A DISCRETE OPTIMIZATION METHOD FOR HIGH-ORDER FIR FILTERS WITH FINITE WORDLENGTH COEFFICIENTS

Kenji Nakayama

Transmission Div., Nippon Electric Co., Ltd.
Kawasaki, 211 JAPAN

ABSTRACT

This paper suggests a discrete optimization method which can solve high order FIR filter problems within a practically reasonable computing time. The error spectrum caused by rounding off the coefficients is shaped through the discrete optimization so to be effectively cancelled, in the L_2 norm sense, by other factors connected in cascade. In order to save computing time, the error spectrum is evaluated in a time domain, and parameters are divided into small groups during searching for the optimum solution. LPF and BPF design examples, with 200 lengths, show the proposed approach can reduce coefficient wordlengths by 2 or 3 bits, compared with results obtained by only rounding off. The execution time on the general purpose computer, ACOS System 900, is 97 seconds.

INTRODUCTION

Digital filter element values are basically expressed with finite precision, that is discrete value. Furthermore, their circuit complexities are highly dependent on a number of quantization steps. For this reason, it has become very important to design discrete valued filters satisfying a desired response with minimum wordlength elements. Several useful discrete optimization methods for IIR and relatively low order FIR filters have been proposed up to now. They include random search [1], [2], univarate search [3], branch and bound [4], [5], Hook-Jeeves method [6], and a combination of rounding off and iterative optimization [7]. On the other hand, discrete optimization for high order FIR filters has been tried by other methods, mainly based on mixed-integer programming algorithms [8] - [10]. However, they require much computing time for high order FIR filters as yet.

This paper proposes one approach, which is particularly useful for high order FIR filters, from the computing time viewpoint.

NEW DISCRETE OPTIMIZATION

Principle

Principle of the proposed discrete optimization algorithm can be summarized as follows:

(1) A transfer function is basically realized in a cascade form

$$H(z) = \prod_{i=1}^{I} H_i(z), \quad z = e^{j\omega T}, \text{ T: Sampling period.} \tag{1}$$

(2) Letting $\Delta H_j(z)$ be the error function for $H_j(z)$, caused by rounding off its coefficients, the coefficients are optimized so that the error spectrum $|\Delta H_j(e^{j\omega})|$ is cancelled by $\overline{H}_j(z)$ in the L_2 norm sense, where T is taken as unity, and

$$\overline{H}_j(z) = \prod_{\substack{i=1 \\ \neq j}}^{I} H_i(z) \tag{2}$$

The discrete optimization can be formulated as

$$E_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\Delta H_j(e^{j\omega})\overline{H}_j(e^{j\omega})|^2 d\omega \tag{3}$$

where E_j is minimized for all $H_j(z)$.

Filter Response Improvement

Let $H(z)$ be expressed as

$$H(z) = H_1(z)H_2(z) \tag{4}$$

The following discussions are valid for the case of large number of factors connected in cascade. It is assumed here that the error spectrum shaping is completely accomplished by the discrete optimization, and parameters are optimized within ℓ bit variable wordlengths from the least significant bit (LSB). The following relations result

$$|\Delta H_{1Q0}(e^{j\omega})H_2(e^{j\omega})| = c_1 \tag{5a}$$

$$|H_1(e^{j\omega})\, \Delta H_{2Q0}(e^{j\omega})| = c_2 \tag{5b}$$

where c_1 and c_2 are constant values. Furthermore, when $H_{1Q0}(z)$ and $H_{2Q0}(z)$ coefficients are assumed to be uniformly distributed in the region $[-\Delta_\ell, \Delta_\ell]$, their power can be expressed as

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}|\Delta H_{iQ0}(e^{j\omega})|^2 d\omega = \frac{\Delta_\ell^2}{3} N_i, \quad i = 1, 2 \tag{6}$$

where N_i is the number of taps for $H_i(z)$. From Eqs. (5) and (6),

$$\frac{c_i^2}{2\pi}\int_{-\pi}^{\pi}|H_j^{-1}(e^{j\omega})|^2 d\omega = \frac{\Delta_\ell^2}{3} N_i, \quad i = 1, 2, \; j = 2, 1. \tag{7}$$

From this relation, c_i^2 can be obtained as

$$c_i^2 = \left(\frac{\Delta_\ell^2}{3} N_i\right) / \|H_j^{-1}(e^{j\omega})\|_2^2, \quad i = 1, 2, \; j = 2, 1 \tag{8}$$

where $\|\cdot\|_p$ is an L_p norm. Assuming the mutual correlation of $\Delta H_{1Q0}(z)$ and $\Delta H_{2Q0}(z)$ to be zero, the optimized $H(z)$ error spectrum is obtained as

$$|\Delta H_{Q0}(e^{j\omega})|^2 = c_1^2 + c_2^2 \tag{9}$$

On the other hand, the error spectrum, caused by only rounding off the $H_1(z)$ and $H_2(z)$ coefficients, becomes

$$c_i'^2 = \frac{\Delta_0^2}{12} N_i |H_j(e^{j\omega})|^2, \quad i = 1, 2, \; j = 2, 1 \tag{10}$$

CH 1746-7/82/0000 - 0484 $ 00.75 © 1982 IEEE

where the $\Delta H_{1Q}(z)$ and $\Delta H_{2Q}(z)$ coefficients are also assumed to be uniformly distributed in the region $[-\Delta_0/2, \Delta_0/2]$. The H(z) error spectrum is expressed

$$|\Delta H_Q(e^{j\omega})|^2 = c_1'^2 + c_2'^2 . \tag{11}$$

If the following relation is satisfied

$$|\Delta H_{Q0}(e^{j\omega})| < |\Delta H_Q(e^{j\omega})| \tag{12}$$

then, filter response improvement can be achieved in the L_2 norm sense. The left and the right hand sides of Eq. (12) are determined by L_2 and L_∞ norms for $H_i(z)$, respectively. The frequency regions satisfying Eq. (12) usually appear in frequency selective filters, because an L_2 norm is well reduced from an L_∞ norm.

Example

For simplicity, the amplitudes of $H_1(z)$ and $H_2(z)$ are assumed to be approximated by

$$|H_1(e^{j\omega})| = \frac{1}{1+\omega^3}, \ -\pi \le \omega \le \pi \tag{13a}$$

$$|H_2(e^{j\omega})| = \frac{1}{1+\omega^4}, \ -\pi \le \omega \le \pi . \tag{13b}$$

From Eqs. (8) and (9),

$$|\Delta H_{Q0}(e^{j\omega})|^2 = \frac{\Delta_\ell^2}{3}\left\{\frac{N_1}{\|H_2^{-1}(e^{j\omega})\|_2^2} + \frac{N_2}{\|H_1^{-1}(e^{j\omega})\|_2^2}\right\}. \tag{14}$$

The L_2 norms for $H_1^{-1}(z)$ and $H_2^{-1}(z)$ are calculated as

$$\|H_1^{-1}(e^{j\omega})\|_2^2 = 153.8 \tag{15a}$$

$$\|H_2^{-1}(e^{j\omega})\|_2^2 = 1094.2 . \tag{15b}$$

Letting ℓ be 2 bits, Δ_ℓ relates to Δ_0 as

$$\Delta_\ell = 3.5\Delta_0 . \tag{16}$$

Furthermore, by setting N_1 and N_2 to 100,

$$|\Delta H_{Q0}(e^{j\omega})|^2 = 3.03\Delta_0^2 . \tag{17}$$

On the other hand, $|\Delta H_Q(e^{j\omega})|^2$ can be expressed as

$$|\Delta H_Q(e^{j\omega})|^2 = \frac{\Delta_0^2}{12}\left\{N_1|H_2(e^{j\omega})|^2 + N_2|H_1(e^{j\omega})|^2\right\}. \tag{18}$$

From Eq. (13),

$$|\Delta H_Q(e^{j\omega})|^2 = \frac{100}{12}\Delta_0^2\left\{\frac{1}{(1+\omega^3)^2} + \frac{1}{(1+\omega^4)^2}\right\} . \tag{19}$$

$|\Delta H_Q(e^{j\omega})|^2$ and $|\Delta H_{Q0}(e^{j\omega})|^2$ are shown in Fig. 1. In this figure, they are normalized by Δ_0^2, and ω is normalized by π. The error spectrum, caused by rounding off the coefficients of the whole transfer function H(z), is also shown with a dashed line.

DISCRETE OPTIMIZATION PROCEDURE

Detailed discrete optimization procedure is described here. The aim of the following procedure is to drastically save computing time. The following two contrivances are introduced for this purpose.

(1) Evaluate the error spectrum in a time domain in order to avoid frequency response calculation at each sarching step, which consumes a large amount of computing time.

(2) Using the low order weighting function $W^*(z)$, divide parameters into small groups during searching for the optimum solution.

For simplicity, H(z) is expressed as

$$H(z) = W(z)F(z) \tag{20}$$

where W(z) is a weighting function for F(z). In the discrete optimization procedure, instead of W(z) another weighting function $W^*(z)$ is utilized in order to decrease a number of parameters and to shape an error function $\Delta F(z)$ so to be effectively cancelled by W(z). Let Δf_n and w_n^* be the impulse response for $\Delta F(z)$ and $W^*(z)$, respectively. The $\Delta H(z)$ impulse response can be expressed as

$$\Delta h_n = \sum_{m=0}^{\tilde{n}} w_m^* \Delta f_{n-m}, \ \tilde{n} = \min\{n, M\} \tag{21}$$

where M is degree of $W^*(z)$. From the Parseval relation, the error E_j by Eq. (3) can be evaluated in a time domain as

$$E = \sum_{n=0}^{N-1} \Delta h_n^2 \tag{22}$$

where N is a number of H(z) taps. From Eqs. (21) and (22),

$$E = \sum_{n=0}^{N-1}\left\{\sum_{m=0}^{\tilde{n}} w_m^* \Delta f_{n-m}\right\}^2, \ \tilde{n} = \min\{n, M\}. \tag{23}$$

Assume that n satisfies

$$M \le n \tag{24}$$

then, Δh_n and Δh_{n+M+1} are expressed as

$$\Delta h_n = \sum_{m=0}^{M} w_m^* \Delta f_{n-m} , \tag{25a}$$

$$\Delta h_{n+M+1} = \sum_{m=0}^{M} w_m^* \Delta f_{n+M+1-m} . \tag{25b}$$

Since they do not contain the same parameters, they can be independently minimized. This basically indicates the possibility of parameter division in the error evaluation. Let the partial sum from Δh_{k-K+1}^2 to Δh_k^2 be $\varepsilon\{k\}$

$$\varepsilon\{k\} = \sum_{i=0}^{K-1} \Delta h_{k-i}^2, \ K < k . \tag{26}$$

$\varepsilon\{k\}$ consists of $\Delta f_{k-K+1-M} \sim \Delta f_k$, $(K+M < k)$, and can be minimized by searching for their optimum solution. In the proposed method, k takes the following value

$$k = L(K-K'), \ K' < K, \ L = 0, 1, \ldots \tag{27}$$

$\varepsilon\{L(K-K')\}$ and $\varepsilon\{(L+1)(K-K')\}$ contain $(K'+M-1)$ common parameters. In the $\varepsilon\{L(K-K')\}$ minimization, the parameters in the set $\{\Delta f_i | L(K-K')-K+1-M \le i \le L(K-K')-L-L'\}$ are already fixed at the $\varepsilon\{k\}$, $(k < L(K-K'))$ minimization step, and the L+L' parameters in the set $\{\Delta f_i | L(K-K')-L-L'+1 \le i \le L(K-K')\}$ are optimized. After minimizing $\varepsilon\{L(K-K')\}$, the L parameters included in the set $\{\Delta f_i | L(K-K')-L-L'+1 \le i \le L(K-K')-L'\}$ are fixed at this step and the remaining L' parameters, included in the set, $\{\Delta f_i | L(K-K')-L'+1 \le i \le L(K-K')\}$ are used in minimizing

$\varepsilon\{(L+1)(K-K')\}$ once more. Using the same parameters in minimizing the adjoining error functions is to minimize error evaluation loss by the parameter division. The number of possible parameter value combinations for $\varepsilon\{L(K-K')\}$ is the $(L+L')$th power of P, where P is a number of discrete value steps for each parameter. Therefore, a small $(L+L')$ value means drastically saving computing time.

Initial Guess: The quantization error, caused by rounding off the approximated coefficients with infinite wordlengths, is taken as the initial guess for Δf_i.

Searching Method: Since, the number of possible assignments is strongly reduced, a global searching method is employed in this paper. Another methods, such as local search and heuristic search methods, cannot avoid a risk of falling into the local minimum solution.

DESIGN EXAMPLES

Design Parameters

A lowpass filter (LPF) and a bandpass filter (BPF), shown in Table 1, are taken as design examples. Design parameters for the discrete optimization are also listed in Table 1. $F(z)$ is approximated by the Remez-exchange method [11] using $W(z)$ as a fixed weighting function. Coefficient wordlengths (*) do not include a sign bit, and the $F(z)$ coefficient values are not normalized. In the LPF case, zeros of $W^*(z)$ are all concentrated at π radian on the ωT axis. However, parameters Δf_i take only discrete values in the restricted range, and $|\Delta F_{Q0}(e^{j\omega})|$ is not strictly proportional to $|W^*(e^{j\omega})|^{-1}$. For this reason, zeros of $W(z)$ are set on $2\pi/3$ and π radian.

Optimized Filter Responses

Optimized filter responses are shown in Fig. 2, for the case of LPF with 10 bit coefficients. The passband ripple in the optimized response, $W(z)F_{Q0}(z)$, is almost the same as that of the original response $W(z)\cdot F_{\infty}(z)$. The stopband attenuation for $W(z)F_Q(z)$ is somewhat improved, from that of $H_Q(z)$, whose coefficients are rounded off only, because the $F_Q(z)$ error spectrum is cancelled by $W(z)$ in the stopband. However, in the frequency range where the condition $|W(e^{j\omega})| \ll 1$ is not satisfied, that is, in the passband and the lower side in the stopband, the improvement rate becomes lower than $W(z)F_{Q0}(z)$.

Frequency Response Improvement Rate

The maximum passband ripple (peak to peak) and the minimum stopband attenuation in $H_\infty(z)$, $H_Q(z)$, $W(z)F_Q(z)$ and $W(z)F_{Q0}(z)$ for the LPF and the BPF are shown in Fig. 3. The solid and dashed lines for $W(z)F_{Q0}(z)$ indicate the variable wordlengths are 2 and 4 bits, respectively.

Passband Ripple: The improvement rates by $W(z)F_{Q0}(z)$ are always very remarkable, compared with others.

Stopband Attenuation: In the case of LPF, $W(z)F_{Q0}(z)$ is superior to $W(z)F_Q(z)$. They are, however, almost the same for the BPF case. However, the efficiency of $F_Q(z)W(z)$ is highly dependent on a desired frequency response.

Coefficient Wordlength Reduction: The proposed approach can shorten the coefficient wordlengths by 3 and 2 bits for the LPF and the BPF, respectively, compared with $H_Q(z)$ and $W(z)F_Q(z)$.

Computing Time

In the case of the LPF with the 4th order weighting function and the 2 bit variable wordlengths, the execution time on the general purpose computer, ACOS System 900. is 97 seconds, which includes the final frequency response calculation.

CONCLUSION

A new discrete optimization method is proposed which can solve high order FIR filter problems within a practically reasonable computing time. The fundamental concept is error spectrum shaping so to be cancelled by other factors. To drastically save computing time, several contrivances are introduced. Design examples for LPF and BPF with 200 taps, show the new approach can decreases coefficient wordlengths by 2 or 3 bits. The computing time, on the general purpose computer, is 97 seconds.

REFERENCES

[1] M.Suk and S.K.Mitra, "Computer-aided design of digital filters with finite word lengths," IEEE Trans. Audio Electroacoust., vol.AU-20, No.5, pp.356-363, Dec. 1972.

[2] N.I.Smith, "A random-search method for designing finite-wordlength recursive digital filters," IEEE Trans. Acoust., Speech, Signal processing, vol.ASSP-27, No.1, pp.40-46, Feb. 1979.

[3] E.Avenhaus, "On the design of digital filters with coefficients of limited word length," IEEE Trans. Audio Electroacoust., vol.AU-20, No.3, pp.206-212, Aug. 1972.

[4] C.Charalambous and M.J.Best, "Optimum of recursive digital filters with finite word length," IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-22, No.6, pp.424-431, Dec. 1974.

[5] J.W.Bandler, B.L.Bardakjian and J.H.K.Chen,"Design of recursive digital filters with optimized word length coefficients," Computer Aided Design, vol.7, No.3, pp.151-156, July 1975.

[6] K.Steiglitz, "Designing shot-word recursive digital filters," Proc. 9th Annu. Allerton Conf. Circuit and System Theory, pp.778-788, Oct. 1971.

[7] F.Brglez, "Digital filter design with short word-length coefficients," IEEE Trans. Circuits Syst., vol.CAS-25, No.12, pp.1044-1050, Dec. 1978.

[8] 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.

[9] D.M.Kodek, "Design of optimal finite wordlength FIR digital filters using integer programming techniques," IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-28, No.3, pp.304-307, June 1980.

[10] V.B.Lawrence and A.C.Salazar, "Finite precision design of linearphase FIR filters," Bell Syst. Tech. J. vol.59, No.9, pp.1575-1598, Nov. 1980.

[11] T.W.Parks and J.H.McClellan,"Chebyshev approximation for nonrecursive digital filters with linear phase," IEEE Trans. Circuit Theory, vol. CT-19, pp.189-194, Oct. 1971.

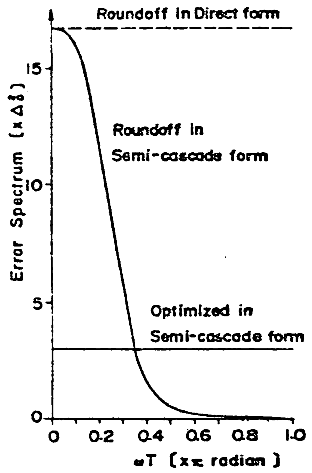

Fig. 1. Example of error spectrum shaping.

Table 1. Filter specifications and discrete optimization parameters.

Parameter	LPF	BPF
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	56-200 Hz	0-50, 150-200 Hz
Attenuation	72.5 dB	80.0 dB
W(z)	$(1+2z^{-1}+z^{-2})$ $x(1+z^{-1}+z^{-2})$	$(1-z^{-2})^2$
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	1, 2 bits	1, 2 bits
K	5	5
K'	1	1
L	4	4
L'	1	1

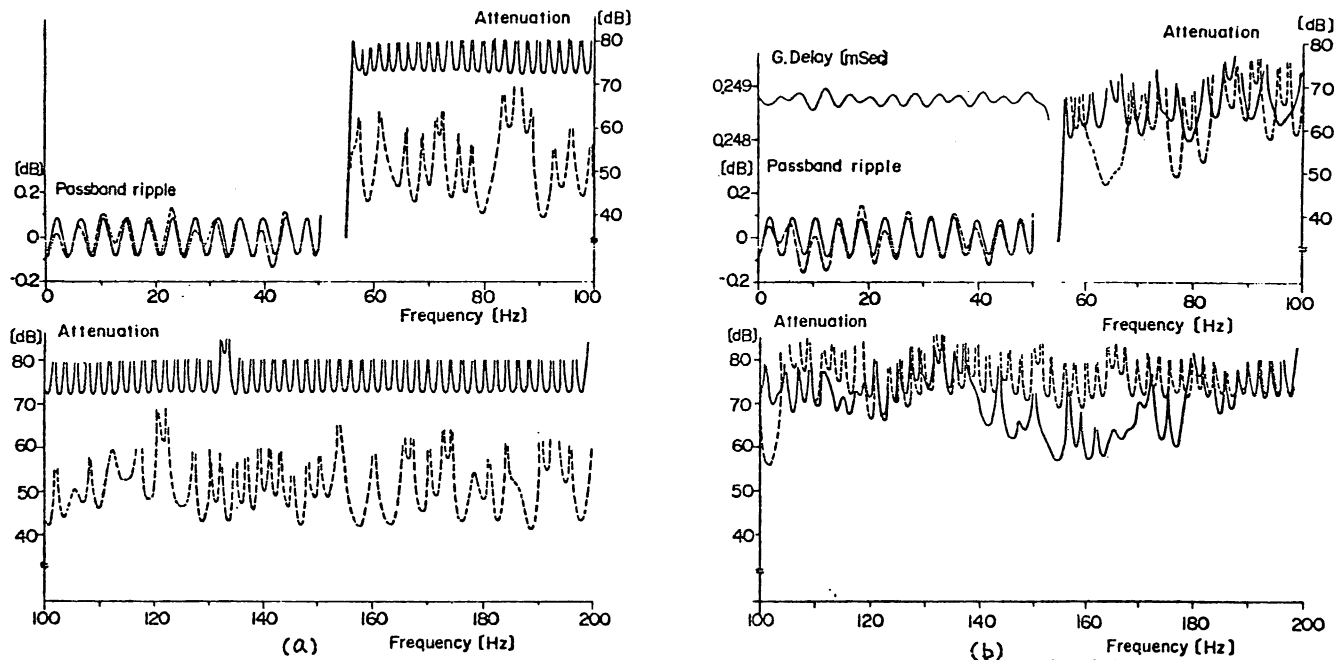

Fig. 2. Optimized frequency responses. (a) Solid line: $W(z)F_\infty(z)$ with infinite precision coefficients, dashed line: $H_Q(z)$ with rounded off coefficients. (b) Solid line: $W(z)F_{Q0}(z)$ with optimized coefficients, dashed line: $W(z)F_Q(z)$ with rounded off coefficients.

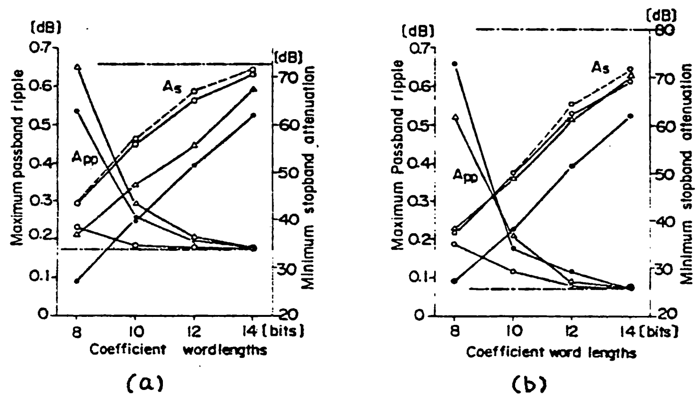

Fig. 3. Frequency response improvements. (a) LPF (b) BPF. Symbols ⊙, △ and O correspond to $H_Q(z)$, $W(z)F_Q(z)$ and $W(z)F_{Q0}(z)$, respectively. $W(z)F_\infty(z)$ is shown by dashed and dotted line.