PAPER

A Discrete Optimization Method for High-Order FIR Filters with Finite Wordlength Coefficients

Kenji NAKAYAMA†, *Member*

SUMMARY This paper proposes a new discrete optimization method which is mainly directed toward saving computing time for high-order FIR filters. In the proposed method, a transfer function is first approximated in a cascade form of a low-order function $W(z)$ with pre-rounded coefficients and a high-order function $F(z)$ with infinite precision coefficients. Second, rounded $F(z)$ coefficients are discretely optimized so as to minimize the mean square error of the amplitude response. In other words, the roundoff error spectrum is shaped so as to be suppressed by a weighting function $W(z)$. In order to save computing time, the error is equivalently evaluated in a time domain, and the $F(z)$ coefficients are divided into small groups in the discrete optimization procedure. Design examples for 200 tap FIR filters demonstrate practical usefullness.

1. Introduction

Coefficient wordlength is one of the factors which determine circuit complexity of digital filters. Therefore, it is important to realize desired filter responses with short wordlength coefficients[1],[2].

There are two approaches to the above design objective. One approach is to statistically estimate filter response deviations due to coefficient roundoff errors. Based on this estimation, tolerance and filter length, which are used in an exact design procedure, are modified such that the filter responses for rounded coefficients just meet specified characteristics[3]-[7].

The other approach is to optimize the finite wordlength coefficients so as to improve the deviated filter responses. Since, in this optimization, the coefficients are restricted to have discrete values, this approach is currently called "Discrete Optimization". This paper is focused on the second approach.

Several useful approaches to infinite impulse response (IIR) filters and relatively low-order finite impulse response (FIR) filters have been reported. They mainly include random search[8],[9], univarate search[10],[11], branch and bound[12]-[14], Hook-Jeeves method[15], and iterative roundoff and optimization[16]. On the other hand, for high-order FIR filters, mixed-integer programming techniques[17]-[19] and a local search method[20] have been mainly applied. These approaches, however, still require a large amount of computing time.

Manuscript received February 2, 1987.

†The author is with C&C Systems Research Laboratories, NEC Corporation, Kawasaki-shi, 213 Japan.

This paper proposes a new discrete optimization method which is mainly directed toward saving computing time for high-order FIR filters[21]. In the proposed method, a transfer function is first approximated in a cascade form of a low-order function $W(z)$ with pre-rounded coefficients and a high-order functions $F(z)$ with infinite precision coefficients. Second, rounded $F(z)$ coefficients are discretely optimized so as to minimize the mean square error of the amplitude response. In order to save computing time, the error is equivalently evaluated in a time domain, and the $F(z)$ coefficients are divided into small groups in the discrete optimization procedure.

2. Discrete Optimization by Error Spectrum Shaping

2.1 Algorithm

A transfer function $H(z)$ is synthesized in a cascade form

$$H(z)=W(z)F(z), \quad z=e^{j\omega T} \tag{1}$$

where T is a sampling period. $W(z)$ is a low-order function with pre-rounded coefficients. $F(z)$ is a high-order function and has infinite precision coefficients in an exact approximation procedure. First, $H(z)$ is approximated through conventional methods, for instance, the Remez-exchange algorithm[22].

Let $F_Q(z)$ be a function with the rounded coefficients of $F(z)$, and let $\Delta F(z)$ be

$$\Delta F(z)=F(z)-F_Q(z). \tag{2}$$

$\Delta H(z)$ is used to expresses

$$\Delta H(z)=W(z)\Delta F(z). \tag{3}$$

In the proposed method, the mean square of $|\Delta H(e^{j\omega})|$ is employed as an error criterion,

$$E=\frac{1}{2\pi}\int_{-\pi}^{\pi}|\Delta H(e^{j\omega})|^2 d\omega. \tag{4 a}$$

From Eq. (3),

$$E=\frac{1}{2\pi}\int_{-\pi}^{\pi}|W(e^{j\omega})\Delta F(e^{j\omega})|^2 d\omega \tag{4 b}$$

where T is assumed to be unity. Minimizing E is equal to shaping $|\Delta F(e^{j\omega})|$ so as to be effectively suppressed by

$|W(e^{j\omega})|$. In order to suppress $|\Delta F(e^{j\omega})|$, the $W(z)$ amplitude response is required to have small values in the stopband.

2.2 Filter Response Improvement

Let $F_{Qo}(z)$ be a function with discretely optimized coefficients of $F(z)$. Furthermore, the following functions are defined,

$$H_Q^W(z)=W(z)F_Q(z) \tag{5 a}$$

$$H_{Qo}(z)=W(z)F_{Qo}(z) \tag{5 b}$$

$$\Delta F_Q(z)=F(z)-F_Q(z) \tag{5 c}$$

$$\Delta F_{Qo}(z)=F(z)-F_{Qo}(z) \tag{5 d}$$

$$\Delta H_Q^W(z)=W(z)\Delta F_Q(z) \tag{5 e}$$

$$\Delta H_{Qo}(z)=W(z)\Delta F_{Qo}(z) \tag{5 f}$$

In these equations, $F_Q(z)$ is assumed to have the coefficients, which are obtained by only rounding off the $F(z)$ coefficients. Furthermore, $H_Q(z)$ and $\Delta H_Q(z)$ are used to denote a function with the rounded $H(z)$ coefficients, and

$$\Delta H_Q(z)=H(z)-H_Q(z) \tag{5 g}$$

respectively. By employing the assumption that the $\Delta H_{Qo}(z)$ amplitude response is flattened through the discrete optimization, it can be estimated as follows:

$$E=\frac{1}{2\pi}\int_{-\pi}^{\pi}|\Delta H_{Qo}(e^{j\omega})|^2 d\omega=C^2 \tag{6 a}$$

where

$$|\Delta H_{Qo}(e^{j\omega})|=C,\quad -\pi\leq\omega\leq\pi,\quad C:\text{constant.} \tag{6 b}$$

From Eqs. (5 f) and (6 a),

$$\frac{C^2}{2\pi}\int_{-\pi}^{\pi}|W(e^{j\omega})|^{-2}d\omega=\frac{1}{2\pi}\int_{-\pi}^{\pi}|\Delta F_{Qo}(e^{j\omega})|^2 d\omega. \tag{7}$$

When the coefficients of $\Delta F_{Qo}(z)$ are uniformly distributed in the region $(-\Delta_l/2, \Delta_l/2)$,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}|\Delta F_{Qo}(e^{j\omega})|^2 d\omega=N_F\frac{\Delta_l^2}{12} \tag{8}$$

where N_F is the filter length of $F(z)$. Therefore, $|\Delta H_{Qo}(e^{j\omega})|$ can be estimated by

$$|\Delta H_{Qo}(e^{j\omega})|=\frac{\left(N_F\frac{\Delta_l^2}{12}\right)^{\frac{1}{2}}}{\|1/W(e^{j\omega})\|_2} \tag{9}$$

where $\|\cdot\|_p$ is an L_p norm.

On the other hand, the deviation $|\Delta H_Q(e^{j\omega})|$ can be estimated by

$$|\Delta H_Q(e^{j\omega})|=\left(N\frac{\Delta_0^2}{12}\right)^{\frac{1}{2}} \tag{10}$$

under the assumption that the $\Delta H_Q(z)$ coefficients are uniformly distributed in the region $(-\Delta_0/2, \Delta_0/2)$, and $|\Delta H_Q(e^{j\omega})|$ is flattened in $-\pi\leq\omega\leq\pi$. N is the filter length of $H(z)$. Furthermore, $\Delta H_Q^W(z)$, given by Eq. (5 e), is estimated by

$$|\Delta H_Q^W(e^{j\omega})|=|W(e^{j\omega})|\left(N_F\frac{\Delta_0^2}{12}\right)^{\frac{1}{2}} \tag{11}$$

under the same assumptions. From Eqs. (9), (10) and (11), the proposed algorithm can improve the filter response from those obtained only rounding off the coefficients of $H(z)$ and $F(z)$ in the frequency band, where the following inequalities are held,

$$\frac{\Delta_l}{\|1/W(e^{j\omega})\|_2}<\sqrt{\frac{N}{N_F}}\Delta_0 \tag{12}$$

$$\frac{\Delta_l}{\|1/W(e^{j\omega})\|_2}<|W(e^{j\omega})|\Delta_0 \tag{13}$$

How to design $W(z)$ with short wordlength coefficients, which satisfies the above conditions, is one of the important design objectives, and will be discussed in the following section.

From Eqs. (9) and (11), $W(z)$ can be regarded as a weighting function or a shaping function for the coefficient roundoff error effects. Hence, $W(z)$ is called "weighting function" in this paper.

3. Design of Weighting Function

3.1 Design Factors for $W(z)$

Several factors exist which must be taken into account in designing $W(z)$. They include (a) pre-rounded short wordlength coefficients, (b) the conditions given by Eqs. (12) and (13), (c) the output noise reduction, which is caused by rounding off the multiplier outputs[26], and (d) optimizing the $H(z)$ characteristics with a short filter length. From factors (b)-(d), it can be roughly concluded that $W(z)$ is required to satisfy

$$|W(e^{j\omega})|\approx 1,\quad \omega\in\Omega_p \tag{14 a}$$

$$|W(e^{j\omega})|\ll 1,\quad \omega\in\Omega_s \tag{14 b}$$

where, Ω_p and Ω_s indicate the passband and stopband, respectively. As will be described in the next paragraph, $W(z)$ is handled as a fixed function in an exact $H(z)$ approximation procedure. Therefore, it is essential to use a low-order function for $W(z)$ to obtain a sub-optimum solution in the weighted Chebyshev sense for the given filter length N. Furthermore, it is desirable to use short wordlength or the power of 2 coefficients for $W(z)$, from the view point of simplifying hardware implementation.

3.2 Filter Response Improvement

One useful approach to designing $W(z)$, which satisfies the conditions given by Eq. (14) with a low-

order function, is to use the zeros, which are located on the unit circle in the stopband. In this paragraph, numerical examples demonstrating filter response improvement are provided by using

$$W(z)=\frac{1}{12}(1+z^{-1}+z^{-2})(1+2z^{-1}+z^{-2}). \tag{15}$$

The L_2 norm of $1/W(z)$ is approximately calculated as

$$\|1/W(e^{j\omega})\|_2^2\approx 5.2\times 10^4 \tag{16}$$

Furthermore, letting Δ_l be $7\Delta_0$†, and letting N and N_F be 200 and 196, respectively, the amplitude response deviations are estimated by

$$|\Delta H_Q(e^{j\omega})|^2=200\frac{\Delta_0^2}{12} \tag{17 a}$$

$$|\Delta H_Q^W(e^{j\omega})|^2=196\frac{\Delta_0^2}{12}|W(e^{j\omega})|^2 \tag{17 b}$$

$$|\Delta H_{Q0}(e^{j\omega})|^2\approx 196\frac{(7\Delta_0)^2}{12}/5.2\times 10^4. \tag{17 c}$$

They are shown in Fig. 1, where the square amplitude response are all normalized by Δ_0^2. This figure shows that $|\Delta H_{Q0}(e^{j\omega})|$ is greatly decreased from $|H_Q(e^{j\omega})|$. Furthermore, in the frequency bands, where $|W(e^{j\omega})|$ is not sufficiently small, $|\Delta H_Q^W(e^{j\omega})|$ is larger than $|\Delta H_{Q0}(e^{j\omega})|$. In other words, the optimized response is superior to that obtained by rounding off the $F(z)$ coefficients in the mini-max sense, as well as in the mean square sense.

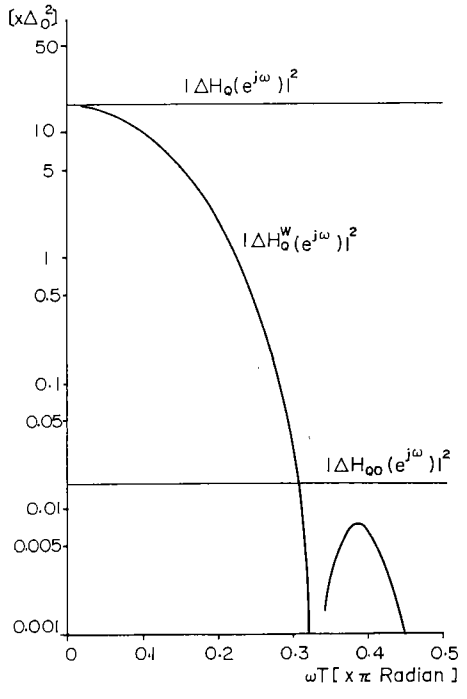

Fig. 1 Example for amplitude responses of squared error functions.

3.3 Transfer Function Approximation

After designing $W(z)$, $H(z)$ is first approximated in a cascade form $W(z)F(z)$ with the infinite precision coefficients of $F(z)$ as variables. Conventional approaches to FIR filter approximation[22],[25] can be basically applied. One approach by the Remez exchange method is described here.

Letting $D(\omega)$ and $U(\omega)$ be a desired amplitude response and a weighting function for error evaluation, respectively, a set of equations is expressed as

$$U(\omega_k)[D(\omega_k)-|H(e^{j\omega_k})|]=(-1)^{k\delta} \tag{18}$$
$$k=0, 1, \cdots, r$$

where $\{\omega_k\}$ is a set of the extremal frequencies[22]. From Eq. (1), Eq. (18) can be rewritten as

$$|W(e^{j\omega_k})|U(\omega_k)[|W^{-1}(e^{j\omega_k})|D(\omega_k)-|F(e^{j\omega_k})|]=(-1)^{k\delta} \tag{19}$$

Since $|W(e^{j\omega})|$ is fixed, $F(z)$ can be approximated using $|W(e^{j\omega})|U(\omega)$ and $|W^{-1}(e^{j\omega})|D(\omega)$ as modified weighting function and desired amplitude response, respectively.

The number of the extremal frequencies is equal to the degrees of freedom in $F(z)$. Therefore, the result obtained by solving Eq. (19) becomes a sub-optimal solution for the order of $H(z)$. For this reason, it is

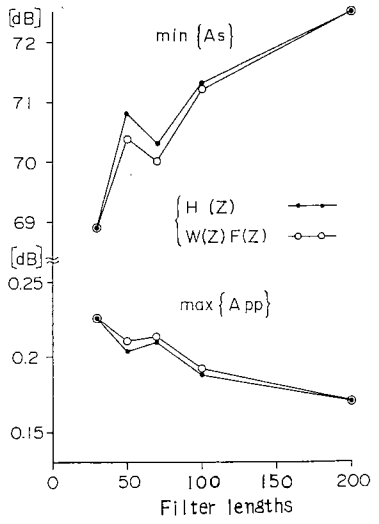

Fig. 2 Comparison between amplitude responses approximated with direct form $H(z)$ and separated form $W(z)F(z)$, where $W(z)$ is fixed, through Remez-exchange method. $\max\{A_{pp}\}$ and $\min\{A_s\}$ mean maximum passband ripple (peak-to-peak) and minimum stopband attenuation, respectively.

†It is assumed that the $F_{Q0}(z)$ coefficients are optimized in the region of $\pm 3\Delta_0$ around the $F_Q(z)$ coefficients. Hence, the $\Delta F_{Q0}(z)$ coefficients can be assumed to be distributed in $(-\Delta_l/2, \Delta_l/2)$, where $\Delta_l=7\Delta_0$.

necessary to employ a low-order function for $W(z)$, as previously mentioned.

Examples:

Differences between amplitude responses of $H(z)$ and $W(z)F(z)$, which are approximated by the Remez exchange method, are given in Fig. 2, where $\max\{A_{pp}\}$ and $\min\{A_s\}$ imply the maximum passband ripple (peak-to-peak) and the minimum stopband, attenuation, respectively. The cutoff frequency is varied according to the filter lengths, so that the resulting amplitude responses become mostly the same in all cases. The function given by Eq. (15), which has the zeros located in the stopband, is used for $W(z)$ in all cases. Figure 2 shows that the differences are sufficiently small in any case, and slightly decrease in ascending order of filter lengths.

4. Discrete Optimization Algorithm

4.1 Saving Computations

Another important feature of the proposed algorithm is to accomplish a considerable saving in computations. For this purpose, first, the mean square of an error function is employed as an error criterion, which can be transformed into the time domain through Parseval's relation[2]. Second, the $F(z)$ coefficients are divided into small groups in the discrete optimization procedure.

Letting Δh_n be an impulse response for $\Delta H(z)$, the error evaluation given by Eq. (4) can be transformed into

$$E=\sum_{n=0}^{N-1}\Delta h_n^2. \tag{20}$$

By letting Δf_n and w_n be impulse responses for $\Delta F(z)$ and $W(z)$, respectively, Δh_n is expressed by

$$\begin{aligned}\Delta h_n&=\sum_{m=n_1}^{n_2} w_m \Delta f_{n-m}\\ n_1&=\max\{o,\ n-N_F+1\}\\ n_2&=\min\{n,\ M-1\}\end{aligned} \tag{21}$$

where M is the order of $W(z)$. From Eqs. (20) and (21), E is rewritten as

$$E=\sum_{n=0}^{N-1}\Big(\sum_{m=n_1}^{n_2} w_m \Delta f_{n-m}\Big)^2. \tag{22}$$

Although all Δf_n are needed to calculate E given by Eq. (22), the idea behind the new approach is to reduce the number of Δf_n which are simultaneously optimized. This can be done by dividing a sum of Δh_n^2 into small groups. In other words, E is successively evaluated by using the partial sums of Δh_n^2 in the discrete optimization procedure. Estimation errors caused by this division are compensated for by overlapping the adjoining partial sums.

4.2 Discrete Optimization Procedure

As described above, E is divided into

$$E=\sum_{l=1}^{\left[\frac{N}{K-K'}\right]}\varepsilon_{l(K-K')},\quad K'<K \tag{23 a}$$

$$\varepsilon_{l(K-K')}=\sum_{i=0}^{K-1}\Delta h_{l(K-K')-i}{}^2. \tag{23 b}$$

The partial sum $\varepsilon_{l(K-K')}$ is individually minimized instead of E. The adjoining partial sums $\varepsilon_{l(K-K')}$ and $\varepsilon_{(l+1)(K-K')}$ include the same Δh_i, $l(K-K')-K'+1\leq i\leq l(K-K')$. This also means that they contain $(K'+M)$ common coefficients Δf_i. In the $\varepsilon_{l(K-K')}$ minimization, the coefficients in the set $\{\Delta f_i | l(K-K')-K+1-M\leq i\leq l(K-K')-L-L'\}$, which are already optimized at the ε_k, $k<l(K-K')$, minimization step, are fixed, and the $L+L'$ coefficients in the set $\{\Delta f_i | l(K-K')-L-L'+1\leq i\leq l(K-K')\}$ are optimized. After minimizing $\varepsilon_{l(K-K')}$, the L coefficients included in the set $\{\Delta f_i | l(K-K')-L-L'+1\leq i\leq l(K-K')-L'\}$ are fixed, and the remaining L' coefficients included in the set $\{\Delta f_i | l(K-K')-L'+1\leq i\leq l(K-K')\}$ are further optimized at the next $\varepsilon_{(l+1)(K-K')}$ minimization step. In the above expression, L is equal to $K-K'$.

Example for Partial Sums:

An example for the partial sums and Δf_i are illustrated here, using parameters $K=5$, $K'=1$, $M=4$, $L=4$, and $L'=1$.

:

ε_{10}:	Δf_2	Δf_3	Δf_4	Δf_5	$\underline{\Delta f_6}$	Δf_7	Δf_8	Δf_9	$\underline{\Delta f_{10}}$
ε_{14}:	Δf_6	Δf_7	Δf_8	Δf_9	$\underline{\Delta f_{10}}$	Δf_{11}	Δf_{12}	Δf_{13}	$\underline{\Delta f_{14}}$
ε_{18}:	Δf_{10}	Δf_{11}	Δf_{12}	Δf_{13}	$\underline{\Delta f_{14}}$	Δf_{15}	Δf_{16}	Δf_{17}	$\underline{\Delta f_{18}}$
ε_{22}:	Δf_{14}	Δf_{15}	Δf_{16}	Δf_{17}	$\underline{\Delta f_{18}}$	Δf_{19}	Δf_{20}	Δf_{21}	$\underline{\Delta f_{22}}$
:	Fixed				Optimized				

For instance, in the ε_{10} minimization, Δf_2-Δf_5 are fixed and Δf_6-Δf_{10} are optimized. In the ε_{14} minimization, Δf_{10} is further optimized using the result obtained in the previous step as the initial guess.

Initial Guess:

Coefficients of $\Delta F_Q(z)$, defined by Eq. (5c), are taken as the initial guess for Δf_n.

Search Method:

A global search method is employed. The reasons are explained as follows: First, local and heuristic search methods cannot avoid the risk of falling into a local minimum solution. Second, since the proposed approach drastically saves the number of assignments to be evaluated, the global search does not require a large amount of computing time.

Search Region:

A search region means the number of grids on which Δf_n are discretely optimized. The number of assignments for $\varepsilon_{l(K-K')}$ evaluation is exponentially proportional to the number of grids. Therefore, a moderate

search region must be chosen.

4.3 Modified Weighting Function

Since the error is evaluated with the mean square of $|\Delta H(e^{j\omega})|$, strictly speaking, the obtained solution is optimum only in the mean square sense. Therefore, when the mini-max error criterion is employed, the weighting function, used in the discrete optimization procedure, must be somewhat modified so that max $\{|\Delta H(e^{j\omega})|\}$ is minimized. An example for the modification is provided in Sect. 5.

4.4 Number of Computations

Since the number of Δf_n, which are included in the ε_k expression, is $L+L'$, all possible combinations of Δf_n in evaluating ε_k become $P^{L+L'}$, where P is the number of grids. Furthermore, the number of ε_k, which are included in the E expression, is $[N_F/(K-K')]$ which implies the maxmum integer not exceeding $N_F/(K-K')$. Hence, the total number of assignments, required in the E evaluation, becomes

$$N(E)=P^{L+L'}\left[\frac{N_F}{K-K'}\right] \tag{25}$$

The numbers of real multiplications and additions, required in the ε_k calculation per Δf_n combination, are both $(M+2)K$. The total numbers are given by

$$N_{\text{Mult}}(E)=N_{\text{Add}}(E)=P^{L+L'}\left[\frac{N_F}{K-K'}\right](M+2)K. \tag{26}$$

On the other hand, the error evaluation by Eq. (4) requires

$$N'_{\text{Mult}}(E)=N'_{\text{Add}}(E)=P^{L+L'}\left[\frac{N_F}{K-K'}\right](L+L')2N_F \tag{27}$$

under the assumption that the number of Δf_n to be simultaneously optimized is $L+L'$, and the number of frequency points, on which $|\Delta H(e^{j\omega})|$ is calculated, is $2N_F$. A saving of computations, achieved by the proposed method, can be expressed by

$$\frac{N_{\text{Mult}}(E)}{N'_{\text{Mult}}(E)}=\frac{(M+2)K}{(L+L')2N_F}. \tag{28}$$

For example, by using $N_F=200$, $K=5$, $K'=1$, $M=4$, $L=4$ and $L'=1$, the ratio becomes 1.5%. This demonstrates that a considerable saving in computing time can be achieved.

4.5 Flow Chart

A flow chart for the proposed algorithm is shown in Fig. 3. Necessary filter length and tolerance are estimated, taking into account filter response improvement rates achieved by the proposed approach.

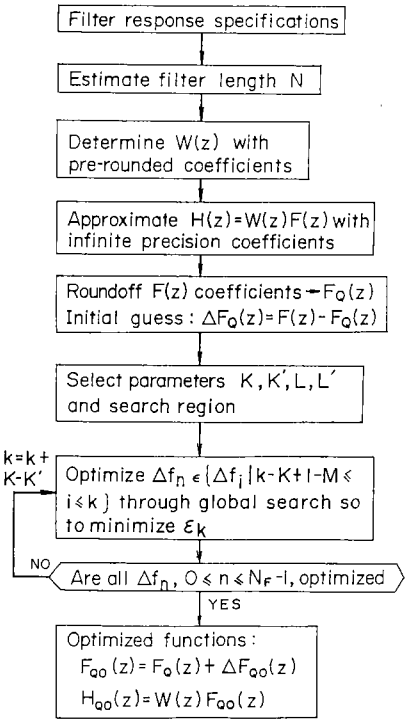

Fig. 3 Flow chart for proposed discrete optimization method.

5. Design Examples

5.1 Design Parameters

Table 1 shows filter responses and design parameters utilized to examine the proposed algorithm performances. The $W(z)$ coefficients are restricted to integer values in these examples. $F(z)$ with infinite precision coefficients is approximated through the Remez exchange method. The rounded $F(z)$ coefficients are represented with 8, 10, 12, and 14 bits wordlengths, which do not include a sign bit. The maximum coefficient value is normalized to unity. Frequency response error is evaluated in the mini-max sense. Thus, a weighting function $W^*(z)$, used in the discrete optimization procedure, is modified from $W(z)$ so that the maximum error is decreased. The search region is $\pm\Delta_0$ or $\pm 3\Delta_0$ around the $\Delta F_Q(z)$ coefficients, where Δ_0 corresponds to the least significant bit. This means that the numbers of grids are three and seven, respectively.

5.2 Optimized Filter Responses

Coefficient Value Scaling:

The $F(z)$ coefficient values become slightly larger than those for $H(z)$, by separating $W(z)$. This is equivalent to improving coefficient accuracy under the same wordlengths. The $F(z)$ coefficient wordlengths are equivalently increased by η bits

Table 1 Filter specifications and discrete optimization parameters.

Parameter	L P F	B P F
H(z)	200 taps	200 taps
Sampling freq.	400 Hz	400 Hz
Passband	0-50 Hz	57-142 Hz
Ripple (Ap)	±0.085 dB	±0.035 dB
Stopband	56-200Hz	0-50 Hz 150-200 Hz
Attenuation	72.5 dB	80.0 dB
W (z)	$(1+2z^{-1}+z^{-2})$ $\times(1+z^{-1}+z^{-2})$	$(1-z^{-2})^2$
$W^*(z)$	$(1+2z^{-1}+z^{-2})^2$	$(1-z^{-2})^2$
F (z)	196 taps	196 taps
Coefficient wordlengths	8,10,12,14 bits	8,10,12,14 bits
Search region	±Δo, ±3Δo	±Δo, ±Δo
Search method	Global	Global
K	5	5
K'	1	1
L	4	4
L'	1	1
Error criterion	mini-max	mini-max

$$\eta=\log_2\{\max_n(f_n)/\max_n(h_n)\}. \tag{29}$$

Numerical examples for the filters and $W(z)$ listed in Table 1 are given in Table 2. The wordlength increases are around 0.1-0.3 bits, which are negligible, compared with improvements gained by the proposed method, as will be demonstrated.

Coefficient Values and Amplitude Responses:

Table 3 gives coefficient values for $F(z)$, $F_Q(z)$ and $F_{Qo}(z)$. Figure 4 shows amplitude responses for $H(z)$, $H_Q(z)$ and $H_{Qo}(z)$, in the case of LPF with 10 bit wordlengths and the search region $\pm 3\Delta_Q$. The Amplitude response for $H_{Qo}(z)$ is sufficiently improved from those for $H_Q(z)$ in a whole frequency band. Furthermore, in the passband and the relatively low stopband, the proposed approach can decrease the deviations from those for $H_Q^W(z)$.

Table 2 Coefficient value scaling effects by separating weighting function $W(z)$ from $H(z)$.

Type	W(Z)	$\max_i\{fi\}$	$\max_i\{hi\}$	η(bits)
LPF	2 nd	0.2695	0.2546	0.08
	4 th	0.2927	0.2546	0.20
BPF	2 nd	0.3461	0.3134	0.14
	4 th	0.3813	0.3163	0.27

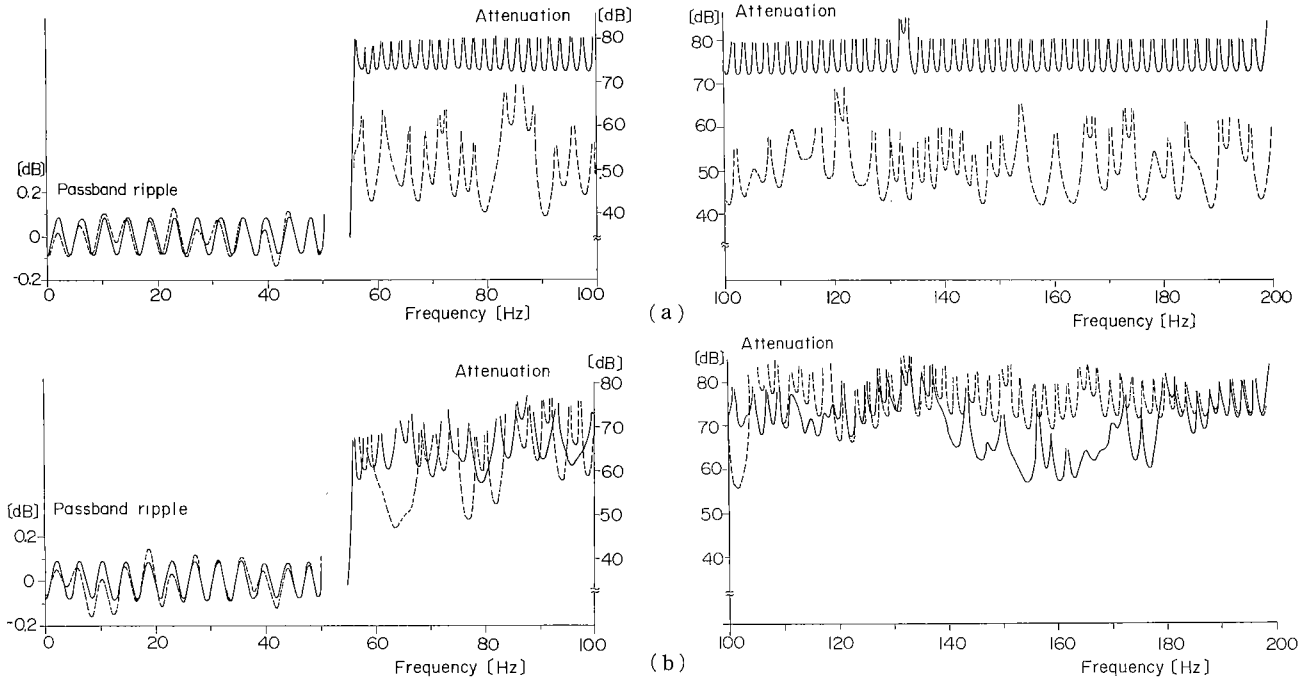

Fig. 4 Optimized frequency responses. (a) solid line: $H(z)$ with infinite precision coefficients, dashed line: $H_Q(z)$ with rounded coefficients. (b) solid line: $W(z)F_{Qo}(z)$ with optimized coefficients, dashed line: $W(z)F_Q(z)$ with rounded coefficients.

Table 3 Optimized $F(z)$ coefficients for LPF with 4th-order $W(z)$ and with seven grids. $F(z)$, $F_Q(z)$ and $F_{QO}(z)$ are transfer functions with infinite precision, rounded and optimized coefficients, respectively. Coefficient values are all scaled by 2^{10}.

NO.	F(z)	$F_Q(z)$	$F_{QO}(z)$	NO.	F(z)	$F_Q(z)$	$F_{QO}(z)$
0	-0.79	-1	-1	50	-19.09	-19	-20
1	5.20	5	6	51	29.33	29	31
2	-5.32	-5	-7	52	-26.15	-26	-28
3	9.90	10	12	53	16.37	16	18
4	-6.64	-7	-8	54	-29.96	-30	-31
5	8.18	8	8	55	19.68	20	20
6	-9.33	-9	-8	56	-20.58	-21	-20
7	10.17	10	9	57	31.96	32	30
8	-14.86	-15	-15	58	-14.16	-14	-11
9	14.34	14	16	59	23.70	24	21
10	-13.78	-14	-16	60	-25.47	-25	-24
11	15.07	15	16	61	10.59	11	10
12	-13.58	-14	-13	62	-30.39	-30	-30
13	17.19	17	16	63	19.58	20	19
14	-19.77	-20	-19	64	-12.53	-13	-11
15	18.34	18	19	65	30.80	31	28
16	-20.39	-20	-22	66	-8.26	-8	-5
17	18.32	18	20	67	18.04	18	16
18	-18.04	-18	-19	68	-28.90	-29	-29
19	22.72	23	23	69	3.57	4	6
20	-21.51	-22	-22	70	-26.78	-27	-30
21	23.39	23	25	71	17.81	18	20
22	-24.11	-24	-27	72	-0.69	-1	-1
23	19.72	20	23	73	32.35	32	31
24	-23.59	-24	-26	74	-5.70	-6	-4
25	24.09	24	25	75	9.50	9	9
26	-23.45	-23	-23	76	-32.50	-33	-33
27	28.26	28	27	77	-8.51	-9	-8
28	-23.48	-23	-22	78	-20.33	-20	-20
29	23.23	23	22	79	21.61	22	20
30	-26.64	-27	-26	80	13.24	13	16
31	22.58	23	23	81	36.36	36	33
32	-28.17	-28	-30	82	-2.57	-3	0
33	28.30	28	31	83	-8.51	-9	-9
34	-22.85	-23	-26	84	-41.94	-42	-43
35	28.21	28	31	85	-23.03	-23	-21
36	-23.84	-24	-26	86	-13.38	-13	-16
37	24.16	24	26	87	36.37	36	39
38	-31.26	-31	-33	88	43.23	43	41
39	23.82	24	26	89	46.13	46	48
40	-27.04	-27	-30	90	-2.48	-2	-4
41	27.90	28	31	91	-47.10	-47	-46
42	-20.10	-20	-23	92	-88.31	-88	-89
43	29.62	30	32	93	-65.79	-66	-65
44	-26.70	-27	-28	94	-0.44	0	-2
45	22.97	23	23	95	115.84	116	118
46	-31.37	-31	-30	96	228.58	229	226
47	20.53	21	18	97	299.70	300	302
48	-23.39	-23	-21	98	299.70	300	298
49	29.94	30	29	99	228.58	229	230
100	115.84	116	114	150	22.97	23	24
101	-0.44	0	2	151	-26.70	-27	-29
102	-65.79	-66	-69	152	29.62	30	32
103	-88.31	-88	-85	153	-20.10	-20	-22
104	-47.10	-47	-50	154	27.90	28	29
105	-2.48	-2	0	155	-27.04	-27	-27
106	46.13	46	44	156	23.82	24	23
107	43.23	43	45	157	-31.26	-31	-30
108	36.37	36	35	158	24.16	24	23
109	-13.38	-13	-12	159	-23.84	-24	-23
110	-23.03	-23	-25	160	28.21	28	28
111	-41.94	-42	-39	161	-22.85	-23	-23
112	-8.51	-9	-12	162	28.30	28	28
113	-2.57	-3	0	163	-28.17	-28	-27
114	36.36	36	36	164	22.58	23	21
115	13.24	13	11	165	-26.64	-27	-25
116	21.61	22	25	166	23.23	23	22
117	-20.33	-20	-23	167	-23.48	-23	-23
118	-8.51	-9	-8	168	28.26	28	29
119	-32.50	-33	-31	169	-23.45	-23	-25
120	9.50	9	7	170	24.09	24	26
121	-5.70	-6	-3	171	-23.59	-24	-25
122	32.35	32	30	172	19.72	20	20
123	-0.69	-1	1	173	-24.11	-24	-23
124	17.81	18	17	174	23.39	23	21
125	-26.78	-27	-27	175	-21.51	-22	-19
126	3.57	4	5	176	22.72	23	21
127	-28.90	-29	-31	177	-18.04	-18	-18
128	18.04	18	20	178	18.32	18	20
129	-8.26	-8	-9	179	-20.39	-20	-23
130	30.80	31	30	180	18.34	18	21
131	-12.53	-13	-11	181	-19.77	-20	-22
132	19.58	20	18	182	17.19	17	19
133	-30.39	-30	-29	183	-13.58	-14	-15
134	10.59	11	10	184	15.07	15	16
135	-25.47	-25	-26	185	-13.78	-14	-14
136	23.70	24	25	186	14.34	14	14
137	-14.16	-14	-16	187	-14.86	-15	-15
138	31.96	32	34	188	10.17	10	11
139	-20.58	-21	-22	189	-9.33	-9	-11
140	19.68	20	20	190	8.18	8	10
141	-29.96	-30	-29	191	-6.64	-7	-7
142	16.37	16	14	192	9.90	10	8
143	-26.15	-26	-23	193	-5.32	-5	-2
144	29.33	29	26	194	5.20	5	2
145	-19.09	-19	-16	195	-0.79	-1	1
146	29.94	30	27				
147	-23.39	-23	-21				
148	20.53	21	19				
149	-31.37	-31	-31				

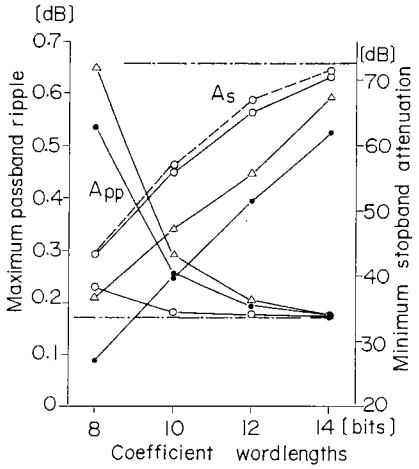

Fig. 5 Frequency response improvements for LPF. Symbols ●, △ and ○ correspond to $H_Q(z)$, $W(z)F_Q(z)$ and $W(z)F_{QO}(z)$, respectively. $H(z)$ with infinite accuracy coefficients is shown by dashed and dotted line.

5.3 Frequncy Response Improvement Rates

Maximum passband ripple (peak-to-peak) A_{pp} and minimum stopband attenuation A_s are shown in Figs. 5 and 6. The solid and dashed lines for $H_{QO}(z)$ indicate the cases where the numbers of grids are three and seven, respectively.

The proposed method can provide sufficiently reduced passband ripples compared with those obtained by only rounding off. $W(z)F_Q(z)$ cannot inherently decrease the passband ripple, because $|W(e^{j\omega})|$ is required to be approximately unity in the passband, from the viewpoint of output noise caused by rounding off the variables in the filter[26].

Stopband Attenuation:

$H_{QO}(z)$ is superior to $H_Q^W(z)$ in the LPF case. The difference between them in the BPF case is, however, small. The reason can be explained as, $|W(e^{j\omega})|$ has relatively small magnitude in the whole stopband.

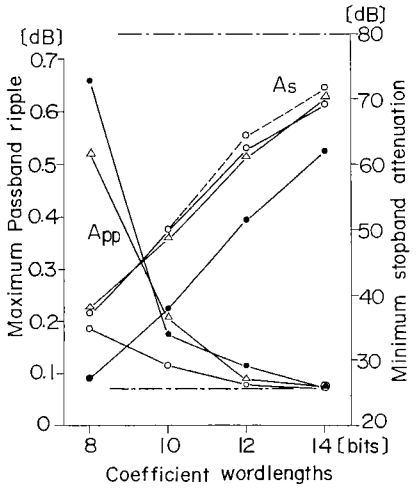

Fig. 6 Frequency response improvements for BPF.

Coefficient Wordlength Reductions:

The coefficient wordlengths, with which the same filter responses are obtained through the proposed method as those for $H_Q(z)$ and $H_Q^W(z)$, can be reduced by 3 and 2 bits in the cases of LPF and BEF, respectively.

Taking into account filter response improvements for arbitrary filter responses in both the passband and stopband, the new approach efficiency can be confirmed.

5.4 Weighting Function

A combination of $W(z)$ and $W^*(z)$ is further optimized by using pre-rounded real values for the $W(z)$ coefficients. Hence, further improvements in filter responses can be expected.

5.5 Computing Time

The execution time required in optimizing LPF with the search region $\pm 3\Delta_0$ on a general purpose computer NEC ACOS 900 was 97 seconds. This result obviously allows practical usage of the proposed method for high-order FIR filters.

6. Conclusion

This paper has presented a discrete optimization method, which is computationally more efficient for high-order FIR filters. The proposed approach has two important features. First, the error spectrum is shaped so as to be effectively suppressed by a weighting function. Second, in order to drastically save the number of computations, the coefficients to be discretely optimized are divided into small groups. Design examples for 200 tap FIR filters demonstrated high efficiency for the proposed method in filter response improvements and computing time.

References

(1) B. Liu: "Effect of finite word length on the accuracy of digital filters—A review", IEEE Trans. Circuit Theory, **CT-18**, pp. 670-677 (Nov. 1971).

(2) L. R. Rabiner and B. Gold: "Theory and Application of Digital Signal Processing", Prentice-Hall, Inc., New Jersey (1975).

(3) J. B. Knowles and E. M. Olcayto: "Coefficient accuracy and digital filter response", IEEE Trans. Circuits & Syst., **CAS-15**, pp. 31-41 (March 1968).

(4) D. S. K. Chan and L. R. Rabiner: "Analysis of quantization errors in the direct form for finite impulse response digital filters", IEEE Trans. Audio & Electroacoust., **AU-21**, pp. 354-366 (Aug. 1973).

(5) R. E. Crochiere: "A new statistical approach to the coefficient word length problem for digital filters", IEEE Trans. Circuits & Syst., **CAS-22**, pp. 190-196 (March 1975).

(6) F. Grenez: "Design of f. i. r. linear phase digital filters to minimise the statistical word length of the coefficients", Electron. Circuits Syst., **1**, 5, pp. 181-185 (Sept. 1977).

(7) U. Heute: "A subroutine for finite wordlength FIR filter design", Programs for Digital Signal Processing, ed. DSP Committee, IEEE ASSP Soc., IEEE Press, pp. 5.4.1-5.4.20 (1979).

(8) M. Suk and S. K. Mitra: "Computer-aided design of digital filters with finite word lengths", IEEE Trans. Audio & Electroacoust., **AU-20**, pp. 356-363 (Dec. 1972).

(9) N. I. Smith: "A random-search method for designing finite-wordlength recursive digital filters", IEEE Trans. Acoust., Speech & Signal Process., **ASSP-27**, pp. 40-46 (Feb. 1979).

(10) E. Avehaus: "On the design of digital filters with coefficients of limited word length", IEEE Trans. Audio & Electroacoust., **AU-20**, pp. 206-212 (Aug. 1972).

(11) H. K. Kwan: "On the problem of designing IIR digital filters with short coefficient word lengths", IEEE Trans. Acoust., Speech & Signal Process., **ASSP-27**, pp. 620-624 (Dec. 1979).

(12) C. Charalambous and M. J. Best: "Optimum of recursive digital filters with finite word length", IEEE Trans. Acoust., Speech & Signal Process., **ASSP-22**, pp. 424-431 (Dec. 1974).

(13) J. W. Bandler, B. L. Bardakjian and J. H. K. Chen: "Design of recursive digital filters with optimized word length coefficients", Computer Aided Design, **7**, pp. 153-156 (July 1975).

(14) P. Chambon and A. Desblache: "Integer coefficients optimization of digital filters", Proc. IEEE Int. Symp. Circuits & Syst., pp. 461-464 (1976).

(15) K. Steiglitz: "Designing short-word recursive digital filters", Proc. 9th Annu. Allerton Conf. Circuit and System Theory, pp. 778-788 (Oct. 1971).

(16) F. Brglez: "Digital filter design with short word-length coefficients", IEEE Trans. Circuits & Syst., **CAS-25**, pp. 1044-1050 (Dec. 1978).

(17) Y. Chen, S. M. Kang and T. G. Marshall: "The optimal design of CCD transversal filters using mixed-integer programming techniques", Proc. IEEE Int. Symp. Circuits Syst., pp. 748-751 (May 1978).

(18) D. M. Kodek: "Design of optimal finite wordlength FIR digital filters using integer programming techniques", IEEE Trans. Acoust., Speech & Signal Process., **ASSP-28**, pp. 304-307 (June 1980).

(19) V. B. Lawrence and A. C. Salazar: "Finite precision design of linear-phase FIR filters", Bell Syst. Tech. J., **59**, pp. 1575-1598 (Nov. 1980).

(20) D. Kodek and K. Steiglitz: "Comparison of optimal and local search methods for designing finite wordlength FIR

digital filters", IEEE Trans. Circuits & Syst., **CAS-28**, pp. 28-32 (Jan. 1981).

(21) K. Nakayama: "A discrete optimization method for high-order FIR filters with finite wordlength coefficients", Paper of Technical Group, **TGCAS 81-58**, IECE Japan (Sept. 1981).

(22) T. W. Parks and J. H. McClellan: "Chebyshev approximation for nonrecursive digital filters with linear phase", IEEE Trans. Circuit Theory, **CT-19**, pp. 189-194 (Mar. 1972).

(23) S. K. Tewksbury and R. W. Hallock: Oversampled, linear predictive and noise-shaping coders of order $N>1$", IEEE Trans. Circuits & Syst., **CAS-25**, pp. 436-447 (July 1978).

(24) T. Thong and B. Liu: "Error spectrum shaping in narrow-band recursive filters", IEEE Trans. Acoust., Speech & Signal Processing, **ASSP-25** pp. 200-203 (April 1977).

(25) L. R. Rabiner: "The design of finite impulse response digital filters using linear programming techniques", Bell Syst. Tech. J., **51**, pp. 1177-1198 (July-Aug. 1972).

(26) D. S. K. Chan and L. R. Rabiner: "An algorithm for minimizing roundoff noise in cascade realizations of finite impulse response digital filters", Bell Syst. Tech. J., **52**, pp. 347-385 (Mar. 1973).

Kenji Nakayama received the B. E. and Dr. deqrees in electronics engineering from the Tokyo Institute of Technology (TIT), Tokyo, Japan, in 1971 and 1983, respectively.

From 1971 to 1972 he was engaged in the research of classical network theory at the TIT. Since he joined the Transmission Div., NEC Corporation in 1972, he has worked on the research and development of filter design techniques for LC, digital and switched-capacitor filters, and computationally efficient algorithms in digital signal processing. His current research interests further include singnal reconstruction and multi-dimensional signal processing. He is now supervisor of the C&C Systems Research Laboratories. He is the author of "Design and Application of *SC* Newtworks" (in Japanese), Tokai Univ. Press, Tokyo, Japan.

Dr. Nakayama is a senior member of the Institute of Electrical and Electronics Engineers.