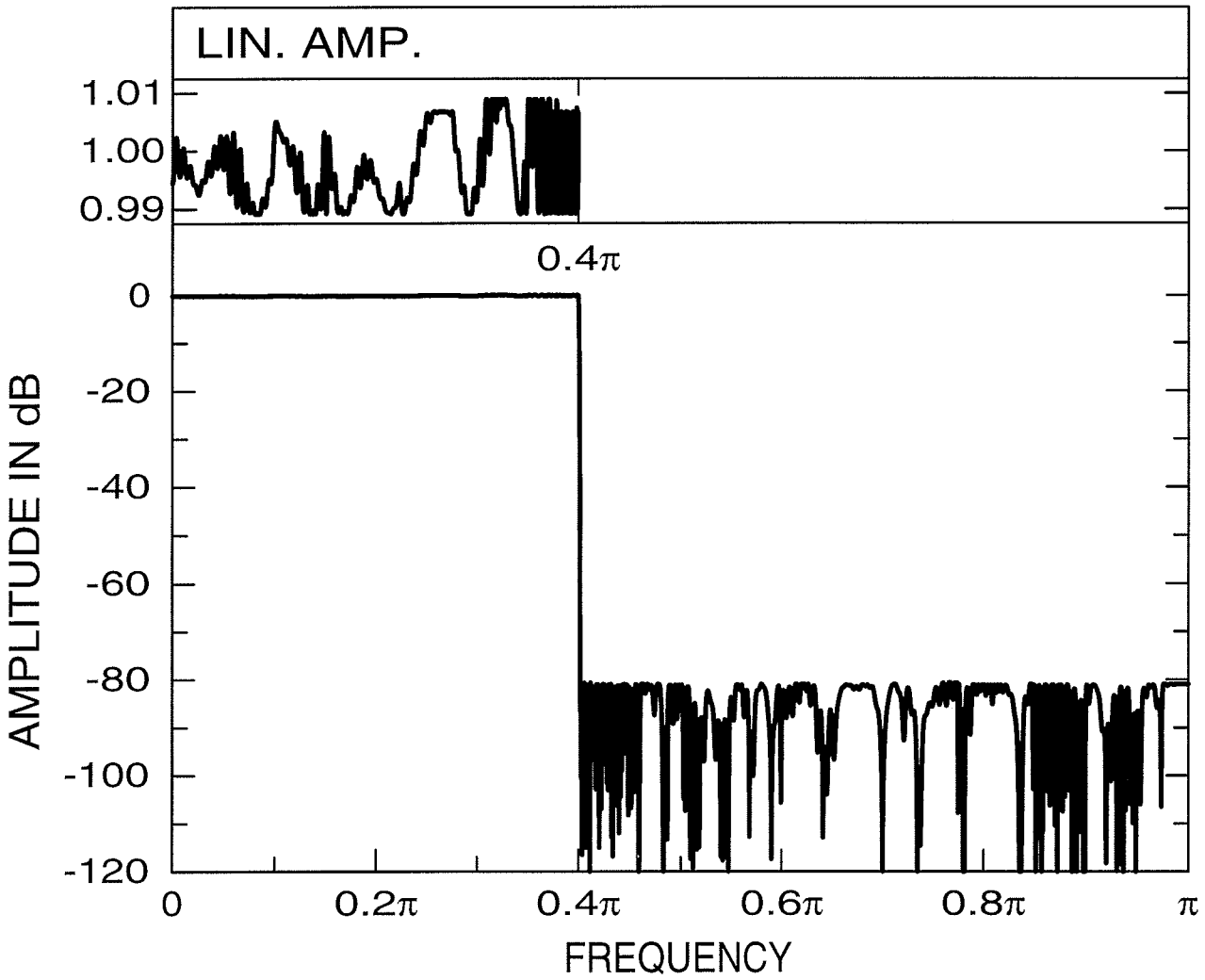
- Use a tapped cascaded interconnection of identical subfilter with large ripple values.

- With a few simple tap coefficients, the large ripples of the subfilter are reduced to small ripples of the overall design.

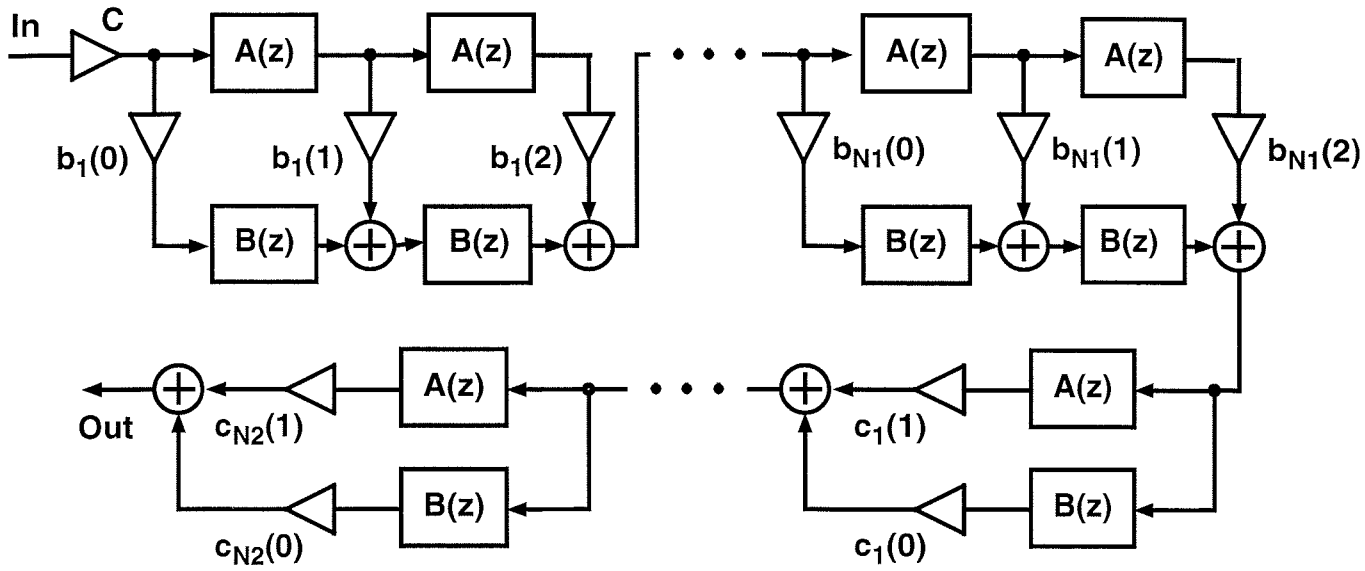
FIR filter: $\omega_p = 0.4\pi$, $\omega_s = 0.402\pi$, $\delta_p = 0.01$, $\delta_s = 10^{-4}$ ($A_p = 0.017$ dB, $A_s = 80$ dB)



FIR filter: $\omega_p = 0.4\pi$, $\omega_s = 0.402\pi$, $\delta_p = 0.01$, $\delta_s = 10^{-4}$ ($A_p = 0.017$ dB, $A_s = 80$ dB)

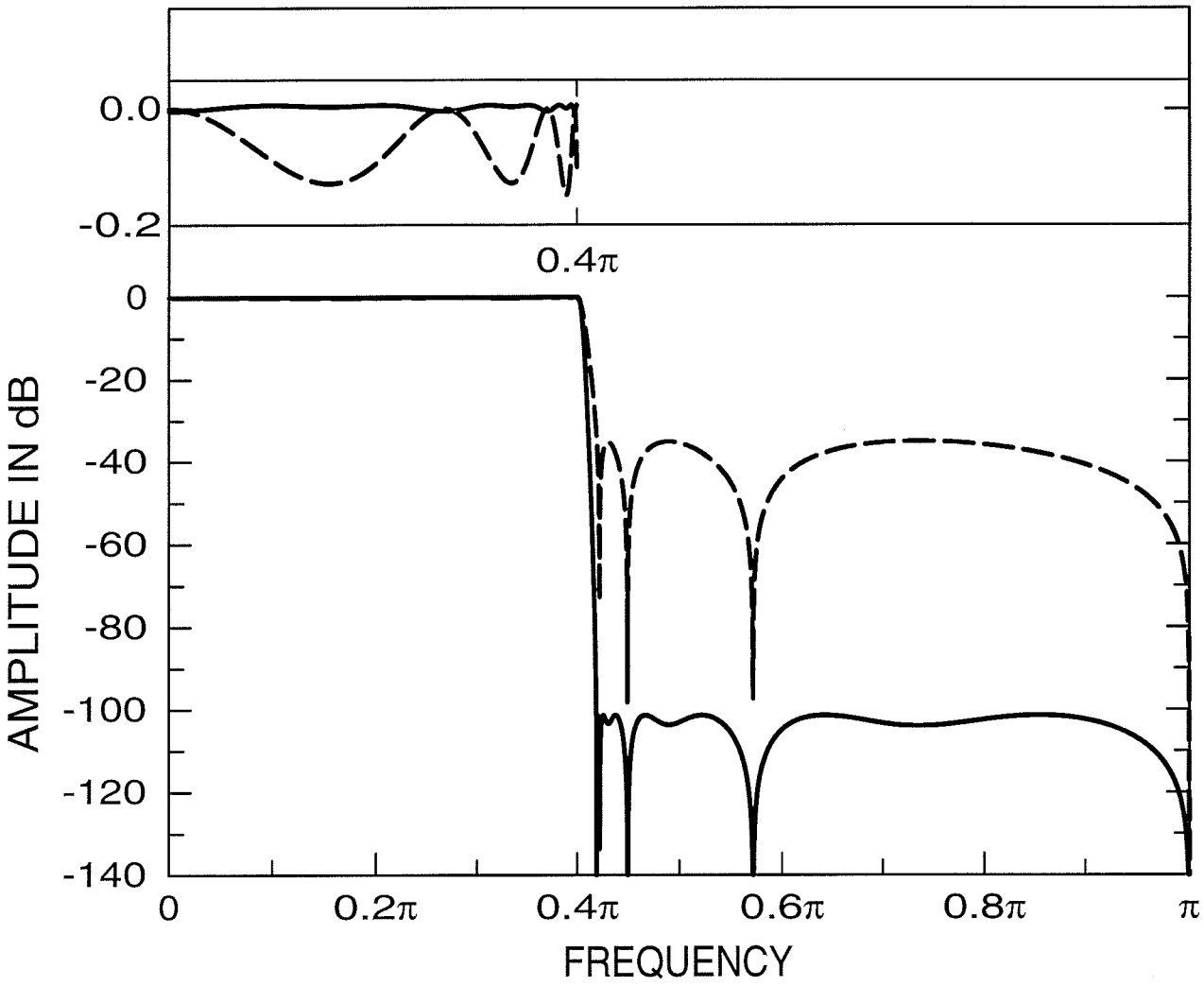


IIR filter: $\omega_p = 0.4\pi,\ \omega_s = 0.42\pi,\ A_p = 0.017$ dB, $A_s = 100$ dB

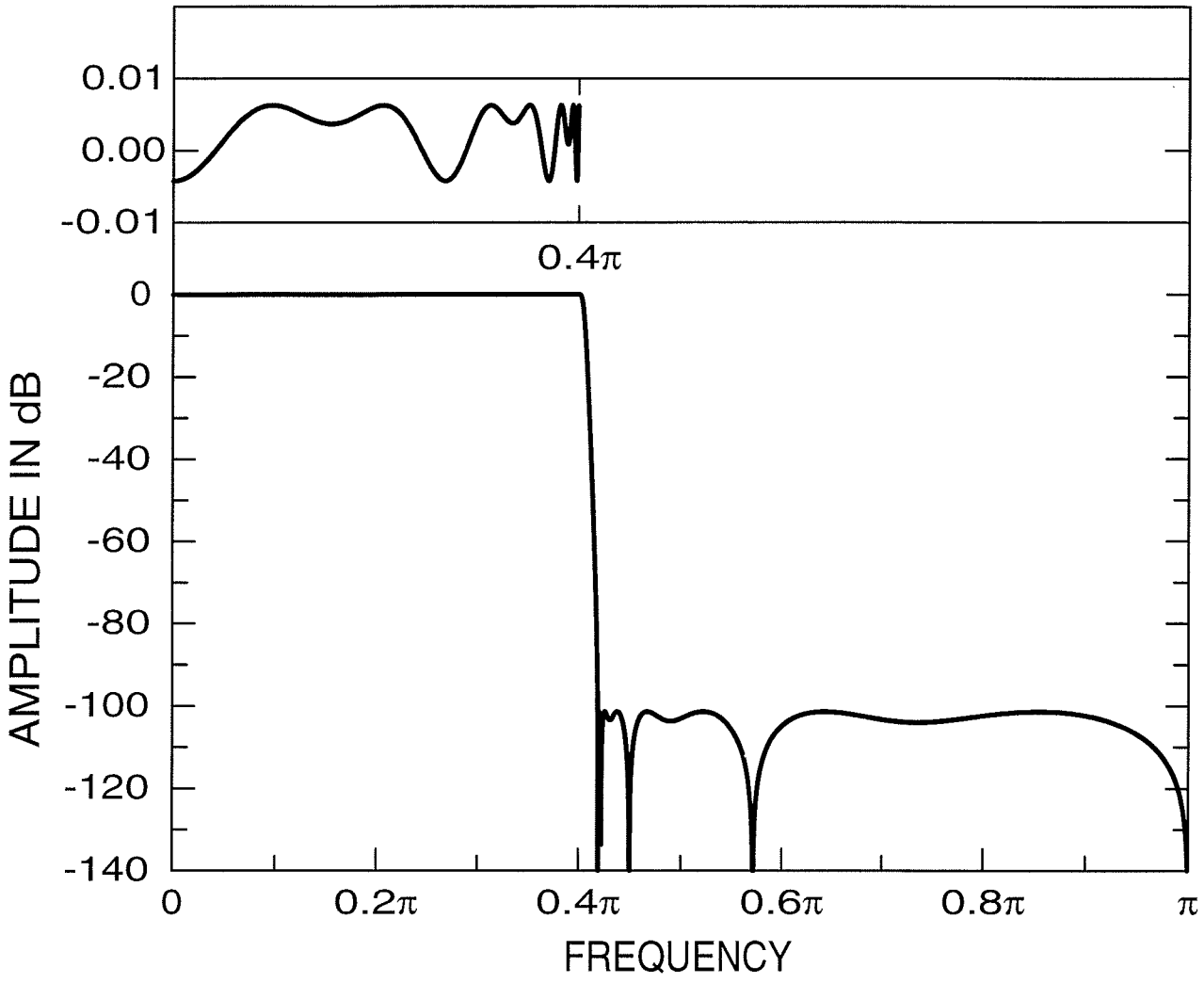


- $C = 1$
- $N_2 = 2,\ c_1(0) = c_1(1) = 2^{-1},\ c_2(0) = 1 + 2^{-1} + 2^{-5},$ $c_2(1) = -2^{-1} - 2^{-5}$
- $N_1 = 1,\ b_1(0) = b_1(2) = 2^{-2},\ b_1(1) = 2^{-1} - 2^{-11}$

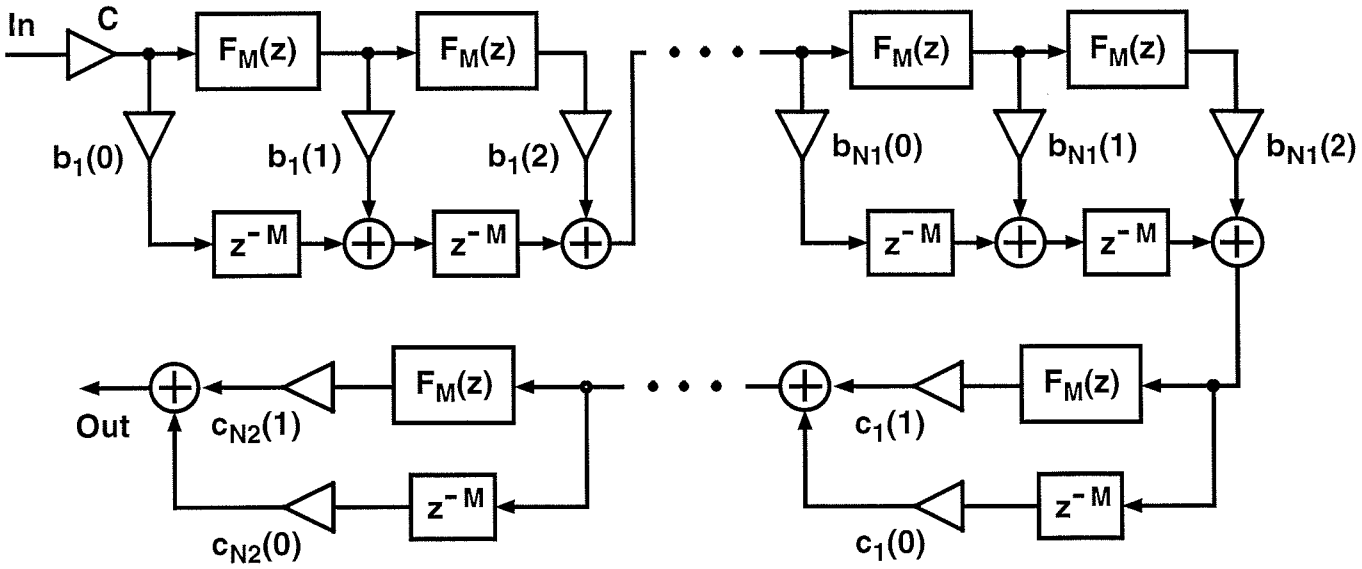
IIR filter: $\omega_p = 0.4\pi$, $\omega_s = 0.42\pi$, $A_p = 0.017$ dB, $A_s = 100$ dB



IIR filter: $\omega_p = 0.4\pi,\ \omega_s = 0.42\pi,\ A_p = 0.017$ dB, $A_s = 100$ dB

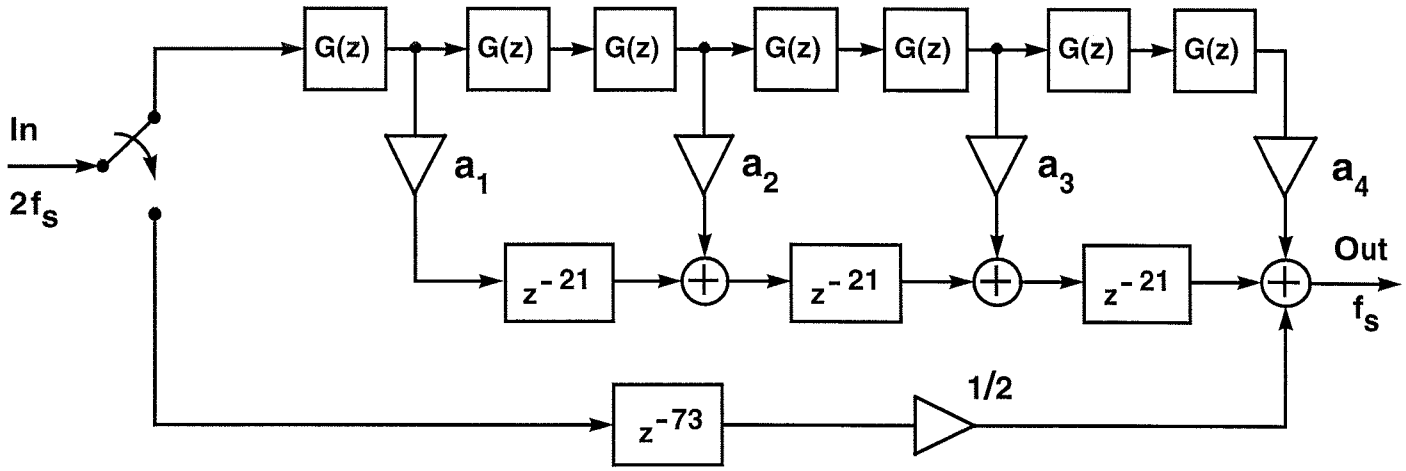


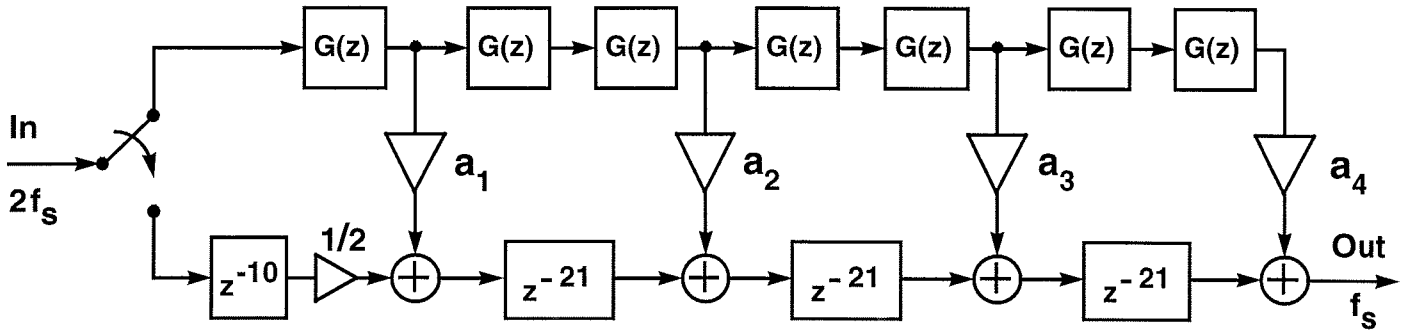
FIR filter: $\omega_p = 0.4\pi$, $\omega_s = 0.402\pi$, $\delta_p = 0.01$, $\delta_s = 10^{-4}$ ($A_p = 0.017$ dB, $A_s = 80$ dB)



$b_1(2) = 2^0 - 2^{-3} + 2^{-6}$	$b_1(1) = -2^1 + 2^{-4} + 2^{-8}$
$b_1(0) = 2^0 + 2^{-3} - 2^{-8}$	
$c_1(1) = 2^{-1} + 2^{-4} + 2^{-7}$	$c_1(0) = -2^{-1} - 2^{-2}$
$c_2(1) = 2^{-1} + 2^{-6}$	$c_2(0) = 2^{-4} - 2^{-8}$
$c_3(1) = 2^{-1} + 2^{-3}$	$c_3(0) = 2^{-5} + 2^{-6}$
$c_4(1) = 2^0$	$c_4(0) = 2^{-7}$
$c_5(1) = 2^0 - 2^{-5}$	$c_5(0) = -2^{-4}$
$c_6(1) = 2^0 - 2^{-5}$	$c_6(0) = -2^{-3} + 2^{-6}$
$C = -2^8 + 2^4 + 2^3 - 2^0$	

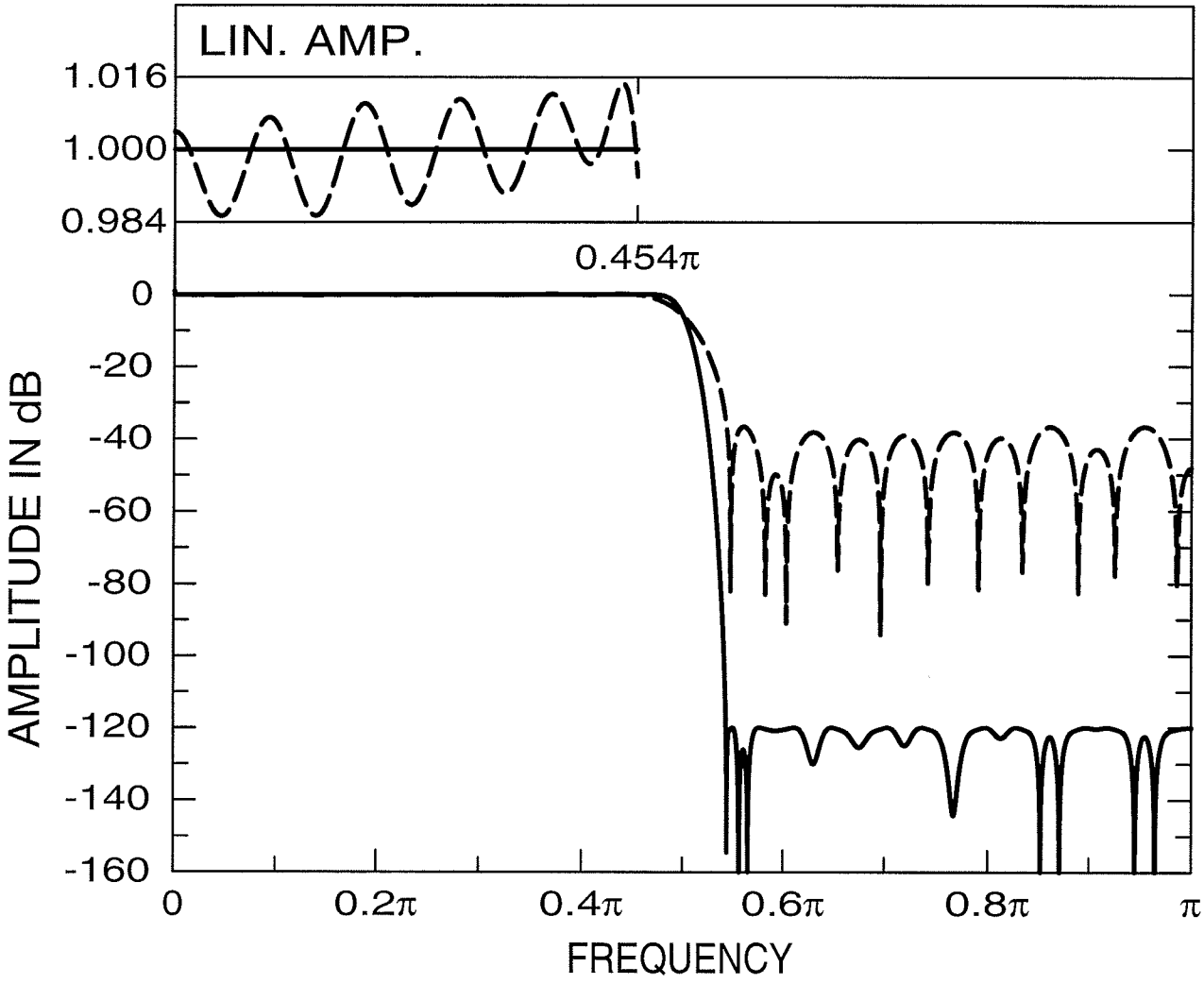
Half-band FIR filter: $\omega_p = 0.454\pi$, $\omega_s = 0.546\pi$, $\delta_p = \delta_s = 10^{-6}$ $(A_s = 120$ dB$)$



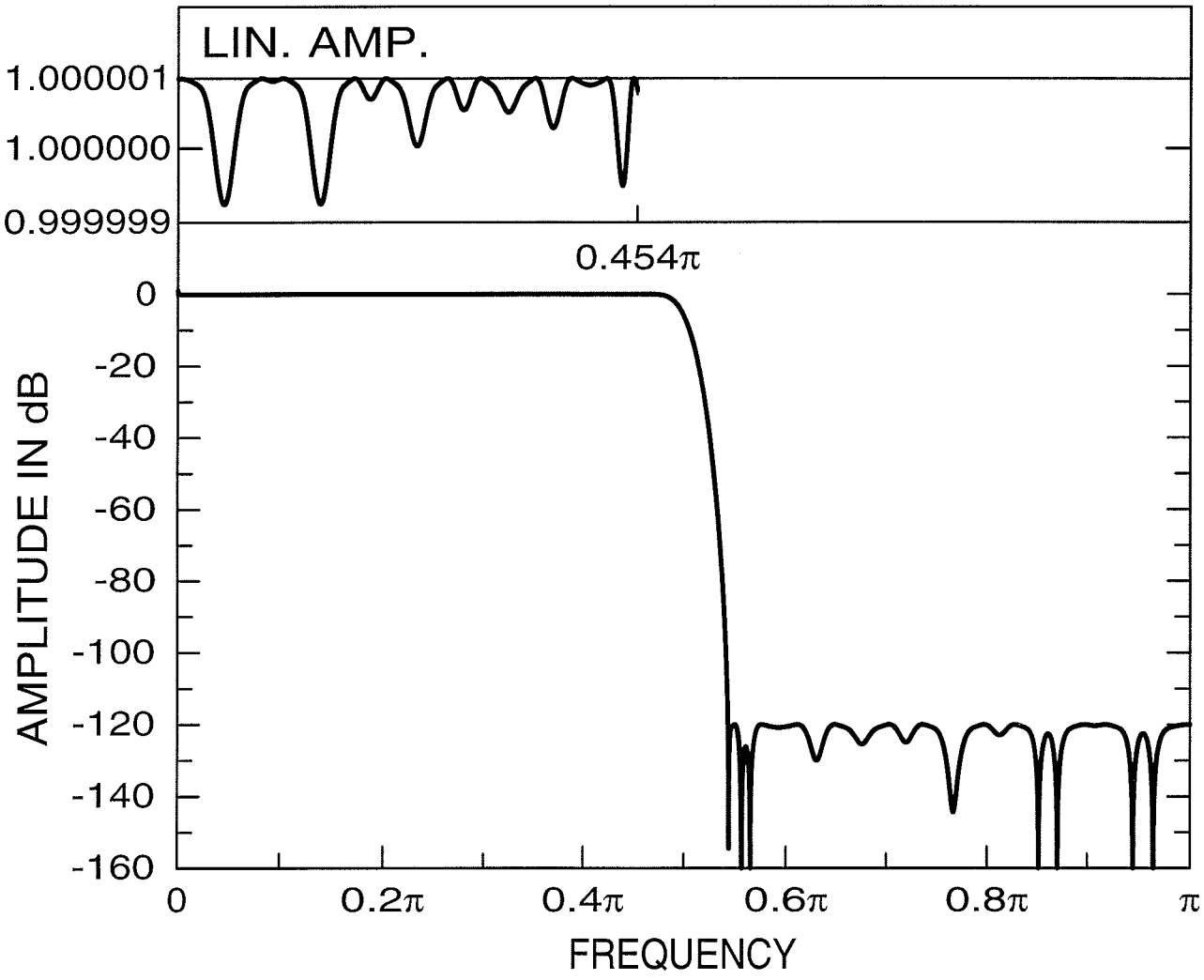


- $a_1 = 2^0 + 2^{-4} + 2^{-5}$, $a_2 = -2^0 - 2^{-4} - 2^{-5}$
- $a_3 = 2^{-1} + 2^{-3} + 2^{-5}$, $a_4 = 2^{-3} + 2^{-5} + 2^{-20}$

Half-band FIR filter: $\omega_p = 0.454\pi$, $\omega_s = 0.546\pi$, $\delta_p = \delta_s = 10^{-6}$ $(A_s = 120$ dB$)$



Half-band FIR filter: $\omega_p = 0.454\pi$, $\omega_s = 0.546\pi$, $\delta_p = \delta_s = 10^{-6}$ $(A_s = 120$ dB$)$



Practical Synthesis

I. Determine tap coefficients such that

1. They have simple representation forms.
2. The allowable ripple variations for the subfilter become huge.

II. Determine subfilter coefficients in two steps:

1. Quantize the coefficients to a few bits.
2. Select the nearest simple representation form for each coefficient value.

Automatika, Journal for Control, Measurements, Electronics, Computing and Communications, vol. 34, pp. 111–116, 1993.1

Optimization of Digital Filter Structures for VLSI Implementation

Tapio Saramäki and Tapani Ritoniemi

VLSI Solution Oy
Kanslerinkatu 6
SF–33720 Tampere, Finland

Summary. *Systematic techniques are reviewed for designing multiplier-free digital filters. The filters under consideration are finite impulse response (FIR) and infinite impulse response (IIR) digital filters as well as decimators and interpolators. In all these cases, the desired result is obtained by building the composite filter using low-order subfilters. Several examples are presented illustrating the usefulness of this approach in significantly reducing the silicon area required in the VLSI implementations of these filters.*

I. INTRODUCTION

In many signal processing applications, it is desired to implement a digital filter without general multipliers. Filters of this kind are very attractive in VLSI implementation where a general multiplier is very costly. If a general multiplier is used, then it is desired to implement all the coefficient values of a filter using a single time-multiplexed multiplier accumulator structure. This limits the attainable sampling rate of the filter and takes a significant silicon area. In addition, the multiplexing of required fast circuit structures causes a high power consumption. For very selective filters, both the number of filter coefficients and the required accuracy for the coefficient representations increase. In this case, both the required accuracy and speed of the multiplier become very demanding.

In order to avoid the use of a costly multiplier, it is preferred to design the filter in such a way that all the coefficient values can be expressed as simple combinations of powers-of-two. A very useful approach to designing such filters is to construct them using low-order subfilters. This paper reviews techniques proposed in [1]–[5] for synthesizing multiplier-free filters. Main emphasis is on examples which illustrate the usefulness of these design methods.

II. FIR FILTER DESIGNS

Low-order linear-phase FIR filters whose stopband attenuation is not very high can be designed without general multipliers in a straightforwad manner. If the filter order is rather low (less than 70) and the stopband attenuation is not very high (less than 50 dB), then a straightforward technique for designing FIR filters with two powers-of-two coefficient values, i.e., values of the form $\pm 2^{-P_1} \pm 2^{-P_2}$, is to use mixed integer linear programming [6], [7]. If the required filter order is higher, then this technique can still be used by implementing the filter using low-order filter stages, e.g., using the Jing-Fam approach [8], the frequency-response masking approach of Lim [9], or the IFIR approach [10].

However, when the required stopband attenuation is higher than 50 dB, there do not usually exist solutions with two powers-of-two coefficient values. This problem can be overcome by implementing the filter using several identical copies of the same subfilter which are interconnected with the aid of a few additional adders and tap coefficients [1], [2]. One alternative to implement the overall filter is depicted in Fig. 1. Here, $F_M(z)$ is a subfilter whose order is $2M$. The delay of this filter is thus M. In this structure, the additional delay terms z^{-M} can be shared with $F_M(z)$ after proper arrangements.

The design of the composite filter consists of two simple steps. In the first step, the additional tap coefficients used for interconnecting the subfilters ($c_k(0)$'s, $c_k(1)$'s, $b_k(0)$'s, $b_k(1)$'s, and $b_k(2)$'s) are determined and quantized in a systematic way to values which are simple combinations of powers-of-two. The second step is then to quantize the subfilter coefficients to similar representation forms. This is rather trivial to

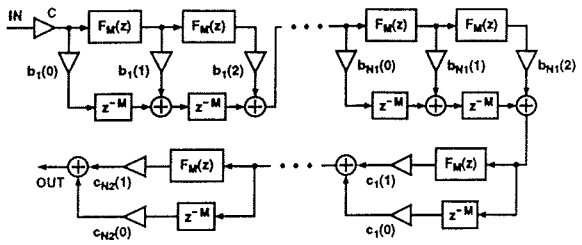


Fig. 1. *A structure for implementing a linear-phase FIR filter as a tapped cascaded interconnection of identical subfilters $F_M(z)$ of order $2M$.*

perform since the ripple values for the subfilter are huge compared to those of the overall design. Also, the subfilter order is much lower than that of the overall design. Even direct rounding can be used. This is based on the following two rules of thumb for direct rounding of the coefficients of FIR filters. The first rule is that if the allowable quantization error is doubled, then one bit is saved. The second rule of thumb for direct rounding is that if there are two filters with the same allowable quantization error and the order of the first filter is one fourth of that of the second filter, then the first filter requires one bit less.

Example 1: We consider the design of a very selective linear-phase FIR filter with specifications: $\omega_p = 0.4\pi$, $\omega_s = 0.402\pi$, $\delta_p = 0.01$, and $\delta_s = 0.0001$ (80-dB attenuation). The estimated minimum order of a conventional direct-form FIR filter to meet these criteria is approximately 3200, requiring 1600 general coefficients values. The required coefficient wordlength is more than 20 bits. The given criteria can be met using the structure of Fig. 1, where eight identical subfilters are used (one second-order block ($N_1 = 1$) and six first-order blocks ($N_2 = 6$)). The easily implementable coefficient values for this structure are shown in Table I.

Figure 2 gives an implementation for a computationally efficient linear-phase subfilter with transfer function of the form [9]

$$F_M(z) = E(z^L)G_1(z) + [z^{-LN_E/2} - E(z^L)]G_2(z). \quad (1)$$

In this structure, the transfer function $E(z^L)$ can be implemented by replacing each unit delay in a conventional transfer function $E(z)$ by L delays. The delay term $z^{-LN_E/2}$ can be shared with $E(z^L)$. Also, $G_1(z)$ and $G_2(z)$ can share the common delays by implementing them using the transposed direct-form structure exploiting the coefficient symmetry. For this subfilter, the passband and stopband edges are the same as those of the overall filter, whereas the required passband and stopband ripples are 0.1787 and 0.1195, respectively. These ripple values are huge compared to those of the overall design, making the quantization of the subfilter coefficients rather trivial. The given criteria are met by $L = 16$ and $E(z)$, $G_1(z)$, and $G_2(z)$ of orders $N_E = 40$, 22, and 30, respectively. It should be noted that in order to make the delays of $G_1(z)$ and $G_2(z)$ the same, instead of $G_1(z)$, $z^{-4}G_1(z)$ has to be used. The transfer functions $E(z)$, $G_1(z)$, and $G_2(z)$ can be synthesized using the technique proposed in [9]. Then direct rounding can be used to quantize the filter coefficients to the 6-bit values shown in Table II. Because of the coefficient symmetry, only half of the coefficients are given for each filter. Only one coefficient ($g_2(14) = 19 \cdot 2^{-6}$) requires a three powers-of-two representation. The resulting overall design requires no multipliers, making it very useful for VLSI implementation.

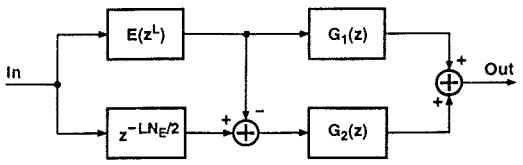


Fig. 2. *A computationally efficient implementation for the subfilter $F_M(z)$.*

Table I
QUANTIZED TAP COEFFICIENTS FOR THE STRUCTURE OF FIG. 1

$b_1(2) = 2^0 - 2^{-3} + 2^{-6}$	$b_1(1) = -2^1 + 2^{-4} + 2^{-8}$
$b_1(0) = 2^0 + 2^{-3} - 2^{-8}$	
$c_1(1) = 2^{-1} + 2^{-4} + 2^{-7}$	$c_1(0) = -2^{-1} - 2^{-2}$
$c_2(1) = 2^{-1} + 2^{-6}$	$c_2(0) = 2^{-4} - 2^{-8}$
$c_3(1) = 2^{-1} + 2^{-3}$	$c_3(0) = 2^{-5} + 2^{-6}$
$c_4(1) = 2^0$	$c_4(0) = 2^{-7}$
$c_5(1) = 2^0 - 2^{-5}$	$c_5(0) = -2^{-4}$
$c_6(1) = 2^0 - 2^{-5}$	$c_6(0) = -2^{-3} + 2^{-6}$
$C = -2^8 + 2^4 + 2^3 - 2^0$	

The upper and lower curves in Fig. 3 give the amplitude responses for the subfilter and the overall filter, respectively. It should be pointed out that in this example very tight specifications have been selected in order to show the effectiveness of our approach. This approach is much easier to apply to milder overall criteria.

III. IIR FILTER DESIGNS

Similarly, for IIR designs, multiplier-free filters can be obtained in a straightforward manner for rather mild criteria provided that low-sensitivity structures are used. For more stringent specifications, there are several alternatives to arrive at the desired goal. This is generally achieved by constructing the overall filter using several low-order multipler-free blocks.

A systematic technique to design IIR filters

Table II
QUANTIZED COEFFICIENTS FOR A SUBFILTER DESIGNED USING THE FREQUENCY-RESPONSE MASKING APPROACH [9]

$e(0) = 2 \cdot 2^{-6}$	$e(1) = -4 \cdot 2^{-6}$
$e(2) = -3 \cdot 2^{-6}$	$e(3) = -2 \cdot 2^{-6}$
$e(4) = 1 \cdot 2^{-6}$	$e(5) = 0$
$e(6) = -1 \cdot 2^{-6}$	$e(7) = -2 \cdot 2^{-6}$
$e(8) = 0$	$e(9) = 2 \cdot 2^{-6}$
$e(10) = 1 \cdot 2^{-6}$	$e(11) = -2 \cdot 2^{-6}$
$e(12) = -3 \cdot 2^{-6}$	$e(13) = 1 \cdot 2^{-6}$
$e(14) = 4 \cdot 2^{-6}$	$e(15) = 1 \cdot 2^{-6}$
$e(16) = -5 \cdot 2^{-6}$	$e(17) = -6 \cdot 2^{-6}$
$e(18) = 5 \cdot 2^{-6}$	$e(19) = 17 \cdot 2^{-6}$
$e(20) = 28 \cdot 2^{-6}$	

$g_1(0) = 2 \cdot 2^{-6}$	$g_1(1) = 3 \cdot 2^{-6}$
$g_1(2) = -1 \cdot 2^{-6}$	$g_1(3) = -2 \cdot 2^{-6}$
$g_1(4) = -1 \cdot 2^{-6}$	$g_1(5) = 3 \cdot 2^{-6}$
$g_1(6) = 2 \cdot 2^{-6}$	$g_1(7) = -4 \cdot 2^{-6}$
$g_1(8) = -5 \cdot 2^{-6}$	$g_1(9) = 4 \cdot 2^{-6}$
$g_1(10) = 20 \cdot 2^{-6}$	$g_1(11) = 28 \cdot 2^{-6}$

$g_2(0) = -3 \cdot 2^{-6}$	$g_2(1) = -1 \cdot 2^{-6}$
$g_2(2) = 1 \cdot 2^{-6}$	$g_2(3) = 2 \cdot 2^{-6}$
$g_2(4) = 1 \cdot 2^{-6}$	$g_2(5) = -1 \cdot 2^{-6}$
$g_2(6) = -2 \cdot 2^{-6}$	$g_2(7) = 0$
$g_2(8) = 3 \cdot 2^{-6}$	$g_2(9) = 3 \cdot 2^{-6}$
$g_2(10) = -1 \cdot 2^{-6}$	$g_2(11) = -5 \cdot 2^{-6}$
$g_2(12) = -2 \cdot 2^{-6}$	$g_2(13) = 7 \cdot 2^{-6}$
$g_2(14) = 19 \cdot 2^{-6}$	$g_2(15) = 24 \cdot 2^{-6}$

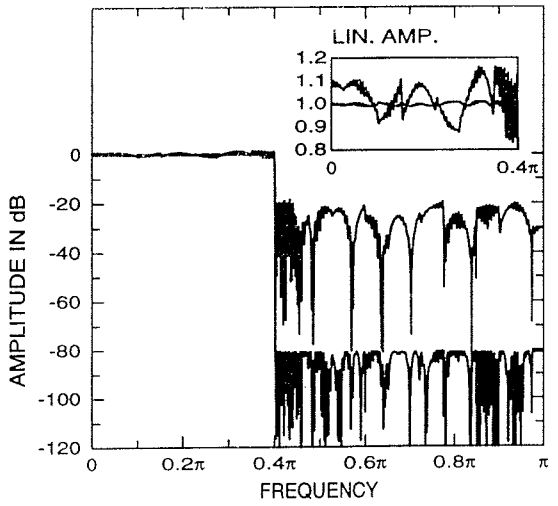


Fig. 3. *Amplitude responses for the overall filter (lower curve) and the subfilter (upper curve) synthesized using the frequency-response masking approach [9].*

without general multipliers is to use identical building blocks, like for linear-phase FIR filters. One alternative is to build the overall filter as

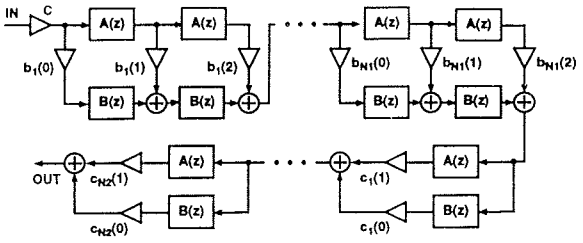


Fig. 4. *A structure for implementing an IIR filter as a tapped cascaded interconnection of identical copies of two allpass sections $A(z)$ and $B(z)$.*

shown in Fig. 4 [3], where the basic building blocks $A(z)$ and $B(z)$ are allpass filters. The design of these filters consists also of two steps. In the first step, the additional tap coefficients in the structure of Fig. 4 are determined and quantized in a systematic manner. The second step is then to determine the allpass sections $A(z)$ and $B(z)$ in such a way that

$$H(z) = \frac{1}{2}[A(z) + B(z)] \qquad (2)$$

satisfies the specifications determined by the overall filter structure. The basic idea in using the structure of Fig. 4 is that the allowable ripple values for $H(z)$ are huge compared to those of the overall filter. This makes the quantization of the coefficients of $A(z)$ and $B(z)$ trivial. Again, an example is used to illustrate the usefulness of this approach.

Example 2: The lowpass filter specifications are: $\omega_p = 0.4\pi$, $\omega_s = 0.42\pi$, $\delta_p = 0.001$ (0.017 dB overall passband variation), and $\delta_s = 0.00001$ (100 dB attenuation). The estimated order of an elliptic filter to meet these criteria is 15.3 so that the minimum order for which the overall filter is implemetable directly in the form of Eq. (2) is 17. In this case, the orders of the allpass filters $A(z)$ and $B(z)$ are 8 and 9, respectively.

If an elliptic subfilter of order 7 is desired to be used, then four copies of $A(z)$ and $B(z)$ are required. The orders of $A(z)$ and $B(z)$ are 4 and 3, respectively. The overall filter can be implemented as shown in Fig. 4, where $C = 1$ and there are two first-order blocks ($N_2 = 2$) with tap coefficients

$$c_1(0) = c_1(1) = 2^{-1} \qquad (3)$$

and

$$c_2(0) = 1+2^{-1}+2^{-5}, \quad c_2(1) = -2^{-1}-2^{-5}, \quad (4)$$

and one second-order block ($N_1 = 1$) with coefficients

$$b_1(0) = b_1(2) = 2^{-2}, \quad b_1(1) = 2^{-1} - 2^{-11}. \quad (5)$$

In this case, it is required that the overall passband variation and the minimum stopband attenuation for $H(z)$ as given by Eq. (2) are 0.19 dB and 31.7 dB, respectively. These variations

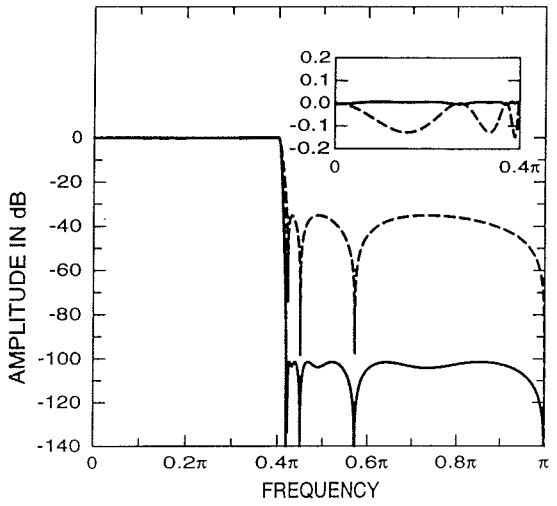


Fig. 5. *Amplitude responses for the overall IIR filter (solid line) and for the subfilter (dashed line).*

are huge compared to those of the overall design, enabling us to find simple coefficient values for $A(z)$ and $B(z)$ through direct rounding. If both $A(z)$ and $B(z)$ are implemented using first- and second-order wave digital allpass structures [11], then $A(z)$ contains two second-order sections with coefficients

$$\gamma_1 = -2^{-1} + 2^{-5} + 2^{-6}, \quad \gamma_2 = 2^{-1} - 2^{-6} - 2^{-7} \tag{6}$$

and

$$\gamma_1 = -2^{0} + 2^{-5} + 2^{-7}, \quad \gamma_2 = 2^{-2} + 2^{-5} + 2^{-6}, \tag{7}$$

respectively. $B(z)$ consists of one first-order section with the adaptor coefficient

$$\gamma = 2^{-2} + 2^{-3} + 2^{-7} \tag{8}$$

and one second-order section with adaptor coefficients

$$\gamma_1 = -2^{-1} - 2^{-2} - 2^{-4}, \tag{9a}$$

$$\gamma_2 = 2^{-2} + 2^{-4} + 2^{-6} + 2^{-7}. \tag{9b}$$

The amplitude responses for $[A(z) + B(z)]/2$ and for the overall design are given by the dashed and solid lines of Fig. 5, respectively. Even though the 17th-order pure elliptic design meets very well the given overall criteria, it requires 18 bits to stay within the given amplitude tolerances. Another advantage of the new filter is that it requires only 7 distinct multipliers, whereas the elliptic design has 17 multipliers. Furthermore, the radius of the outermost pole for the pure elliptic design is 0.9948, whereas that of the subfilter is 0.9802. This means that the multiplication roundoff noise generated by the pure elliptic filter is significantly higher and it requires a significantly longer internal data wordlength.

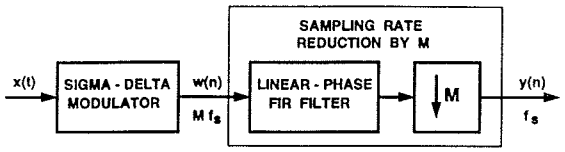


Fig. 6. *Block diagram for the A/D converter consisting of an oversampled sigma-delta modulator and a decimator filter.*

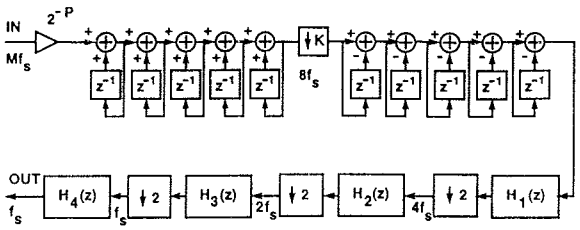


Fig. 7. *Implementation of the proposed decimator.*

IV. DECIMATOR AND INTERPOLATOR DESIGNS

We consider the design of FIR decimators for filtering the output signal of a sigma-delta modulator. The block diagram for the overall system is depicted in Fig. 6. The output sampling rate of the sigma-delta modulator is M times the final sampling rate f_s. We state the amplitude requirements for the decimator in the form

$$1 + \delta_p \le \left|H(e^{j\frac{2\pi f}{Mf_s}})\right| \le 1 - \delta_p \quad \text{for} \quad 0 \le f \le \alpha\frac{f_s}{2} \tag{10a}$$

$$\left|H(e^{j\frac{2\pi f}{Mf_s}})\right| \le \delta_s \quad \text{for} \quad (2-\alpha)\frac{f_s}{2} \le f \le M\frac{f_s}{2}. \tag{10b}$$

When these specifications are satisfied, then the signal components aliasing into the passband $[0, \alpha f_s/2]$ are attenuated at least by $1/\delta_s$. We consider the following criteria:

$$\delta_p = 0.0001, \quad \delta_s = 0.000001, \quad \alpha = 0.907.$$

In this case, the stopband attenuation is at least 120 dB. α has been selected such that the passband edge becomes 20 kHz for the final output sampling rate of $f_s = 44.1$ kHz.

To reduce the arithmetic complexity of the decimator, it is preferred to construct it using several low-order stages, instead of one high-order stage. An efficient multistage implementation for the decimator is given in Fig. 7 for $M = 8K$ [4]. By selecting $K = 2, 4, 8, 16, 32, \ldots$, the same structure can be used for the overall decimation ratios of $M = 16, 32, 64, 128, 256, \ldots$, respectively. The first decimation stage providing the decimation by an adjustable factor of K is constructed using five accumulators and five differentiators. If $P = 5\log_2 K$ for the scaling multiplier 2^{-P} and two's complement arithmetic is

used, then the output of this filter stage is correct even though there occur overflows in the feedback loops. This filter stage requires only 10 adders and 10 delay elements regardless of the value of K.

In order to avoid the use of general multipliers also in the remaining stages, the subfilters $H_1(z)$, $H_2(z)$, $H_3(z)$, and $H_4(z)$ have been designed to be special tailored filters. Their transfer functions are:

$$H_1(z) = 2^{-1}z^{-9} + \widehat{H}_1(z^2), \qquad (11a)$$

where

$$\widehat{H}_1(z) = F_1(z)\left[(2^{-1}+2^{-2})z^{-3} + (-2^{-2}+2^{-20})[F_1(z)]^2\right] \qquad (11b)$$

with

$$F_1(z) = (-2^{-4}-2^{-7}-2^{-11})(1+z^{-3}) + (2^{-1}+2^{-4}+2^{-7})(z^{-1}+z^{-2}). \qquad (11c)$$

$$H_2(z) = 2^{-1}z^{-21} + \widehat{H}_2(z^2), \qquad (12a)$$

where

$$\widehat{H}_2(z) = F_2(z)\left[(2^0+2^{-4}+2^{-5})z^{-9} + (-2^0-2^{-4}-2^{-5})z^{-6}[F_2(z)]^2 + (2^{-1}+2^{-3}+2^{-5})z^{-3}[F_2(z)]^4 + (-2^{-3}-2^{-5}+2^{-20})[F_2(z)]^6\right] \qquad (12b)$$

with

$$F_2(z) = (-2^{-4}-2^{-5})(1+z^{-3}) + (2^{-1}+2^{-4}+2^{-5})(z^{-1}+z^{-2}). \qquad (12c)$$

$$H_3(z) = 2^{-1}z^{-147} + \widehat{H}_3(z^2), \qquad (13a)$$

where

$$\widehat{H}_3(z) = F_3(z)\left[(2^0+2^{-4}+2^{-5})z^{-63} + (-2^0-2^{-4}-2^{-5})z^{-42}[F_3(z)]^2 + (2^{-1}+2^{-3}+2^{-5})z^{-21}[F_3(z)]^4 + (-2^{-3}-2^{-5}+2^{-20})[F_3(z)]^6\right] \qquad (13b)$$

with

$$\begin{aligned} F_3(z) &= 2^{-6}(1+z^{-21}) \\ &+ (-2^{-7}-2^{-8})(z^{-1}+z^{-20}) \\ &+ 2^{-6}(z^{-2}+z^{-19}) \\ &+ (-2^{-6}-2^{-7})(z^{-3}+z^{-18}) \\ &+ 2^{-5}(z^{-4}+z^{-17}) \\ &+ (-2^{-5}-2^{-7}-2^{-8})(z^{-5}+z^{-16}) \\ &+ (2^{-4}-2^{-8})(z^{-6}+z^{-15}) \\ &+ (-2^{-4}-2^{-6}-2^{-8})(z^{-7}+z^{-14}) \\ &+ (2^{-3}-2^{-8})(z^{-8}+z^{-13}) \\ &+ (-2^{-2}+2^{-5}+2^{-7})(z^{-9}+z^{-12}) \\ &+ (2^{-1}+2^{-3}+z^{-7})(z^{-10}+z^{-11}). \end{aligned} \qquad (13c)$$

$$H_4(z) = z^{-7} + (2^{-6}+2^{-10})\widehat{H}_4(z), \qquad (14a)$$

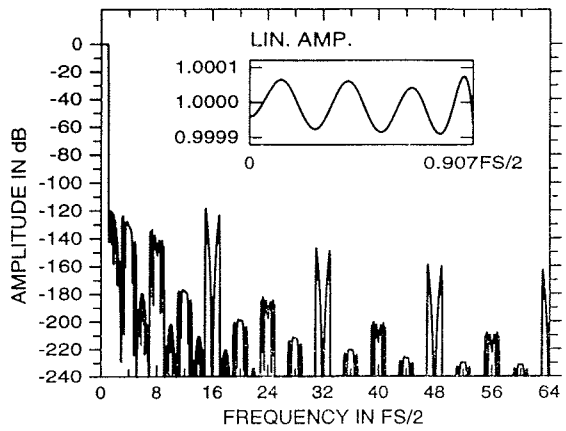


Fig. 8. *Amplitude response for the overall decimator for* $M = 64$.

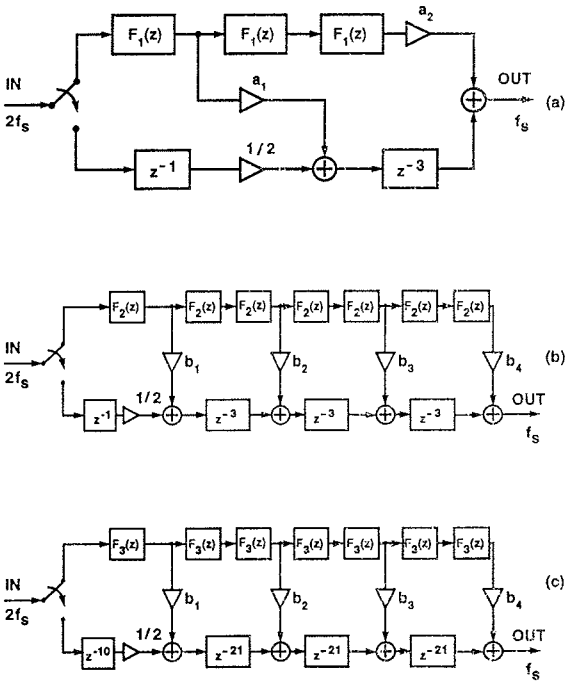


Fig. 9. *Efficient implementations for the half-band subfilters. (a)* $H_1(z)$. *(b)* $H_2(z)$. *(c)* $H_3(z)$.

where

$$\begin{aligned} \widehat{H}_4(z) &= (-2^{-8}-2^{-11})(1+z^{-14}) \\ &+ (2^{-8}+2^{-9}+2^{-11})(z^{-1}+z^{-13}) \\ &+ (-2^{-7}-2^{-8}+2^{-11})(z^{-2}+z^{-12}) \\ &+ (2^{-6}+2^{-8}+2^{-9})(z^{-3}+z^{-11}) \\ &+ (-2^{-5}-2^{-7}-2^{-8})(z^{-4}+z^{-10}) \\ &+ (2^{-3}-2^{-6}-2^{-8})(z^{-5}+z^{-9}) \\ &+ (-2^{-1}+2^{-4}+2^{-9})(z^{-6}+z^{-8}) \\ &+ (2^{-1}+2^{-2}-2^{-5})z^{-7}. \end{aligned} \qquad (14b)$$

The amplitude response of the overall design is depicted in Fig. 8 for $M = 64$ ($K = 8$). The subfilters $H_1(z)$, $H_2(z)$, and $H_3(z)$ are special half-band filters which can be implemented effec-

tively using a polyphase structure based on the commutative model [12]. The structures resulting by properly sharing the delays between the two branches are shown in Fig. 9. Fig. 9(a) gives the structure for $H_1(z)$, Fig. 9(b) for $H_2(z)$, and Fig. 9(c) for $H_3(z)$. The tap coefficients are $a_1 = 2^{-1}+2^{-2}$, $a_2 = -2^{-2}+2^{-20}$, $b_1 = 2^0+2^{-4}+2^{-5}$, $b_2 = -2^0-2^{-4}-2^{-5}$, $b_3 = 2^{-1}+2^{-3}+2^{-5}$, and $b_4 = 2^{-3}+2^{-5}+2^{-20}$. Note that the tap coefficients are the same for $H_2(z)$ and $H_3(z)$. One of the branches for all the three filters is a pure delay term. For $H_1(z)$, the other branch is a tapped cascaded interconnection of three identical subfilters of order 3, for $H_2(z)$, the other branch consists of seven identical subfilters of order 3, and, for $H_3(z)$, seven identical subfilters of order 21. The role of $H_4(z)$ working at the output sampling rate of f_s is to equalize the passband distortion caused by the earlier filter stages in the overall system.

The details of how the filter stages share the frequency-response-shaping responsibilities can be found in [4]. In [5], it is shown how efficient interpolators can be designed based on the use of similar ideas.

ACKNOWLEDGEMENT

The authors wish to thank the Median-Free Group International for excellent working atmosphere and fruitful discussions during the course of this work.

REFERENCES

[1] T. Saramäki, "Design of FIR filters as a cascaded tapped interconnection of identical subfilters," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 1011–1029, Sept. 1987.

[2] T. Saramäki, "A systematic technique for designing highly selective multiplier-free FIR filters," in *Proc. 1991 IEEE Int. Symp. Circuits Syst.* (Singapore), pp. 484–487, June 1991.

[3] T. Saramäki and M. Renfors, "A novel approach for the design of IIR filters as a tapped cascaded interconnection of identical allpass subfilters," in *Proc. 1987 IEEE Int. Symp. Circuits Syst.* (Philadelphia, PA), pp. 629–632, May 1987.

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

[5] T. Saramäki, T. Karema, T. Ritoniemi, and H. Tenhunen, "VLSI-realizable multiplier-free interpolators for high-order sigma-delta D/A converters," in *Proc. Sixth Mediterranean Electrotechnical Conference* (Ljubljana, Yugoslavia), pp. 295–298, May 1991.

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

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

[8] Z. Jing and A. T. Fam, "A new structure for narrow transition band, lowpass digital filter design," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 362-370, Apr. 1984.

[9] Y. C. Lim, "Frequency-response masking approach for the synthesis of sharp linear phase digital filters,", *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 357–364, Apr. 1986.

[10] T. Saramäki, Y. Neuvo, and S. K. Mitra, "Design of computationally efficient interpolated FIR filters," *IEEE Trans. Circuits Syst.*, vol. CAS-35, pp. 70–88, Jan. 1988.

[11] L. Gazsi, "Explicit formulas for lattice wave digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 68-88, Jan. 1985.

[12] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1983.